ClipCrop: Conditioned Cropping Driven by Vision-Language Model

Zhihang Zhong[1,2]   Mingxi Cheng[3]   Zhirong Wu[3]   Yuhui Yuan[3]   Yinqiang Zheng[1]
Ji Li[3]   Han Hu[3]   Stephen Lin[3]   Yoichi Sato[1]   Imari Sato[1,2]
[1]The University of Tokyo, Japan   [2]National Institute of Informatics, Japan   [3]Microsoft Research

Abstract

Image cropping has progressed tremendously under the data-driven paradigm. However, current approaches do not account for the intentions of the user, which is an issue especially when the composition of the input image is complex. Moreover, labeling of cropping data is costly and hence the amount of data is limited, leading to poor generalization performance of current algorithms in the wild. In this work, we take advantage of vision-language models as a foundation for creating robust and user-intentional cropping algorithms. By adapting a transformer decoder with a pre-trained CLIP-based detection model, OWL-ViT, we develop a method to perform cropping with a text or image query that reflects the user's intention as guidance. In addition, our pipeline design allows the model to learn text-conditioned aesthetic cropping with a small cropping dataset, while inheriting the open-vocabulary ability acquired from millions of text-image pairs. We validate our model through extensive experiments on existing datasets as well as a new cropping test set we compiled that is characterized by content ambiguity.

1. Introduction

Images often need to be cropped when circulating among different devices and social media apps. The key requirement is to maintain the aesthetics of the image, while not ruining the intended content. Many datasets [4, 8, 54, 59, 65] with aesthetics-aware annotations have been constructed for benchmarking automatic cropping algorithms. Nevertheless, how to accurately reflect users' intentions of image cropping has not been well studied.

In recent years, data-driven learning has dramatically boosted the performance of image cropping methods. Existing methods can be broadly divided into anchor (sliding window) evaluation methods [5, 28, 42, 43, 54, 65] and coordinate regression methods [11, 16, 27, 37]. There are two problems with the practical application of these methods. First, desirable crop of a given image highly depends on a user's intention, such as how the user want to use the re-

sulting image or from what perspective the user sees the image. If the composition of the input image is complex, such as the presence of many objects, there is no way for general cropping methods to accurately predict the user's intention, but only to make random guesses. Second, due to the tedious process of labeling a cropping dataset, the size of existing datasets is rather small. This results in current approaches being poorly generalized and often failing when faced with unseen scenarios.

The success of vision-language modeling (VLM) [40, 46] gives us inspiration to address the above two challenges. Today, we have access to considerable amounts of multi-modal data, *e.g.*, images with labels or titles. To tackle the first challenge, text can be introduced as additional input to reflect the user's intention, which naturally eliminates content ambiguity during learning. As for the second challenge, transferring VLM's open-vocabulary knowledge from millions of image-text pairs to cropping algorithms is a promising but unexplored direction for more effective cropping in the wild.

In this work, we propose ClipCrop: a vision-language model for generating aesthetic image crops based on text or image queries. The proposed model is built on OWL-ViT [40], which is one of the state-of-the-art open-vocabulary object detection models. We adopt a transformer decoder like DETR [2] to learn aesthetic-aware offsets given the initially predicted bounding box from OWL-ViT. Since no text description is available for cropping training datasets, we provide annotations by complementing GAIC [65] with text descriptions for each image. When fine-tuning our model, we adopt a mosaics composition training strategy, which encourages the model to localize from text queries. The improved model can comprehensively understand input text while retaining the ability of one-shot image-conditioned querying. Taking the source image in Fig. 1 as an example, there are multiple contents of interest in the image, *e.g.*, a woman, three dogs, or a woman and three dogs on the boat. The proposed ClipCrop takes the user text/image query as input to understand the user's intent and generates several crop solutions.
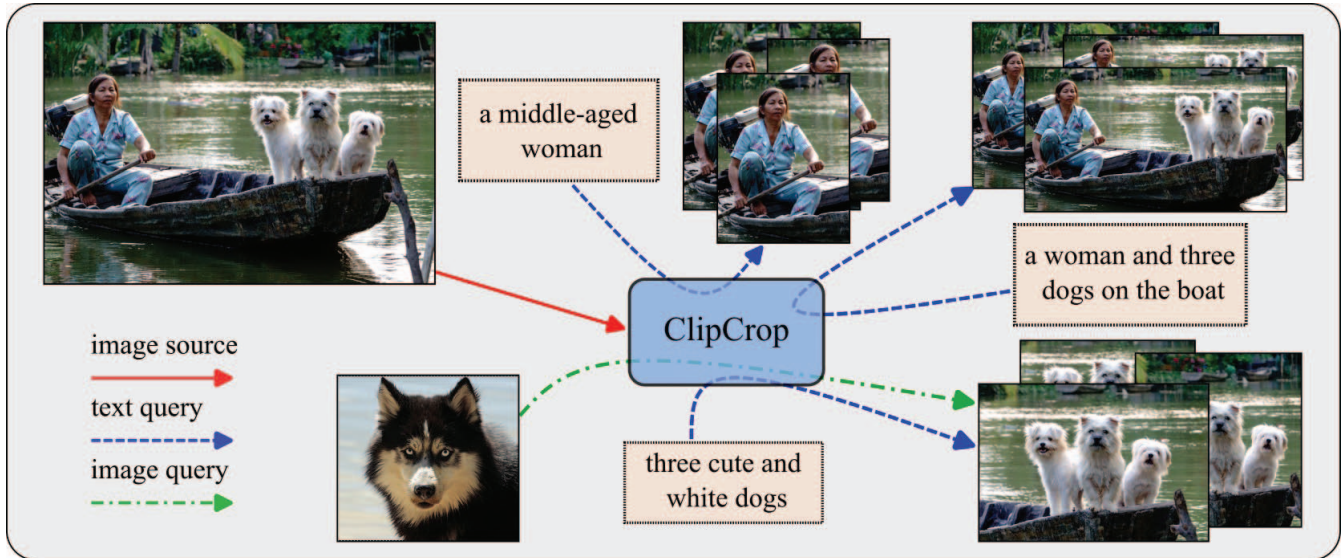
Horanyi-PR [13] is the only cropping evaluation bench-

