Learning Subject-Aware Cropping by Outpainting Professional Photos

James Hong[1], **Lu Yuan**[1], **Michaël Gharbi**[2], **Matthew Fisher**[2], **Kayvon Fatahalian**[1]

[1]Stanford University
[2]Adobe Research
{ james.hong , luyuan , kayvonf }@cs.stanford.edu, { mgharbi , matfishe }@adobe.com

Abstract

How to frame (or crop) a photo often depends on the image subject and its context; e.g., a human portrait. Recent works have defined the subject-aware image cropping task as a nuanced and practical version of image cropping. We propose a weakly-supervised approach (GenCrop) to learn what makes a high-quality, subject-aware crop from professional stock images. Unlike supervised prior work, GenCrop requires no new manual annotations beyond the existing stock image collection. The key challenge in learning from this data, however, is that the images are already cropped and we do not know what regions were removed. Our insight is to combine a library of stock images with a modern, pre-trained text-to-image diffusion model. The stock image collection provides diversity, and its images serve as pseudo-labels for a good crop. The text-image diffusion model is used to outpaint (i.e., outward inpainting) realistic uncropped images. Using this procedure, we are able to automatically generate a large dataset of cropped-uncropped training pairs to train a cropping model. Despite being weakly-supervised, GenCrop is competitive with state-of-the-art supervised methods and significantly better than comparable weakly-supervised baselines on quantitative and qualitative evaluation metrics.

1 Introduction

Framing a photo is compositional skill that professional photographers hone with years of experience, and cropping is a key way to adjust framing or experiment with alternative compositions after capture. Framing and cropping decide what elements of a scene to include or exclude from the image, and how to frame or crop an image is influenced by the subject that one wishes to portray. Subject-aware cropping takes a notion of a subject in addition to pixels and has been studied in recent work on data-driven approaches (Yang et al. 2023) and in the context of human-centric images (Zhang et al. 2022). High-quality solutions to this problem are based on supervised learning, from large datasets of manual annotations created specifically for cropping (Zeng et al. 2019; Wei et al. 2018; Yang et al. 2023).

We explore an alternative approach to the subject-aware cropping problem that is only *weakly-supervised*. We observe that millions of professional images are easily accessible online in the form of stock image collections and that these collections cover a wide range of subject-matter that people want to capture — e.g., portraits of people. We then ask, *to crop better portraits, can one seek out a relevant set of professional portraits from these collections and teach a model to replicate that distribution?* The key challenge is that, although every professional image provides an expert label (i.e., a good crop), the original uncropped photo is unknown and cannot be recovered from the crop. Unlike typical weakly-supervised computer vision tasks where input images are plentiful but labels are scarce and/or unreliable, in our setting this assumption is reversed.

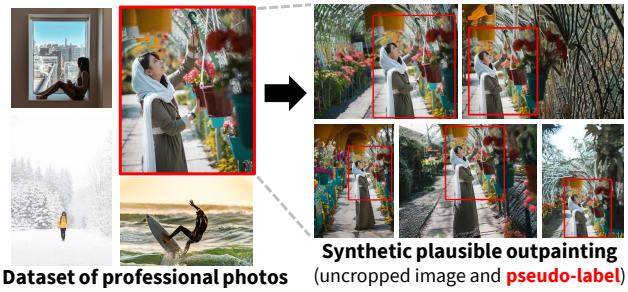


Figure 1: **Generated training pairs.** We outpaint professional images (left) to obtain plausible, uncropped input images (right). The original image is treated as a pseudo-label crop (red). Since the images come from stock image collections, each pseudo-label is an acceptable, professional crop.

Our proposed method, GenCrop, addresses this challenge by combining a readily available dataset of stock images with powerful, pre-trained image generation models to synthesize the required inputs. Specifically, we use text-to-image diffusion to "out-paint" (i.e., outward pixel generate or outward inpaint) stock images and generate plausible uncropped-and-cropped pairs (Fig. 1). By scaling this automatic process, we can generate a large and diverse set of images to train our subject-aware cropping model. The key advantage of GenCrop is that it is weakly-supervised, requiring no new manual crop or scoring annotations beyond access to the original professional image collection.

To demonstrate the effectiveness of GenCrop, we evaluate it on the human-centric (portrait) and the subject-aware cropping tasks proposed by (Zhang et al. 2022) and (Yang et al. 2023). We show that GenCrop yields competitive results against fully supervised approaches (Zhang et al.

2022) on the existing datasets (Fang et al. 2014; Chen et al. 2017a; Yang et al. 2023) (under quantitative metrics such as Intersection-over-Union and boundary displacement (Zhang et al. 2022)), while being superior to the best weakly/unsupervised method (Chen et al. 2017b). We also evaluate GenCrop on additional subject categories such as cats, dogs, etc. to test the generalization of our approach beyond just humans. On qualitative evaluation, GenCrop is comparable to or better than supervised prior work on the rate of cropping errors, while prior weakly-supervised/unsupervised baselines fall substantially short.

Lastly, we conduct additional analysis and ablations to assess the effectiveness and limitations of learning to crop from our generated data. The code for our data generation pipeline and cropping models is publicly available.

2 Background and Related Work

Prior works on image cropping have proposed a broad set of methods, which include optimizing for attention/saliency (Chen et al. 2016; Fang et al. 2014), heuristics (Zhang et al. 2005), and user-interaction (Santella et al. 2006). Recently, end-to-end data-driven approaches have shown strong performance on benchmarks. GenCrop follows this paradigm and we compare to that body of work.