Figure 1. **The concept of conditioned smart cropping.** Given an image with complex composition, the proposed ClipCrop can accept different text queries, such as "a middle-aged woman", "three cute and white dogs" and "a woman and three dogs on the boat", as conditions to return corresponding aesthetic-aware crops. In addition, ClipCrop can also accept an image as a query condition for cropping.

mark that provides text descriptions, with a single sentence per image. To emphasize the content ambiguity problem for cropping, we further provide an evaluation benchmark, TextCrop for short, with two distinct text annotations and corresponding aesthetics-aware bounding box annotations for each image. This new evaluation benchmark consists of about 200 crowded photos containing multiple contents. Experimental results demonstrate that the proposed ClipCrop achieves significant improvement in terms of average and maximum IoUs in comparison to the prior arts on the Horanyi-PR and the proposed TextCrop test sets. Even in the general cropping setting of GAIC [65], we show that our baseline model performs at a SoTA level and reveal the mismatch between evaluation on this dataset and practical applications.

Our contributions can be summarized as follows.

- We are the *first* to adapt VLMs for conditioned image cropping. The proposed method makes it possible to let the user provide a text or image query to crop, better reflecting the user's intention.

- We present a pipeline to generalize smart cropping models that are trained with expensive yet limited human-annotated data to unseen photos in the wild.

- We verify the performance of our ClipCrop via extensive experiments, and provide a novel evaluation dataset focusing on the challenge of content ambiguity.

2. Related Works

In this section, we review the related literature on image cropping, visual grounding, and vision-language modeling.

2.1. Image Cropping

Image cropping algorithms aim to improve the aesthetic quality of images by removing redundant content and can serve many image editing tasks, *e.g.*, thumbnail generation [14]. The algorithms for cropping can be roughly divided into anchor evaluation methods and coordinate (bounding box) regression methods.

Anchor evaluation is a two-step process of generating and scoring a crop proposal. The first step can be realized through sliding windows [5, 18, 21, 38], predefined anchor boxes [54] similar to detection models [34], or grid anchors [65]. For the second step, distinct quality scorers [42] are created to rank the crop proposals. Many works like Ni *et al*. [41] propose algorithms to mine the implicit composition rules behind photos with high aesthetic quality, such as the "rule of thirds", "rule of space", and "rule of symmetry" popular in the field of photography. Fang *et al*. [8] build a system consisting of three models that learn a visual composition rule, boundary simplicity rule, and content preservation rule for comprehensive scoring. Also, saliency maps [8] and foreground identity [14] are often considered as critical factors for scoring. Yan *et al*. [59] highlight the importance of removed or changed content from the original image for crop quality assessment, while Zeng *et al*. [65] adopt an RoD operation to consider the region of discard. Chen *et al*. [5] presume that a profes-

sional photograph on a website is generally better than its corresponding crop in terms of aesthetics. Therefore, instead of direct scoring, they propose VFN to learn how to compare the quality between two input images by harvesting many examples of unambiguous paired aesthetic rankings from a website. Li *et al*. [28] exploit mutual relations between good crops and bad crops with a graph convolution operation. More recently, TransView [43] provides a transformer-based solution that achieves strong performance and presents interesting visualizations about the attraction and repulsion between visual elements to explain the reasons behind the scores.

For coordinate regression methods, the goal is to directly predict the cropping boxes. Guo *et al*. [10] propose a cascaded pipeline to iteratively regress the final box, while Li *et al*. [26] introduce a reinforcement learning framework to mimic human adjustment decisions for dynamic adjustment of the predicted box. In addition, Li *et al*. propose Mars [27] to realize arbitrary aspect ratio cropping box prediction by meta learning. Instead of fully supervised learning, Lu *et al*. [37] propose a weakly supervised cropping framework guided by the distribution dissimilarity between the predicted cropped images and high-quality images, which eliminates the need for ground-truth boxes annotated by humans. Then, Hong *et al*. propose CACNet [11] that combines an anchor-point-based cropping branch and an interpretable composition branch to learn the offsets of each anchor point to the ground-truth boxes. Inspired by recent advances in object detection [2], Jia *et al*. [16] explicitly address the diversity problem in coordinate regression-based cropping methods by learning in a set prediction fashion.

The closest work to ours is CAGIC [13], which to the best of our knowledge is the first text-conditioned image cropping model. Horanyi *et al*. [13] propose to directly re-purpose pre-trained networks to realize text-conditioned cropping without any fine-tuning. We begin with a VLM perspective and implement more general conditioned cropping, including text querying and image querying. More importantly, our model inherits an open vocabulary capability, which is more stable and accurate for both the dataset presented with GAIC and in practical applications.

2.2. Visual Grounding

The basic goal of visual grounding [19] is to predict and identify a set of bounding boxes surrounding the target objects that are referred to by the given text query. According to the task formulations, we can categorize the existing efforts into two paths, namely phrase grounding [45] which maps a sentence consisting of multiple entities into a set of bounding boxes, and referring expression comprehension [39, 63] which only maps a referring expression to a single bounding box. Most of the previous

works [3, 6, 29, 35, 36, 51, 52, 61, 62] address the visual grounding tasks by exploiting a well-trained existing object detector (*e.g*., DarkNet [48] and DETR [2]) to generate the candidate bounding boxes and then search for the target based on semantic similarities between region features and textual features. Recently, MDETR [17], GLIP [31, 67], and OWL-ViT [40] choose to embrace a scheme that performs end-to-end training and alignment between texts and objects in the image directly. The key difference between visual grounding and conditioned image cropping is that the latter requires consideration of aesthetic elements, and often the box may include multiple objects combined with a background. In this work, we construct our approach based on the very recent OWL-ViT [40] that already exploits rich datasets for visual grounding including OpenImages-V4 [24] and Visual Genome [22].

2.3. Vision-Language Modeling

Vision-language modeling (VLM) attempts to exploit one data modality as the supervision for the other modality. The research on VLM follows the success of BERT [7], where image-conditioned BERT models [35, 52] are developed to facilitate the downstream task of visual question answering and image captioning. Later works [3,20,29,30,32, 53] further incorporate image masked modeling, image text matching, and multimodal fusion for visual grounding and image and text retrieval. BERT-like models require fine-tuning in order to reach satisfactory results.

CLIP [46] revolutionized the field by enabling zero-shot capabilities through large-scale image-caption contrastive learning. The pre-trained CLIP model shows a strong generalization ability to open-set problems with prompt tuning [25] and prompt engineering [33]. The work following CLIP extends the zero-shot capability for detection [40,68], segmentation [9,47,57], and many other applications [49,55,56]. Models that improve upon CLIP include LiT [66], Align [15], UniCL [60], and BASIC [44]. Our work leverages the language interface of CLIP for smart cropping, enabling the user to express intention with regard to image regions.

3. Methodology

In this section, we first review the working mechanism of OWL-ViT [40], which serves as an encoder for our model. Then, we present the structure of our ClipCrop model. Finally, we further elaborate on the training strategy with conditioned cropping.

3.1. Review of OWL-ViT

OWL-ViT uses a standard Vision Transformer as the image encoder $\mathcal{F}_{img}$ and the text encoder $\mathcal{F}_{txt}$. Training is conducted in two stages. In the first stage, $\mathcal{F}_{img}$ and
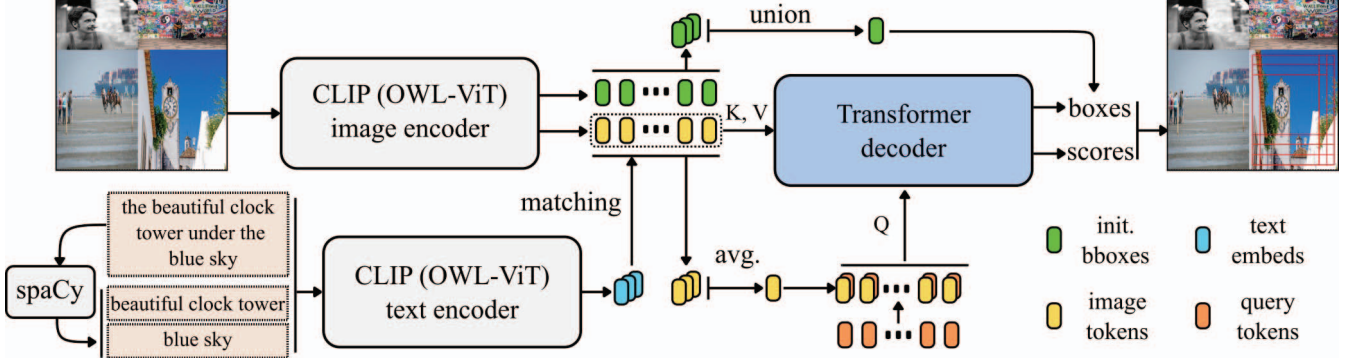
Figure 2. **The architecture of ClipCrop.** In addition to the full input text, ClipCrop uses spaCy [12] to extract keywords and then generates a series of text embeddings via the text encoder of OWL-ViT [40]. These text embeddings are used to match image tokens and corresponding initial bounding boxes from the image encoder of OWL-ViT. To achieve conditioned cropping, a transformer decoder similar to DETR [2] is introduced to regress a series of offsets for the union of the selected initial boxes and predict the corresponding quality scores.

$\mathcal{F}_{txt}$ are contrastively pre-trained with large-scale image-text pairs following the standard CLIP [46] process. In the second stage, the pooling operation and the final projection layer are removed from $\mathcal{F}_{img}$. Then, the output image tokens $\{t_{img}^n\}$ are projected to per-object classification embeddings $\{e_{cls}^n\}$ by a linear layer as well as to the corresponding bounding boxes $\{b^n\}$ by a MLP. In the case of using ViT-B/32 backbone, $n \in \{1, \cdots, 576\}$. This stage is trained using publicly available detection datasets including OpenImages-V4 [24], Objects 365 (O365) [50], and Visual Genome [22]. In inference, a text embedding e_{txt} from $\mathcal{F}_{txt}$ is used as guidance to match the most similar classification embedding. Then, the corresponding box of the matched classification embedding becomes the predicted box.