Data-driven approaches for cropping. Most recent works (Hong et al. 2021; Zeng et al. 2019; Zhang et al. 2022; Jia et al. 2022; Pan et al. 2021; Wang et al. 2023) utilize **direct supervision** and require large amounts of human-annotated crops and crop scores to train. GAIC (Zeng et al. 2019), CPC (Wei et al. 2018), and FCDB (Chen et al. 2017a) are commonly used datasets. These datasets are expensive to annotate (CPC and GAIC have 259K and 106K crops; on average 24 and 86 crops per image) and the quality of these images and crowd-sourced crops can vary (see supplemental §B.3). Since these datasets are not specific to a subject type (e.g., human), prior work on human-centric cropping (Zhang et al. 2022) is limited by the number of images available for training and evaluation. Only 1.1K, 339, 176, and 39 of the images in CPC, GAICD, FCDB and FLMS (Fang et al. 2014) are human-centric (Zhang et al. 2022); the human-centric evaluation sets consist of only 50 images from GAICD and 215 from FLMS and FCDB combined. SACD (Yang et al. 2023) is a recent dataset for subject-aware cropping that does not focus on a particular subject type but contains 24K+ labels and 5.2 million ranking pairs generated using their annotation procedure. Our approach differs in that we generate a dataset to provide weak-supervision for a given subject type. We focus on the subject-aware task because cropping better portraits of people is a subtle and important use case, and choosing a subject type (e.g., people) is a simple way to select the most relevant portion of a stock image dataset for the task. Extra experiments in §5.1 and §5.2 suggest that subject-type can be a general object category and that GenCrop can generalize to subject types not targeted during training.

GenCrop is comparable to VFN (Chen et al. 2017b), which is also weakly-supervised by high-quality professional images; VFN generates likely bad crops from within these good images to form ranking pairs. We compare against VFN trained on our stock images and find that GenCrop performs better, showing that our generated dataset is richer than the ranking pairs mined by VFN.

Other works have also experimented with weak or external supervision in addition to the fully-supervised data. Despite being human-centric, HCIC (Zhang et al. 2022) trains on all of the images and annotations in CPC and GAICD. CACNet (Hong et al. 2021) utilizes a second composition classification dataset (Lee et al. 2018). (Wang et al. 2023) trains on the test images, without labels. GenCrop is only weakly-supervised — we only train on the uncropped, pseudo-labeled data generated by our pipeline.

(Zhong et al. 2021) uses outpainting to enlarge the set of possible crops at test time but utilizes GAICD for training.

Models for data-driven approaches can be categorized by which variant of the cropping task they attempt to solve: (1) learning to rank a set of crop candidates (Zhang et al. 2022; Zeng et al. 2019; Wei et al. 2018; Pan et al. 2021) and (2) regressing crops directly (Hong et al. 2021; Jia et al. 2022). GenCrop directly regresses crops and we use an architecture inspired by CACNet (Hong et al. 2021). Like other regression approaches, we quantitatively evaluate using the standard Intersection-over-Union (IoU) and boundary displacement (Disp) metrics against human-annotated crops. We do not focus on ranking metrics (e.g., SRCC) but provide those results in supplemental §A.9.

Dataset generation using text-image diffusion. (Sarıyıldız et al. 2023; Tian et al. 2023) use Stable Diffusion (SD) (Rombach et al. 2022) to synthesize data for ImageNet (Deng et al. 2009) and other generic image classification tasks. InstructPix2Pix (Brooks, Holynski, and Efros 2023) generates a dataset for text-driven image edits. GenCrop also leverages the image generation capabilities of SD but to transform stock images into labels for cropping. Techniques to generate data for other spatial tasks (e.g., object grounding or placement) are interesting future work.

Other works have studied the **outpainting task** directly (Teterwak et al. 2019; Yang et al. 2019; Cheng et al. 2022). Recently, outpainting using text-image diffusion (Ramesh et al. 2022; Adobe 2023) has been shown to be a powerful interactive tool in the hands of artists. Future work on automatic outpainting would benefit the quality of the training data produced by our pipeline.

Pre-trained models used by GenCrop. In addition to Stable Diffusion (SD) (Rombach et al. 2022), we use other off-the-shelf models in our pipeline. These include an image captioner (Li et al. 2023) and an instance segmenter (Ultralytics 2023). Text-conditioning, even with noisy captions, helps SD produce more plausible outpainted images. The instance segmenter is used to detect and segment the subject as an input to our model. While the YOLOv8 model that we use is limited to the COCO (Lin et al. 2014) object classes, we anticipate that advances in models such as SAM (Kirillov et al. 2023) will enable arbitrary subject classes.