3.2. Structure of ClipCrop

ClipCrop processes the image tokens $\{t_{img}^n\}$ and the predicted boxes $\{b_{ini}^n\}$ (renaming of $\{b^n\}$) from the image encoder of OWL-ViT as the initial ingredients, as illustrated in Fig. 2. The target of OWL-ViT is visual grounding. This means that the predicted boxes are for single objects without any consideration of aesthetics. To adapt OWL-ViT to conditioned cropping, in addition to the original text input, ClipCrop further uses spaCy [12] to extract keywords from it and generate several text embeddings $\{e_{txt}^w \mid w \in \{1, \cdots, W\}\}$ via the text encoder. Then ClipCrop follows the matching mechanism of OWL-ViT by using $\{e_{txt}^w\}$ to select a series of image tokens $\{t_{img}^w\}$ and the corresponding bounding boxes $\{b_{ini}^w\}$. ClipCrop introduces a new DETR-like transformer decoder $\mathcal{F}_{dec}$ to regress a set of offsets $\{b_{ofs}^m\}$ for the union box b_u of $\{b_{ini}^w\}$ and the corresponding quality scores $\{s^m\}$, as follows:

$$\{b_{ofs}^m\}, \{s^m\} = \mathcal{F}_{dec}(Q, K, V), \qquad (1)$$

where $m \in \{1, \cdots, M\}$ and $M = 90$ by default, as in Jia *et al.* [16]. Key K and value V are from complete image tokens $\{t_{img}^n\}$, and the query Q is the sum of the average of selected image tokens $\{t_{img}^w\}$ and M learnable query tokens $\{t_{qry}^m\}$:

$$Q = \left\{t_{qry}^m + \frac{1}{W}\sum_{w=1}^{W} t_{img}^w\right\}. \qquad (2)$$

The final predicted boxes are the sum of the union box and the regressed offsets:

$$\{b_{pred}^m\} = b_u + \{b_{ofs}^m\}. \qquad (3)$$

3.3. Training Strategy

Our training strategy basically follows that of Jia [16] to supervise the predicted boxes and scores. The Hungarian algorithm [23] is used to find a bipartite matching between predicted bounding boxes and high-quality ground-truth boxes. The high-quality boxes in the GAIC [65] dataset represent boxes with a score of no less than 4. The scores in this dataset range from 1 to 5, with higher scores indicating better. Regarding quality guidance, we also leverage local redundancy [64], *i.e.*, human perception is insensitive to small changes of the crop, to achieve label smoothing. Specifically, if there exists a ground-truth box that has high IoU (IoU $\geq$ 0.9) with a predicted box, the quality score of the predicted box can be supervised using the quality score of that ground-truth box.

To train a conditioned cropping algorithm, there are still several problems. First, there are no paired texts for training. Therefore, we manually annotated a text for each image of GAIC [65] using a crowdsourcing platform. With this enhanced GAIC dataset, there is sufficient data for the community to train conditioned cropping algorithms. Second, because the ground-truth boxes of the GAIC [65] dataset are generated with a fixed grid pattern, the boxes generated by

direct training using a single image will have bias. Hence, we deploy mosaics augmentation [1] to overcome this issue. We randomly sample single images, 2×2 grids, and 3×3 grids with equal probability, while the input text and ground-truth only belong to a single image. However, mosaic augmentation can introduce ambiguity for the input text when different images in the composed mosaic image have similar content. To deal with this, we filter out the initial boxes $\{b_{ini}^w\}$ that are far from the box with the highest quality score in the target image or contain less than half the area of that box.

4. Experiments

4.1. Dataset

Many benchmark datasets [4, 8, 54, 59, 65] have been constructed for image cropping. Among these datasets, CPC [54] and GAIC [65] are densely annotated datasets, while only GAIC is dense enough to directly train a transformer decoder for predicting a set of boxes, as indicated by Jia *et al*. [16]. We train ClipCrop on GAIC [65] and evaluate it on GAIC's test set, Horanyi-PR [13], and the proposed TextCrop.

GAIC [65]   This dataset contains 2636, 200, and 500 images for training, validation, and testing, respectively. Each image has about 90 crop proposals generated along a predefined grid pattern, which is intended to reflect the non-uniqueness of image cropping in practice. For each crop proposal, there is also a quality score ranging from 1 to 5. To train the text-conditioned cropping algorithm, we complemented GAIC by labeling each image with a text description.

Horanyi-PR [13]   This dataset is designed to evaluate the ability to automatically crop images for a given caption. There are 100 images in total for testing, without training and evaluation sets. Each image has a text description and eight aesthetics-aware bounding boxes annotated by different human annotators.

TextCrop   We propose a new cropping test set that is characterized by content ambiguity, named TextCrop for short. TextCrop has 208 crowded photos with various scenarios and styles. We labeled each image with two different text descriptions and four aesthetics-aware bounding boxes from different human annotators for each text.

4.2. Results on Horanyi-PR

First, we compare ClipCrop to the prior arts for visual grounding, general cropping, and text-conditioned cropping on the Horanyi-PR dataset [13]. The visual grounding algorithms that are included, such as ReSC-Large [61]

Table 1. **Comparison on the Horanyi-PR [13] dataset.**

Method	IoU-Mean	IoU-Max
Visual Grounding		
ReSC-Large [61]	0.4531	0.6079
TransVG [6]	0.4398	0.6078
OWL-ViT [40]	0.4131	0.5789
Cropping		
A2-RL [26]	0.3017	0.4506
VPN [54]	0.3231	0.4830
VEN [54]	0.3394	0.4882
Li *et al*. [28]	0.3357	0.4805
GAIC [65]	0.3407	0.4863
CACNet [11]	0.3385	0.4899
Jia *et al*. [16]	0.3333	0.4818
Conditioned Cropping		
CAGIC [13]	0.4160	-
GAIC [65]+TransVG [6]	0.4914	0.6290
ClipCrop	**0.5551**	**0.7284**

and TransVG [6], mainly deal with expression comprehension and return a single bounding box. Regarding the original OWL-ViT [40], we take the bounding box with the highest matching score as the output. The compared general cropping algorithms cover both anchor evaluation methods [28, 54, 65] and coordinate regression methods [11, 16, 26]. All these algorithms are trained in their respective official settings and then tested on the unseen Horanyi-PR dataset. To the best of our knowledge, the only text-conditioned cropping algorithm that exists is the baseline algorithm CAGIC [13] provided with this dataset. Since the implementation code of CAGIC is not published, we directly report the performance given in their paper. In addition, we implemented a cascaded scheme by connecting GAIC [65] and TransVG [6]. The returned crop is the one with the highest GAIC score and can cover the crop generated by TransVG. In this experiment, ClipCrop, CAGIC, the cascaded scheme (GAIC+TransVG), and visual grounding algorithms can utilize the text input of this test set, while the general cropping algorithms cannot.

The quantitative results are shown in Table 1. In addition to the average IoU (IoU-Mean) in [13], we calculate the maximum IoU (IoU-Max) between the bounding boxes labeled by human annotators and the predicted bounding box to reflect the performance. We can see that ClipCrop outperforms CAGIC, general cropping algorithms and visual grounding algorithms by a large margin. We find that the visual grounding algorithms are superior to general cropping algorithms and even CAGIC. The reason why visual grounding algorithms are better than the general cropping algorithms is that the text descriptions of this dataset are biased towards small objects in the image, while general
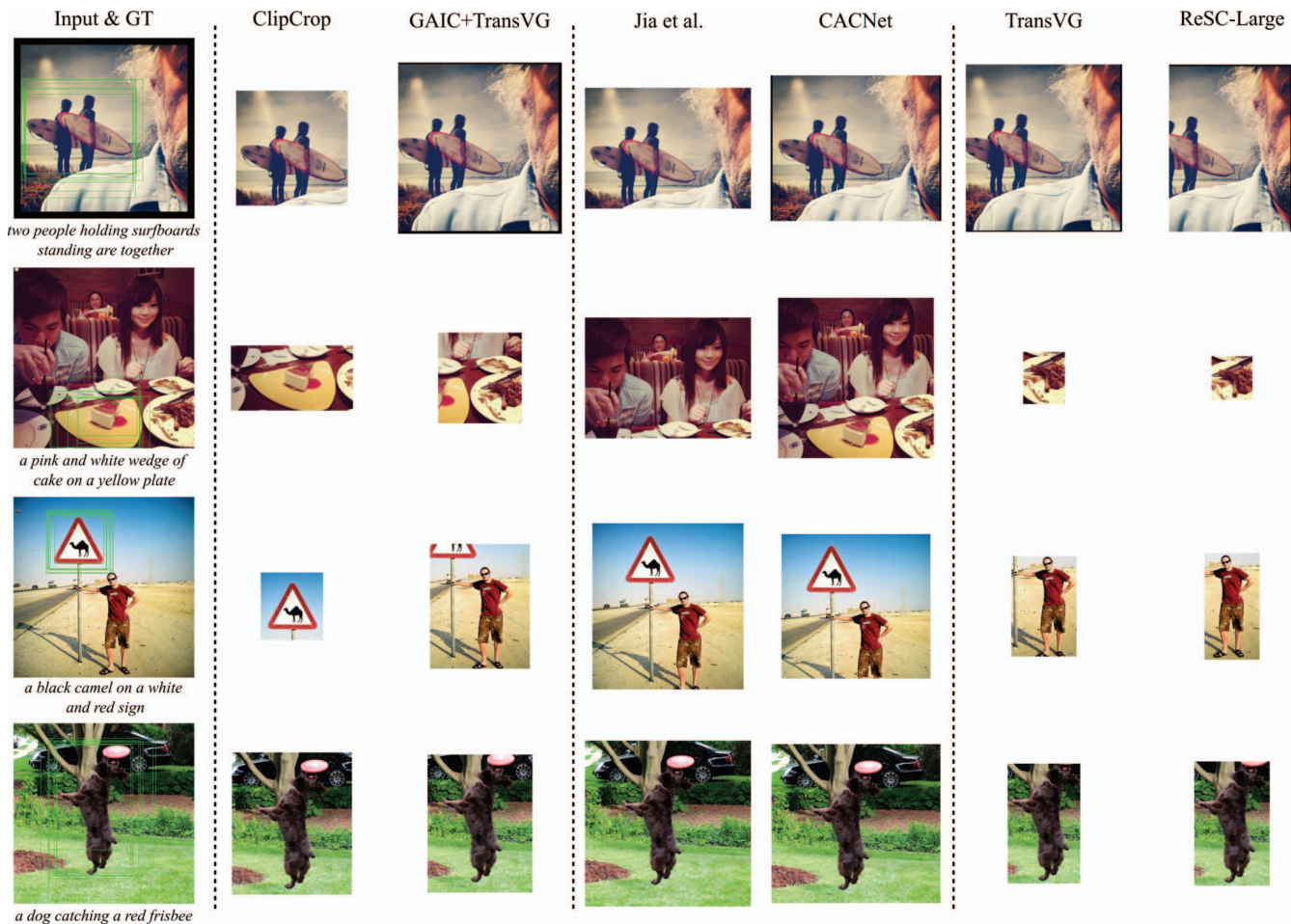
Figure 3. **Visual results on Horanyi-PR [13].** The first column presents the image and text description of the input. The green bounding boxes in the image represent the ground-truth (GT) crops. We show the results of our ClipCrop, cascaded scheme (GAIC [65]+TransVG [6]), general cropping algorithms (Jia *et al.* [16], CACNet [11]), and visual grounding algorithms (TransVG [6] and ReSC-Large [61]) sequentially in the subsequent columns. ClipCrop outperforms other methods by accurately reflecting the intent of the text and ensuring the aesthetic quality of the crop.

cropping algorithms tend to return a large crop that considers the global image. The reason that even CAGIC is inferior to visual grounding algorithms is due to its algorithm design. CAGIC combines a pre-trained aesthetic evaluation network [5] and a pre-trained image captioning network [58] to optimize a bilinear sampler to find a suitable crop, which leads to a low sum of aesthetic loss and caption loss. The match between the input text and the caption generated by [58] is not accurate, resulting in the generated bounding box having a rather poor performance compared to the tight bounding box generated by the visual grounding algorithms, which have a more accurate understanding of the input text. The cascaded scheme can combine the advantages of visual grounding algorithms and anchor evaluation cropping algorithms, but the improvement is limited.