3 Methods

Given an image, GenCrop generates possible crop rectangles that lead to aesthetically pleasing compositions. Our method

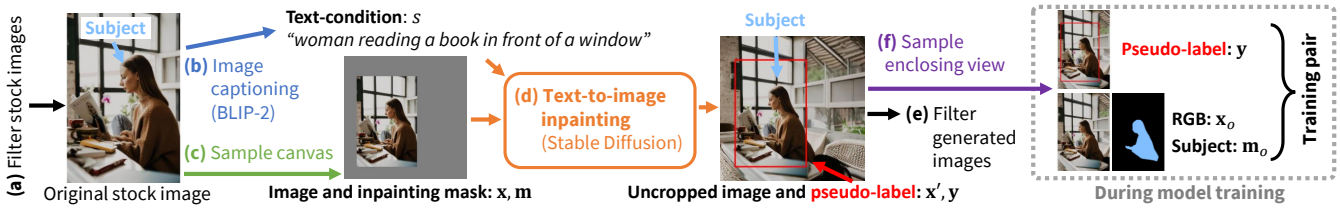
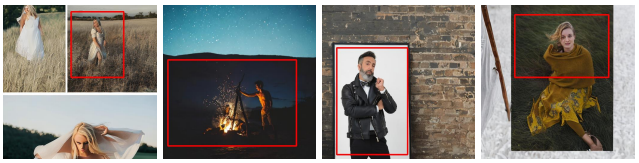


Figure 2: **Dataset generation pipeline.** Stages are marked (a-f). Refer to §3.1 for detailed explanation. We start with a stock image (a) and estimate its text caption (b). To determine the region to be outpainted, we sample a blank canvas to outpaint around the image (c). Outpainting is done using a text-to-image inpainting model (Rombach et al. 2022) and results in a square image (d). Afterwards, we apply automated filters to remove poorly generated images (e). Later, when training a cropping model, we sample an enclosing view (f) in the uncropped image from a common aspect (e.g., $3:4$) so that the model generalizes beyond square images. The region containing the original image is treated as a pseudo-label when training a cropping model.

(a) Additional outpainted subject



(b) Tiled, composite, or framed images

Figure 3: **Common outpainting failure cases.** The original image is marked with a red rectangle. (a) An extra person was synthesized in the outpainted region. This can alter the ideal composition of the scene. (b) The outpainted region is a grid or composite of multiple images (col 1, 2), frames the original image (col 3), or has a border (col 4). These artificial edges can bias the model towards detecting sharp borders.

is "subject-aware". We condition the cropper on both the input image and an estimated pixel mask denoting the location of a given subject. The core of our approach is to generate a dataset of cropped and uncropped image training pairs using a pre-trained generative model (§3.1). We then use the generated data to train a subject-aware cropping model (§3.2).

3.1 Dataset Generation with Image Outpainting

Our first goal is to construct a dataset of image pairs, one casually-framed and one expertly-framed, to supervise our cropping model. Because our dataset generation is automated, we synthesize images according to a subject type (i.e., cropping portraits like professional portraits). We filter the stock image collection (Unsplash 2023) to a set of relevant images. These photos, by nature of their inclusion in the stock photo collection, have been vetted for good composition and high aesthetic quality by human experts. We then use a pre-trained diffusion model to hallucinate plausible out-of-frame content for each image. Fig. 2 illustrates our pipeline and a generated cropped-uncropped training pair.

For each stock image, we apply the following operations:

1. *Pre-processing and filtering.* We filter for images that include an identifiable subject (e.g., person in portraiture; Fig. 2a). This is done with metadata tags first and then with an object detector (Ultralytics 2023). We also discard the image if it contains too many possible subjects (e.g., > 5). For simplicity, if there are multiple possible subjects, we select the largest one as the dominant subject (by bounding box area).
2. *Estimating image captions.* We use an automatic image captioning algorithm (Li et al. 2023) to estimate a text-conditioning string: s (Fig. 2b). The purpose of this text conditioning is to constrain the content generated by the diffusion model. Omitting it can lead to unrelated contexts in the outpainting; see supplemental Fig. 6.
3. *Outpainting.* We randomly downscale the image with bilinear interpolation and paste it into a surrounding 512×512 canvas to obtain an image $\mathbf{x}$ (Fig. 2c). We also compute a binary mask $\mathbf{m}$ with 1's in the area corresponding to valid pixels. We then pass $\mathbf{x}$ and $\mathbf{m}$ to a pre-trained diffusion inpainting model (Rombach et al. 2022) to obtain a 512×512 outpainted image (Fig. 2d):

$$\mathbf{x}' := \texttt{StableDiffusionInpaint}(\mathbf{x}, \mathbf{m}, s)$$

   Because we wish to learn to crop images of different aspect ratios, not just the 1:1 square images produced by Stable Diffusion, we sample a rectangular crop $\mathbf{x}_o$ from inside $\mathbf{x}'$ that also encloses the original image. $\mathbf{x}_o$ is chosen to have a common aspect ratio (e.g., 2:3, 4:5, etc.). We refer to the coordinates of the original image region in $\mathbf{x}_o$ as $\mathbf{y} \in \mathbb{R}^4$ and treat this as a crop pseudo-label. We also run an object segmenter (Ultralytics 2023) to update the subject's bounding box (since a cropped subject may grow in size due to outpainting) and to produce the subject mask, $\mathbf{m}_o$, needed for GenCrop. Together, $(\mathbf{x}_o, \mathbf{m}_o)$ and $\mathbf{y}$ form a weakly-labeled training pair. (Fig. 2f).
4. *Outpainting quality filtering.* Not all outpainting attempts lead to plausible images (see Fig. 3); we discard the most striking failure cases (Fig. 2e) using two automatic heuristics described later in this section.

We repeat the pipeline multiple times per image to sample additional variations for data amplification ($4\times$ to $8\times$ depending on the initial number of stock images available).

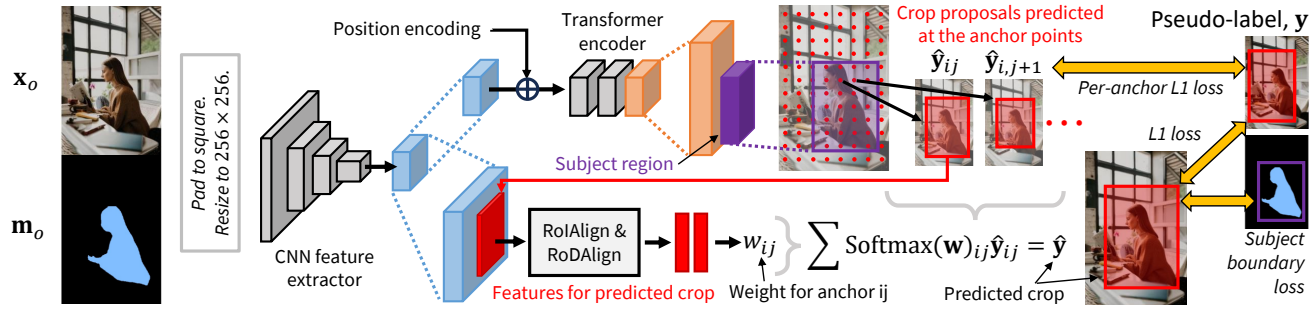


Figure 4: **Cropping model architecture.** Our design is inspired by CACNet (Hong et al. 2021): details in §3.2 and supplemental §C.2. We extract CNN features from the input image, $\mathbf{x}_o$, and subject mask, $\mathbf{m}_o$. These features are used by a transformer-encoder (Vaswani et al. 2023) to generate crop proposals at a grid of anchor points. The crop proposal at each anchor point contained in the subject region is weighted by a second branch. A softmax-weighted sum computes the final crop prediction $\hat{\mathbf{y}}$.

Filtering the uncropped images. We discard two prominent classes of images where Stable Diffusion (SD) (Rombach et al. 2022) fails to outpaint a reasonable scene.

1. Images with a new subject (e.g., another person) in the outpainted region (Fig. 3a). This can happen if the subject or if the original region (in $\mathbf{x}_o$) is small.
2. Images depicting a grid of images or scenes that "frame" the original image (Fig. 3b). These images leak information about the crop label.

These images account for about 20% of the outpainted portrait images. To suppress images of category (1), we run a heuristic that compares the size of the largest object of the subject class (e.g., person) in the outpainted region with the size of the originally identified subject. If another object has an area larger than $\frac{1}{4}$ of the subject's area, the image is discarded. To address category (2), we separately train a binary classifier, $D_{quality}$, to identify grid, composite, and bordered images. $D_{quality}$ is a standard ResNet-50 (He et al. 2016) that we trained on 3K generated images (of which 15% are 'bad') and it identifies approximately 90% of the images with such issues (details in supplemental §C.1). GenCrop can still train without these filters, but we show via ablations in supplemental §A.4 that removing these problematic data boosts results slightly (up to 0.03 IoU and 0.008 Disp).

Images produced by current text-image diffusion models often have artifacts (e.g., in fine details such as faces; zoom into Fig. 1-right). With Stable Diffusion inpainting, these artifacts appear even in the non-inpainted regions; we do not correct them and find that they do not impede training of the cropping model. Blending the original (cropped) image into the outpainted synthetic image would reduce artifacts, but it would also leak information about the crop pseudo-label due to the absence of artifacts (Wang et al. 2020).

3.2 Cropping Model and Training

We describe the cropping model used in our experiments.

Model inputs and outputs. The input is an uncropped image $\mathbf{x}_o$ and subject mask $\mathbf{m}_o$, which contain the box specified by the crop pseudo-label $\mathbf{y}$. We scale and zero pad $\mathbf{x}_o$ and $\mathbf{m}_o$ to 256×256, since our model is not fully convolutional. The output is a vector $\hat{\mathbf{y}} \in \mathbb{R}^4$.

Architecture. At a high level, our model design is inspired by CACNet (Hong et al. 2021), sharing a similar three-part structure of a multi-scale CNN feature extractor and two branches: one for regressing crops at a grid of anchor points ('cropping branch') and the other for producing blending weights for the crops ('composition branch').

The feature extractor is a ResNet-50 that accepts four-channel inputs: the RGB image and the binary subject mask. From the 256×256 inputs, we obtain a 16×16 grid of features. Our cropping branch is a small 2-layer transformer-encoder (Vaswani et al. 2023) that can easily learn global interactions between distant parts of the image. Each of the 16×16 transformer inputs and outputs corresponds to an anchor point, and each anchor point produces one crop proposal: $\hat{\mathbf{y}}_{ij} \in \mathbb{R}^4$ at the ij'th anchor point. In our additional experiments (§5.3), we extend this component to inject conditional control into our cropping model by replacing the transformer-encoder with a transformer-decoder (Vaswani et al. 2023). In our composition branch, we predict blending weights for the crop proposals. Since we want to prioritize the subject, we use zero weights for proposals by anchor points that lie outside the subject's bounding box. For anchor points within the subject bounding box, we use RoIAlign (He et al. 2017) and RoDAlign (Zeng et al. 2019) layers to pool the spatial CNN features from inside and outside of a crop proposal, similar to prior cropping work (Zhang et al. 2022). A feed-forward network uses these features to produce a weight: w_{ij} for the ij'th proposal. The final crop prediction, $\hat{\mathbf{y}}$, is the softmax-weighted sum:

$$\hat{\mathbf{y}} = \sum_{i=1}^{16}\sum_{j=1}^{16} \texttt{Softmax}(\mathbf{w})_{ij}\hat{\mathbf{y}}_{ij}$$

Fig. 4 shows our architecture and supplemental §C.2 provides implementation details and comparison to CACNet.