We show qualitative results in Fig. 3. The case in the first row demonstrates that ClipCrop can effectively reflect the

user's intention by removing extraneous background. The examples in the second and third rows demonstrate the powerful generalization capabilities of ClipCrop. ClipCrop can understand complicated descriptions such as "a pink and white wedge of cake on a yellow plate" and "a black camel on a white and red sign", while TransVG and ReSC-Large are unable to localize the expected area. In the case that the main content of the image and input text are more consistent, such as the fourth row, the results of ClipCrop and clipping algorithms are closer.

4.3. Results on TextCrop

We followed the test setup of the Horanyi-PR dataset to test on the TextCrop dataset.

The quantitative results are shown in Table 2. Unlike Horanyi-PR, each image of TextCrop has two different descriptions. Besides local descriptions, global descriptions
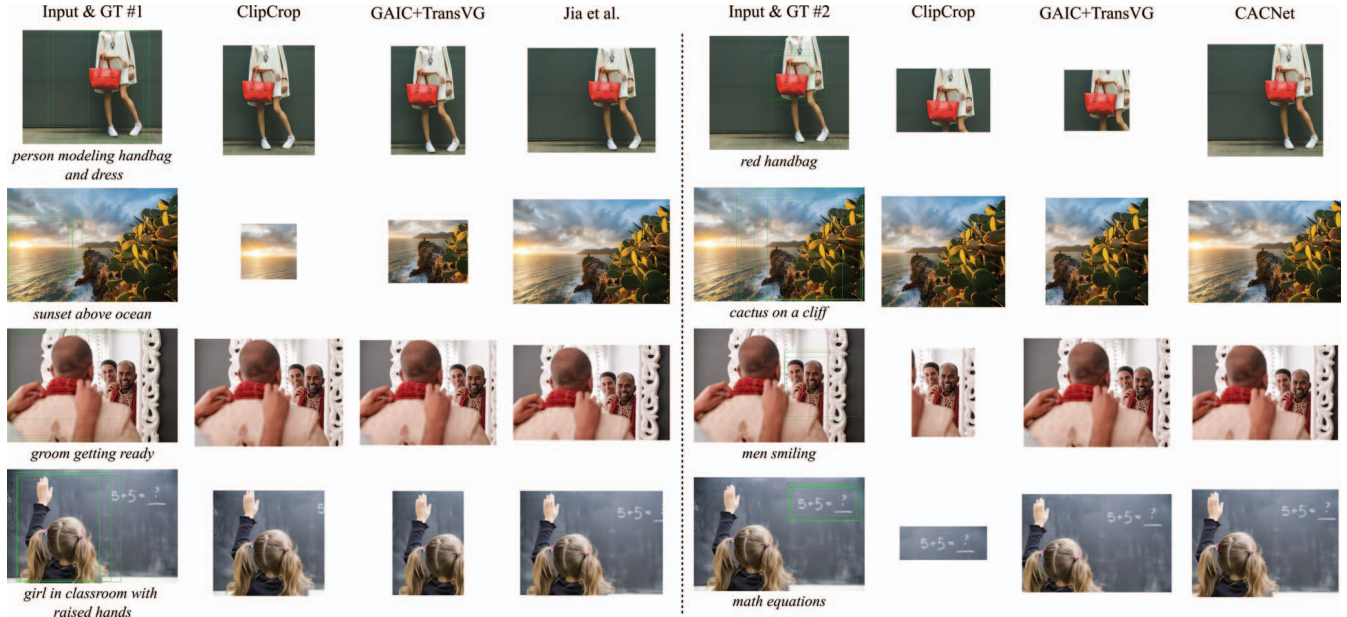
Figure 4. **Visual results on TextCrop.** The left side and the right side show the results of the same image with different text input as query. The green bounding boxes in the input represent the ground-truth. We show the results of our ClipCrop, cascaded scheme (GAIC [65]+TransVG [6]), and general cropping algorithm (Jia *et al.* [16]) for reference. ClipCrop can robustly discern intentions from different texts and crop accordingly.

Table 2. **Comparison on our TextCrop dataset.**

Method	IoU-Mean	IoU-Max
Visual Grounding		
ReSC-Large [61]	0.3116	0.3784
TransVG [6]	0.3363	0.4056
OWL-ViT [40]	0.3923	0.4758
Cropping		
A2-RL [26]	0.3919	0.4800
VPN [54]	0.4052	0.4934
VEN [54]	0.3758	0.4531
Li *et al.* [28]	0.4185	0.5034
GAIC [65]	0.4141	0.4978
CACNet [11]	0.4271	0.5162
Jia *et al.* [16]	0.4317	0.5219
Conditioned Cropping		
GAIC [65]+TransVG [6]	0.4230	0.5031
ClipCrop	**0.4992**	**0.5956**

Table 3. **Comparison on the GAIC [65] dataset.**

Method	$ACC_{1/5}$	$ACC_{1/10}$
VFN [5]	26.6	40.6
A2-RL [26]	23.2	39.5
VPN [54]	36.0	48.5
VEN [54]	37.5	50.5
GAIC [65]	68.2	85.8
TransView [43]	69.0	85.4
Jia *et al.* [16]	85.0	92.6
ClipCrop-base	**87.2**	**94.4**
ClipCrop	79.8	89.6

(GAIC+TransVG) can understand different text queries, such as "person modeling handbag and dress" and "red handbag" in the first row, to crop images that highlight the corresponding subject. However, ClipCrop is more robust than the cascaded scheme in other cases. With "groom getting ready" in the third row and "math equations" in the fourth row, ClipCrop again demonstrates its outstanding open vocabulary capability for text-conditioned cropping in the wild.