Losses. We train using a *regression loss*: the L1 error between $\hat{\mathbf{y}}$ and $\mathbf{y}$. We also add two additional losses.

1. *Per-anchor regression loss.* We apply L1 loss between the predicted crop at each anchor point and $\mathbf{y}$, with a low loss weight ($\frac{1}{10}$). This reduces variance by encouraging all anchors to make predictions similar to the ground truth (i.e., reasonably shaped boxes).

Subject	Train # imgs	Train # outpainted	Val # imgs	Test # imgs	Test # labeled
Human	52.0K	188K (3.6×)	11.0K	10.3K	1000
Cat	6.7K	18K (2.7×)	838	804	204
Dog	8.9K	24K (2.7×)	1.1K	1.2K	200
Bird	9.0K	25K (2.8×)	1.3K	1.8K	201
Horse	2.2K	12K (5.3×)	221	412	200
Car	9.2K	23K (2.5×)	891	811	100

Table 1: **Dataset statistics and splits for each subject class.** # outpainted is the number of synthetic training images that pass the automatic quality filters in §3.1. # labeled is the size of our hand-labeled evaluation subset (of the test split).

2. *Subject boundary loss.* A common error that we notice when training on the aforementioned losses is that the model produces crops that cut subjects at their extremities (e.g., tip of the feet; Fig. 5-left). This is unflattering and is difficult to penalize because the predicted crop can be within pixels of the true crop. Manually-annotated ranking datasets may explicitly label such crops as bad (Wei et al. 2018). Without relying on manual annotations, we introduce a margin L1 loss to discourage crops within 2.5% of the subject mask's bounding box. To avoid overriding real labels, this loss is applied only if the label also does not crop on or near the subject.

Implementation. We optimize the network end-to-end using AdamW (Loshchilov and Hutter 2019) and cosine annealing (Loshchilov and Hutter 2017). In addition to the input processing described above, we apply standard image augmentations (e.g., flip, color jitter, blur, distortion, etc.). See supplemental §C.2 for hyper-parameters and details.

4 Stock Image Dataset

Unsplash contains over three million curated images and is publicly accessible for research use (Unsplash 2023). While our primary motivation is to crop human portraits, we also experiment with five other categories: cats, dogs, horses, birds, and cars. We select the relevant images using provided metadata and off-the-shelf object detection as described in §3.1. Because Unsplash reflects a real-world distribution of images and subject matter submitted by photographers, the amount of data by subject varies. After filtering, we are left with 73K, 8K, 11K, 2.8K, 12K and 11K images for portraits, cats, dogs, birds, horses, and cars. We designate a fraction of the images for test and validation; see Tab. 1.

4.1 Evaluation Sets for Subject-Aware Cropping

Prior cropping datasets such as FLMS (Fang et al. 2014), FCDB (Chen et al. 2017a), and SACD (Yang et al. 2023) lack the quantity of images in any particular subject category needed to serve as evaluation (having only 500, 348, and 290 test images total). In order to evaluate GenCrop, we construct new evaluation sets for the six aforementioned subjects, derived from the Unsplash testing images. We select Unsplash images with space for tighter crops (alternative framings) and task the model to produce crops that preserve the aesthetic qualities of a good composition; *a good cropping model should not produce bad crops that violate compositional norms* (e.g., by cutting through a person at a joint). Therefore, to create an evaluation set, one author of this paper, who is a photography domain expert, annotated crops for 1,905 images of the six aforementioned subjects (see Tab. 1), producing 2.3 good crops per image on average. The annotations for this data and the images are all publicly available. We also use these images for qualitative evaluation in §5.2, where we measure the rate at which cropping methods produce common framing errors.

Method	Trained on	WS	FCDB+FLMS IoU ↑	FCDB+FLMS Disp ↓	SACD IoU ↑	SACD Disp ↓
VFN	CPC		*0.6509	*0.0876	-	-
LVRN	CPC		*0.7373	*0.0674	†0.6962	†0.0765
GAIC	CPC		*0.7260	*0.0708	-	-
GAIC	GAICD		-	-	†0.7124	†0.0696
CGS	CPC		*0.7331	*0.0689	-	-
CACNet	FCDB,KUPCP		*0.7364	*0.0676	0.7109	0.0716
HCIC	GAICD		-	-	0.7120	0.0683
HCIC	CPC		***0.7469**	***0.0648**	0.7109	0.0712
FRCNN-m	SACD		-	-	†0.7306	†0.0587
SAC-Net	SACD		-	-	†**0.7665**	†**0.0491**
VFN	Flickr	✓	*0.5115	*0.1257	†0.6690	†0.0887
VFN	Unsplash	✓	0.5783	0.1114	0.6555	0.0775
GenCrop	Unsplash	✓	**0.7334**	**0.0687**	**0.7301**	**0.0632**

Table 2: **Quantitative comparison on existing benchmarks.** The best result in each category is **bold**; WS indicates weakly-supervised. * and † are results reported by (Zhang et al. 2022) and (Yang et al. 2023), respectively. FCDB+FLMS is the human-centric test-set used by (Zhang et al. 2022). SACD is the subject-aware dataset from (Yang et al. 2023), which includes non-human subjects. GenCrop refers to our model trained on outpainted human portraits.