4.4. Results on GAIC

In addition to the text-conditioned test sets, we also present a comparison with the GAIC [65], as shown in Table 3. Besides ClipCrop, we show a variant called ClipCrop-base, which does not involve the matching of text
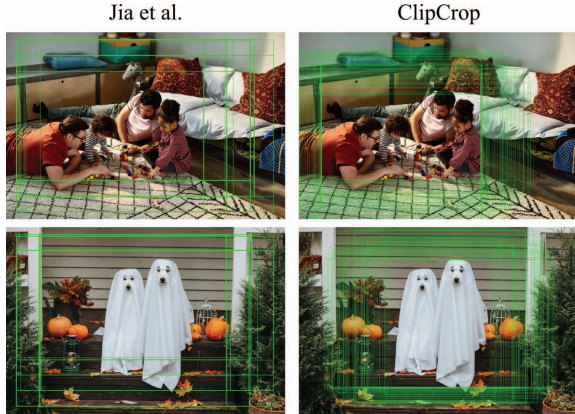
exist for most images. Because of this change, the performance of general cropping algorithms becomes better than visual grounding algorithms. ClipCrop still achieves the best performance, demonstrating its ability to provide a crop that accurately reflects the user's intention and has an aesthetic appeal.

We further present the qualitative results on TextCrop in the Fig. 4. Only ClipCrop and the cascaded scheme

Figure 5. **Pattern of predicted bounding boxes.**

Table 4. **Ablation study on Horanyi-PR [13].**

Method	IoU-Mean	IoU-Max
ClipCrop-base	0.3233	0.4742
ClipCrop-main	0.4550	0.6007
ClipCrop-key	0.5170	0.6848
ClipCrop w/o mosaic	0.5231	0.6901
ClipCrop	0.5551	0.7284

query but rather regresses directly with query tokens and image tokens. With the boost brought by the introduction of the image embedding from the OWL-ViT encoder, this baseline model achieves state-of-the-art performance without considering text. The reasons why ClipCrop has lower performance under this dataset are twofold. First, the annotated description of GAIC images is subjective, which may introduce some noise to the model. Second, the proposed training strategy results in the generated bounding boxes not overfitting to the grid pattern. As shown in Fig. 5, the predicted bounding box pattern of Jia *et al*. [16] is similar to the ground-truth grid pattern of the GAIC [65], while the prediction of ClipCrop is more flexible. Overfitting to the ground-truth pattern is beneficial for the metrics evaluation of this dataset, but harmful for practical applications.

4.5. One-shot Image-conditioned Querying

The design of ClipCrop ensures that it inherits the ability from OWL-ViT to perform one-shot image-conditioned querying. We follow the pipeline of OWL-ViT to extract image embeddings from the query image instead of text embeddings to match the initial bounding boxes and image embeddings from the source image. There are some examples for reference in Fig. 6. The first, second, and third columns represent input image, query image, and crop, respectively. The image-conditioned cropping is a powerful extension to text-conditioned cropping, as it enables cropping the desired areas that are difficult to describe by text but easy to convey by an image.
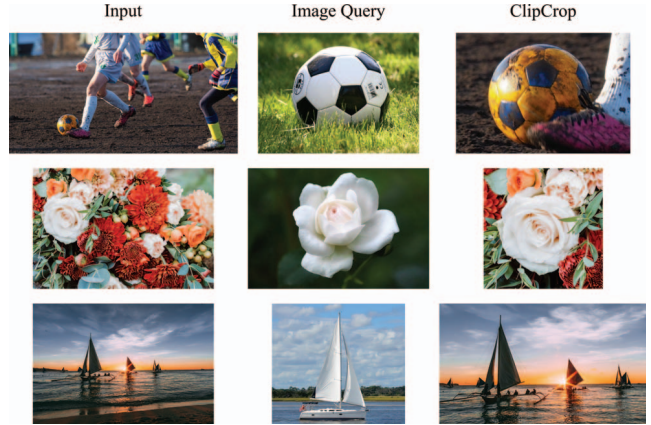

Figure 6. **Results of one-shot image-conditioned querying.** The size of the crop was adjusted for better viewing.

4.6. Ablation Study

To verify the validity of our model design, we conducted an ablation study on Horanyi-PR, as shown in Table 4. ClipCrop w/o mosaic denotes training without mosaic augmentation. ClipCrop-key denotes only using keywords extracted by SpaCy [12] for matching. Clip-main denotes only using the original text description for matching. ClipCrop-base does not use additional text description as input. The results show that additional text input, combination of original text and keywords, as well as mosaic augmentation all facilitate the text-conditioned cropping.

4.7. Limitations

ClipCrop inherits the advantages of VLM while also becoming highly dependent on the pre-trained VLM model. Thus, ClipCrop may include biases introduced by text-image pairs in the contrastive learning stage. Then, because OWL-ViT highly relies on the nouns in the description, the current ClipCrop may not work for descriptions that are too abstract or indirect.

5. Conclusion

In this work, we focus on the challenging problem of how to better reflect the user's intention in the cropping task. Drawing inspiration from vision-language modeling, we presented a novel model, ClipCrop, for text/image-conditioned image cropping. We provided training strategies and designed a network structure that enables ClipCrop to realize conditioned cropping in the wild, with strong open-vocabulary ability. ClipCrop yields superior and robust performance on both the public datasets and the dataset we provide. We believe that introducing pre-trained VLM models into the cropping domain, where data annotation is exceptionally expensive, is a promising direction.