5 Experiments

In §5.1, we evaluate GenCrop quantitatively on the existing subject-aware cropping benchmarks (Yang et al. 2023; Zhang et al. 2022) and our class-specific test sets described in §4.1. In order to control subjectivity and to provide clearer insight into cropping model failures, we measure the violation rate for crops on a set of five pre-determined, objective aesthetic-quality guidelines (§5.2). We also conduct additional experiments and ablations in §5.3 and §5.4.

5.1 Quantitative Evaluation

Metrics. Intersection-over-Union (IoU) and boundary displacement (Disp) are common metrics for cropping evaluation used in prior work (Zhang et al. 2022; Yang et al. 2023). If there are multiple ground-truth labels in an image, we follow the standard protocol of using the label with the top IoU to evaluate the prediction. While IoU and displacement metrics are not fully informative of cropping quality, they provide a standardized way to compare to existing methods.

Method	Trained on	Weak sup	Human		Cat		Dog		Bird		Horse		Car	
			IoU ↑	Disp ↓	IoU ↑	Disp ↓	IoU ↑	Disp ↓	IoU ↑	Disp ↓	IoU ↑	Disp ↓	IoU ↑	Disp ↓
CACNet	FCDB,KUPCP		0.749	0.062	0.740	0.065	0.742	0.063	0.696	0.076	0.757	0.060	0.727	0.068
HCIC	GAICD		0.723	0.065	0.733	0.065	0.740	0.061	0.696	0.074	0.754	**0.059**	0.730	**0.064**
HCIC	CPC		**0.750**	**0.060**	**0.769**	**0.056**	**0.759**	**0.057**	**0.714**	**0.069**	**0.759**	**0.059**	**0.735**	0.065
VFN	Unsplash	✓	0.622	0.095	0.633	0.093	0.621	0.095	0.573	0.106	0.623	0.093	0.633	0.088
GenCrop	Unsplash	✓	**0.750**	**0.061**	**0.777**	**0.054**	**0.758**	**0.058**	**0.712**	**0.071**	**0.757**	**0.059**	**0.744**	**0.063**
GenCrop-6 (all 6)		✓	**0.752**	**0.060**	0.767	0.057	0.748	0.061	**0.719**	**0.069**	**0.760**	**0.058**	0.742	0.062
GenCrop-H (humans)		✓	0.750	0.061	0.767	0.056	0.751	0.059	0.711	0.070	0.748	0.061	**0.748**	**0.060**

Table 3: **Quantitative comparison on different subject types.** Best results per category are **bold**. We evaluate HCIC (Zhang et al. 2022) without its human-specific feature partitioning scheme for non-human subjects. GenCrop, trained on synthesized data of the subject category (middle), is competitive with supervised methods (top). To test whether specialization of the training data is necessary, we test GenCrop-6, trained jointly on all six categories, and GenCrop-H, trained on humans only but applied to other categories (bottom). The results are similar between all three GenCrop variations, suggesting a degree of generalization.

Method	Human	Cat	Dog	Bird	Horse	Car	Mean
VFN	52	41	59	60	55	40	51.2
CACNet	10	9	13	17	23	9	13.5
HCIC (CPC)	9	10	12	12	7	**3**	8.8
GenCrop	11	7	5	3	15	**3**	**7.3**
GenCrop-H (human)	11	6	5	9	10	4	7.5
GenCrop-6 (all 6)	**8**	**3**	**3**	**1**	**4**	8	**4.5**

Table 4: **Qualitative results.** Percentage of images with *one or more* violations (↓ is better; definitions in §5.2). GenCrop-H and GenCrop-6 are trained on human and all 6 classes. See supplemental Tab. 1 for the full breakdown.

Method	V1	V2	V3	V4	V5
VFN	19.7	10.1	11.0	2.6	5.3
CACNet	6.8	1.8	4.5	1.0	1.8
HCIC (CPC)	3.7	**0.8**	2.8	**0.6**	2.1
GenCrop	2.5	2.8	1.0	1.5	**0.6**
GenCrop-H (human)	1.7	2.5	1.5	2.0	1.7
GenCrop-6 (all 6)	**0.8**	1.3	**0.3**	1.7	1.0

Table 5: **Qualitative results.** Percentage of images with violations, by violation type and aggregated across the six subject classes (↓ is better; definitions in §5.2). GenCrop-H and GenCrop-6 are trained on human images and all of the data.

Baselines. We compare GenCrop to HCIC (Zhang et al. 2022), CACNet (Hong et al. 2021), and VFN (Chen et al. 2017b) and reported results in prior work (Yang et al. 2023).

HCIC and CACNet are recent, supervised methods with publicly available code; HCIC is trained on CPC (Wei et al. 2018) or GAICD (Zeng et al. 2019) and is subject-aware, while CACNet is trained on FCDB (Chen et al. 2017a) and KUPCP (Lee et al. 2018). VFN (Chen et al. 2017b) is weakly supervised (though it can also be trained in a supervised manner on CPC). For direct comparison to GenCrop, we train VFN on Unsplash images, including our subject masks and using the same ResNet-50 backbone.

Figure 5: **Examples of crops with subtle mistakes** (input on left; crop on right). First pair: the crop cuts through the subject's feet. Second pair: the crop leaves clutter on the edges and places the subject neither centered for left-right symmetry nor at a third, resulting in an unbalanced image.

Figure 6: **Conditional cropping model.** We are able to sample crop variations by varying the "area" (top) and "aspect ratio" (bottom) conditioning at inference time.

Comparison on prior benchmarks. Tab. 2 compares our GenCrop to prior work on existing, published benchmarks. FCDB + FLMS is the human-centric test set used by (Zhang et al. 2022). SACD is the test set from (Yang et al. 2023), which also contains non-human images. We use GenCrop trained on outpainted human portraits for these experiments.

GenCrop is significantly better than the comparable VFN (up to 0.15 IoU and 0.04 Disp) and is competitive with supervised methods on both datasets. GenCrop is within 0.014 IoU and 0.004 Disp to HCIC on FCDB + FLMS and better than HCIC, GAIC, and CACNet on SACD (by around 0.02 IoU and 0.005 Disp). On SACD, GenCrop is within 0.036 IoU and 0.014 Disp of SAC-Net (Yang et al. 2023), a method tailored around the human-annotated label structure in the SACD training data. Despite SACD not being human-subject exclusive, GenCrop trained on human portraits is able to show generalization ability — more so than other supervised methods trained on GAICD, CPC, and FCDB.

Comparison on the Unsplash test sets from §4.1. Tab. 3 shows results on the six categories: humans, cats, dogs, birds, horses, and cars. GenCrop is trained on generated data filtered by subject category, and it significantly outperforms VFN, while remaining competitive with supervised methods. We also test two additional versions of GenCrop: (1) trained jointly on the six categories (GenCrop-6) and (2) trained on humans only (GenCrop-H). All three variations of GenCrop perform similarly, showing that specialization of the synthetic dataset to the subject category is not necessary (though it can obtain similar results with less data).

5.2 Qualitative Evaluation

Image cropping is challenging to evaluate as the best crop is subjective. Quantitative metrics such as IoU and Disp are also not fully informative (Zeng et al. 2019). Narrowing the evaluation by subject type (e.g., human portraits) can be more objective since there are well-established technical rules on what makes a bad crop. For instance, cropping a person through a joint (e.g., ankle, knee, elbow) is generally regarded as unflattering (Popular Photography 2016; Northrup and Northrup 2019). We perform a qualitative evaluation that counts the number of such violations in a sample of 100 images per class. Specifically, we consider five (non-mutually-exclusive) questions: *Does the crop:*

1. *cut unnaturally through the subject?*
2. *cut unnaturally through the scene (e.g., other objects)?*
3. *have too much or too little negative space?*
4. *have or create unnecessary clutter around the edges?*
5. *lack balance (e.g., missing symmetries, rule-of-thirds)?*

Fig. 5 shows two crops with subtle mistakes (to casual observers). Please refer to supplemental §A.2 for more detailed explanations and additional visual examples.

While these five criteria are not exhaustive, they reflect common errors that we observed and critiques that a poorly-composed image might receive. Also, although experts may deliberately violate these rules for artistic expression, we observe empirically that cropping model failures on these criteria are indicative of bad crops.

Tab. 4 and Tab. 5 compare the methods by the subject category and the violation type, respectively. Like in Tab. 3, we compare the default (specialized) GenCrop, GenCrop-H trained on human images only, and GenCrop-6 trained jointly on all six classes. All three versions of GenCrop are better than or competitive with (supervised) HCIC and CACNet. GenCrop-6 is the best in five of six subjects and overall, suggesting that more training data is helpful. This can be seen on the horse class (which has the fewest training images), where the specialized GenCrop makes $2\times$ as many errors as HCIC, but GenCrop-6 makes only $0.6\times$ as many.

By violation type, all three versions of GenCrop do well at preserving the subject (V1) and managing negative space (V3). However, HCIC and CACNet are better at preserving context (V2) and managing edge clutter (V4).

5.3 Extension: Conditioning Signal

A common limitation of methods that directly regress a crop is that they lack user control: one cannot specify an aspect ratio or tightness. We propose a small extension of GenCrop that is conditioned on the above properties and hypothesize that, given a large number of training images, GenCrop-C can disentangle aspect ratio and tightness (area) while still learning to predict good crops. To achieve this, we swap the transformer-encoder in GenCrop with a transformer-decoder (Vaswani et al. 2023) and encode the conditioning with a feed-forward network (details in supplemental §C.3). Fig. 6 shows examples with different conditioning applied. While GenCrop-C does not enforce exact adherence to the signal, it does generally respect orientation (portrait vs. landscape), and changing the signal varies the crops (in aspect and tightness). We anticipate that future work may improve adherence or could learn to condition on other interesting properties from the data (e.g., composition patterns).

5.4 Additional Results, Ablations & Images

Please refer to supplemental §A.1 – §A.9 for additional analysis (e.g., computational cost, different cropping model architectures, results on cropping generic images, ranking metrics), ablations (e.g., subject-awareness, our data-quality filters, training dataset size), and example crops.

6 Discussion & Conclusion

We have demonstrated that it is possible to equal or surpass fully-supervised performance on subject-aware image cropping using a weakly-supervised approach that requires only stock photos and a pre-trained generative model.

GenCrop has its limitations. Training and evaluating on generic data (unknown or arbitrary subjects) is difficult because of the distribution mismatch between professional images in Unsplash and the cropping datasets created by crowd-sourcing (Chen et al. 2017a; Fang et al. 2014). Methods to calibrate Unsplash to these distributions could improve performance but are not a focus of this paper. GenCrop's pseudo-labels are sparse (1-per-image) and do not reproduce the dense crop-ranking annotations in GAICD, CPC, and SACD. Other composition issues such as parallax and occlusions cannot be fixed by cropping alone. Generative methods show promise for more advanced tasks such as searching for good perspectives in a NeRF (Martin-Brualla et al. 2021) or a 3D-transformed photo (Niklaus et al. 2019).