References

[1] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 5

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1, 3, 4

[3] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 3

[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 *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 226–234. IEEE, 2017. 1, 5

[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 *Proceedings of the 25th ACM international conference on Multimedia*, pages 37–45, 2017. 1, 2, 6, 7

[6] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779, 2021. 3, 5, 6, 7

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[8] Chen Fang, Zhe Lin, Radomir Mech, and Xiaohui Shen. Automatic image cropping using visual composition, boundary simplicity and content preservation models. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1105–1108, 2014. 1, 2, 5

[9] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Open-vocabulary image segmentation. *arXiv preprint arXiv:2112.12143*, 2021. 3

[10] 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. 3

[11] Chaoyi Hong, Shuaiyuan Du, Ke Xian, Hao Lu, Zhiguo Cao, and Weicai Zhong. Composing photos like a photographer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7057–7066, 2021. 1, 3, 5, 6, 7

[12] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python. 2020. 4, 8

[13] Nora Horanyi, Kedi Xia, Kwang Moo Yi, Abhishake Kumar Bojja, Aleš Leonardis, and Hyung Jin Chang. Repurposing existing deep networks for caption and aesthetic-guided image cropping. *Pattern Recognition*, 126:108485, 2022. 1, 3, 5, 6, 8

[14] Jingwei Huang, Huarong Chen, Bin Wang, and Stephen Lin. Automatic thumbnail generation based on visual representativeness and foreground recognizability. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 253–261, 2015. 2

[15] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 3

[16] Gengyun Jia, Huaibo Huang, Chaoyou Fu, and Ran He. Rethinking image cropping: Exploring diverse compositions from global views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2455, 2022. 1, 3, 4, 5, 6, 7, 8

[17] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 3

[18] Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1733–1740, 2014. 2

[19] Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. *Advances in neural information processing systems*, 27, 2014. 3

[20] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 3

[21] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *European conference on computer vision*, pages 662–679. Springer, 2016. 2

[22] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 3, 4

[23] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 4

[24] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 3, 4

[25] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 3

[26] Debang Li, Huikai Wu, Junge Zhang, and Kaiqi Huang. A2-rl: Aesthetics aware reinforcement learning for image cropping. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8193–8201, 2018. 3, 5, 7

[27] Debang Li, Junge Zhang, and Kaiqi Huang. Learning to learn cropping models for different aspect ratio requirements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12685–12694, 2020. 1, 3

[28] Debang Li, Junge Zhang, Kaiqi Huang, and Ming-Hsuan Yang. Composing good shots by exploiting mutual relations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4213–4222, 2020. 1, 3, 5, 7

[29] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344, 2020. 3

[30] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 3

[31] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 3

[32] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 3

[33] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*, 2021. 3

[34] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 2

[35] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 3

[36] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446, 2020. 3

[37] Peng Lu, Jiahui Liu, Xujun Peng, and Xiaojie Wang. Weakly supervised real-time image cropping based on aesthetic distributions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 120–128, 2020. 1, 3

[38] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z Wang. Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 457–466, 2014. 2

[39] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 3

[40] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection with vision transformers. *arXiv preprint arXiv:2205.06230*, 2022. 1, 3, 4, 5, 7

[41] Bingbing Ni, Mengdi Xu, Bin Cheng, Meng Wang, Shuicheng Yan, and Qi Tian. Learning to photograph: A compositional perspective. *IEEE Transactions on Multimedia*, 15(5):1138–1151, 2013. 2

[42] Masashi Nishiyama, Takahiro Okabe, Yoichi Sato, and Imari Sato. Sensation-based photo cropping. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 669–672, 2009. 1, 2

[43] Zhiyu Pan, Zhiguo Cao, Kewei Wang, Hao Lu, and Weicai Zhong. Transview: Inside, outside, and across the cropping view boundaries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4218–4227, 2021. 1, 3, 7

[44] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Hanxiao Liu, Adams Wei Yu, Minh-Thang Luong, Mingxing Tan, and Quoc V Le. Combined scaling for zero-shot transfer learning. *arXiv preprint arXiv:2111.10050*, 2021. 3

[45] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 3

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 3, 4

[47] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022. 3

[48] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 3

[49] Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18603–18613, 2022. 3

[50] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 4

[51] Fengyuan Shi, Ruopeng Gao, Weilin Huang, and Limin Wang. Dynamic mdetr: A dynamic multimodal transformer decoder for visual grounding. *arXiv preprint arXiv:2209.13959*, 2022. 3

[52] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 3

[53] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 3

[54] 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 the IEEE conference on computer vision and pattern recognition*, pages 5437–5446, 2018. 1, 2, 5, 7

[55] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4563–4567. IEEE, 2022. 3

[56] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 3

[57] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 3

[58] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015. 6

[59] Jianzhou Yan, Stephen Lin, Sing Bing Kang, and Xiaoou Tang. Learning the change for automatic image cropping. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 971–978, 2013. 1, 2, 5

[60] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19163–19173, 2022. 3

[61] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive subquery construction. In *European Conference on Computer Vision*, pages 387–404. Springer, 2020. 3, 5, 6, 7

[62] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018. 3

[63] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016. 3

[64] Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. Reliable and efficient image cropping: A grid anchor based approach. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5949–5957, 2019. 4

[65] Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. Grid anchor based image cropping: A new benchmark and an efficient model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 2, 4, 5, 6, 7, 8

[66] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022. 3

[67] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *arXiv preprint arXiv:2206.05836*, 2022. 3

[68] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022. 3