Pairing inputs to labels is a common approach to learning, and obtaining both often requires an expensive annotation process. Weakly-supervised learning typically relies on lower-quality-but-cheap labels to generalize and, by itself, is often not competitive with supervised approaches — e.g., on ImageNet classification (Deng et al. 2009). We have shown an instance where generative foundation models (Rombach et al. 2022) can invert this norm and convert plentiful, expert *labels* (i.e., finished products) into otherwise difficult-to-obtain *inputs*. This has implications for other learning problems where obtaining complete training data is hard. As the capabilities of image and text foundation models advance, we anticipate this paradigm will become a viable data creation strategy for other applications as well.

Acknowledgements

This work is supported by gifts from Adobe and Meta as well as computing support from the Stanford Institute for Human-Centered Artificial Intelligence (HAI). We thank Unsplash for providing access to their stock image dataset and the photographers who contributed their work to the Unsplash platform. We also thank the anonymous reviewers for their helpful comments and feedback.

Ethics Statement

All of the images needed to reproduce this paper are publicly accessible on the Unsplash platform, under the terms of the Unsplash license (https://unsplash.com/license). The images are designated by Unsplash for AI and academic use (https://unsplash.com/data). Please refer to Unsplash for the full terms and conditions and access to image data.

There is growing ethical discussion around AI trained on public image data. We use Unsplash images because they are highly moderated by Unsplash for artistic merit; inappropriate and offensive material; and data provenance. Our methods are not limited to Unsplash images — other professional stock image platforms, such as Adobe Stock, could also provide high-quality images that are suitable for training GenCrop. We encourage users to responsibly source their stock images according to the license terms of the image provider.

References

Adobe. 2023. Adobe Firefly.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. InstructPix2Pix: Learning To Follow Image Editing Instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, J.; Bai, G.; Liang, S.; and Li, Z. 2016. Automatic Image Cropping : A Computational Complexity Study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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 *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Chen, Y.-L.; Klopp, J.; Sun, M.; Chien, S.-Y.; and Ma, K.-L. 2017b. Learning to compose with professional photographs on the web. In *Proceedings of the ACM international conference on Multimedia (MM)*.

Cheng, Y.-C.; Lin, C. H.; Lee, H.-Y.; Ren, J.; Tulyakov, S.; and Yang, M.-H. 2022. InOut: Diverse Image Outpainting via GAN Inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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 ACM international conference on Multimedia (MM)*.

He, K.; Gkioxari, G.; Dollar, P.; and Girshick, R. 2017. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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 (CVPR)*.

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 (CVPR)*.

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. *arXiv preprint arXiv:2304.02643*.

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, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Loshchilov, I.; and Hutter, F. 2017. SGDR: Stochastic Gradient Descent with Warm Restarts. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Martin-Brualla, R.; Radwan, N.; Sajjadi, M. S. M.; Barron, J. T.; Dosovitskiy, A.; and Duckworth, D. 2021. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Niklaus, S.; Mai, L.; Yang, J.; and Liu, F. 2019. 3D Ken Burns Effect from a Single Image. *ACM Transactions on Graphics*, 38(6): 184:1–184:15.

Northrup, T.; and Northrup, C. 2019. *Stunning Digital Photography*. Mason Press.

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 (ICCV)*.

Popular Photography. 2016. *The Complete Portrait Manual (Popular Photography): 200+ Tips and Techniques for Shooting Perfect Photos of People*. Weldon Owen.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv:2204.06125.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Santella, A.; Agrawala, M.; DeCarlo, D.; Salesin, D.; and Cohen, M. 2006. Gaze-based interaction for semi-automatic photo cropping. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*.

Sarıyıldız, M. B.; Alahari, K.; Larlus, D.; and Kalantidis, Y. 2023. Fake It Till You Make It: Learning Transferable Representations From Synthetic ImageNet Clones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Teterwak, P.; Sarna, A.; Krishnan, D.; Maschinot, A.; Belanger, D.; Liu, C.; and Freeman, W. T. 2019. Boundless: Generative Adversarial Networks for Image Extension. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Tian, Y.; Fan, L.; Isola, P.; Chang, H.; and Krishnan, D. 2023. StableRep: Synthetic Images from Text-to-Image Models Make Strong Visual Representation Learners. arXiv:2306.00984.

Ultralytics. 2023. YOLOv8.

Unsplash. 2023. Unsplash Dataset.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2023. Attention Is All You Need. arXiv:1706.03762.

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 (CVPR)*.

Wang, S.-Y.; Wang, O.; Zhang, R.; Owens, A.; and Efros, A. A. 2020. CNN-Generated Images Are Surprisingly Easy to Spot... for Now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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.

Yang, Z.; Dong, J.; Liu, P.; Yang, Y.; and Yan, S. 2019. Very Long Natural Scenery Image Prediction by Outpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

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 (CVPR)*.

Zhang, B.; Niu, L.; Zhao, X.; and Zhang, L. 2022. Human-centric Image Cropping with Partition-aware and Content-preserving Features. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

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.

Zhong, L.; Li, F.-H.; Huang, H.-Z.; Zhang, Y.; Lu, S.-P.; and Wang, J. 2021. Aesthetic-guided outward image cropping. *ACM Transactions on Graphics*, 40(6): 1–13.