Supplemental Material: 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

A Additional Results and Ablations

We include additional results omitted from the main paper. Fig. 1 and Fig. 2 show cropping results on images from Unsplash (Unsplash 2023) and FCDB (Chen et al. 2017a), respectively. Fig. 3 shows additional examples by GenCrop-C, our cropping model trained with area and aspect ratio conditioning. This rest of the section is organized as follows.

- §A.1 analyzes the cost of our image generation pipeline.
- §A.2 provides our full qualitative results on our five evaluation criteria introduced in §5.2 of the main text.
- §A.3 – §A.5 ablate subject awareness, data filtering, and training dataset size in order to better understand the performance of GenCrop.
- §A.6 and §A.7 modify our cropping model with additional inputs and different CNN architectures to investigate the extent that these reasonable deviations to model architecture affect performance.
- §A.8 compares GenCrop on generic images in FCDB (Chen et al. 2017a), without the human-centric focus and subject-awareness.
- §A.9 compares GenCrop on human-centric crop ranking in GAICD (Zeng et al. 2019).

We also provide examples of generated images from our pipeline. Figs. 5, 6, and 7 show outpainted images produced by our pipeline and used for training; images outpainted without text-conditioning; and images that are discarded by filtering, respectively. We hope that these additional results will provide useful baselines and commentary for future research on weakly-supervised image cropping.

A.1 Computational Cost of Image Generation

The most expensive stage of our pipeline is the generation of outpainted images. This is dominated by Stable Diffusion (Rombach et al. 2022) (approximately 2 seconds per image on a NVIDIA V100 GPU (Nvidia 2017) at 512×512 with 50 de-noising steps). The other pre-trained models that we use for instance segmentation, YOLOv8x (Ultralytics 2023), and image captioning, BLIP-2 (6.7B) (Li et al. 2023), take 26 and 380 ms per image, respectively. The overhead of our data quality filtering using the CNN and subject heuristic is negligible (2K images per second).

Our pipeline is trivially parallelizable across GPUs and machines, and costs approximately \$60 per 100K images, using spot VMs, making it both extremely fast and low-cost, even compared to crowd-sourcing annotations.

A.2 Full Qualitative Results

In the main results, we reported the number of cropping mistakes by method and category, aggregated by the five criteria: *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 unnecessary clutter around the edges?*
5. *lack balance (e.g., missing symmetries, rule-of-thirds)?*

Fig. 4 provides full definitions and examples of these criteria and their application to cropped images.

While these criteria still require an expert photographer to judge, they reduce the subjectivity of qualitative evaluation to key technical aspects of the image — which can be consistently be found in reference books (such as (Popular Photography 2016; Northrup and Northrup 2019)). This assessment was performed by one of the paper authors, a photography domain expert, on 600 unique images (100 per subject category; 3,600 crops from the six methods) in a blinded experiment, with all method names withheld and their ordering randomized.[1]

We report the number of violations by each criterion in Tab. 1. VFN (Chen et al. 2017b), the prior weakly-supervised method, performs very poorly in comparison to GenCrop and the supervised HCIC (Zhang et al. 2022) and CACNet (Hong et al. 2021) baselines; performance is especially bad on cropping through the subject (V1) and negative space (V3), often by a factor of 2-4× or more. We are unable to compare to SAC-Net (Yang et al. 2023) due to code not being available at time of submission. On all subjects except horses, GenCrop produced fewer unnatural cuts through the subject (V1) than HCIC, the second best. GenCrop also

[1] Prior work (Hong et al. 2021) also assessed the quality of cropped images using coarse ‘good’, ‘normal’, and ‘bad’ buckets, but because our subject-aware task definition is more restrictive, we can apply a more focused set of technical criteria from the photography literature.

	Method	Weak sup	V1	V2	V3	V4	V5
(a)	*Portrait (Human)*						
	VFN	✓	31	13	6	6	2
	CACNet		7	2	1	1	0
	HCIC (CPC)		5	3	0	1	1
	GenCrop	✓	2	4	0	5	2
	GenCrop-H	✓	(same as GenCrop)				
	GenCrop-6	✓	1	3	0	4	0
(b)	*Cat*						
	VFN	✓	26	6	9	1	2
	CACNet		6	1	3	0	1
	HCIC (CPC)		3	1	4	1	2
	GenCrop	✓	2	2	0	1	2
	GenCrop-H	✓	2	0	0	3	3
	GenCrop-6	✓	0	0	0	2	1
(c)	*Dog*						
	VFN	✓	28	19	15	5	8
	CACNet		4	1	7	2	1
	HCIC (CPC)		7	1	7	1	0
	GenCrop	✓	1	3	1	1	0
	GenCrop-H	✓	1	3	0	2	0
	GenCrop-6	✓	0	0	1	2	1
(d)	*Horse*						
	VFN	✓	40	8	14	0	9
	CACNet		13	4	5	1	6
	HCIC (CPC)		3	0	2	1	2
	GenCrop	✓	9	4	5	1	0
	GenCrop-H	✓	2	4	3	0	1
	GenCrop-6	✓	2	2	0	0	1
(e)	*Bird*						
	VFN	✓	46	8	13	1	7
	CACNet		10	1	5	2	1
	HCIC (CPC)		4	0	4	0	5
	GenCrop	✓	1	1	0	1	0
	GenCrop-H	✓	3	2	4	1	4
	GenCrop-6	✓	0	0	0	0	1
(f)	*Car*						
	VFN	✓	26	7	9	3	4
	CACNet		1	2	6	0	2
	HCIC (CPC)		0	0	0	0	3
	GenCrop	✓	0	3	0	0	0
	GenCrop-H	✓	0	2	2	1	0
	GenCrop-6	✓	2	3	1	2	2

Table 1: **Full qualitative results.** Number of images with quality violation (↓ is better) in 100 images sampled per class (a – f). Refer to Fig. 4 for detailed explanation of the evaluation criteria (V1 – V5). GenCrop is our method trained on generated data of the subject class; GenCrop-H is trained on generated data for humans; and GenCrop-6 is trained jointly on generated data from all six classes. Analysis of these results is provided in §A.2. These results are summarized in Tab. 4 and Tab. 5 of the main text.

is similar or better than HCIC at managing negative space (V3) in the human, cat, dog, bird, and car images. However, GenCrop falls behind HCIC on V2 and V4, relating to contextual objects and clutter on the edges. We believe that HCIC is able to better avoid these mistakes having learned on CPC (Wei et al. 2018) using their content-preservation scheme. By contrast, GenCrop regresses crops directly, and a crop with and without edge clutter can be very similar in L1 distance due to the difference being only a few pixels. While our subject boundary loss penalizes unnatural cuts through the subject, defining such a loss over generic backgrounds and context is more challenging. Horses are a challenging subject for GenCrop, the number of initial stock images is smaller by a factor of 3x (2.2K) than the next smallest category, cats (6.7K). In this case, GenCrop-H trained on human images (22× more data), performs better than GenCrop[2]. On the other hand, we note that GenCrop-H is worse than GenCrop targeted for birds, suggesting that given a sufficient domain gap (human vs. bird appearance) and more specialized training data, training on more (out-of-domain) images alone does not guarantee improved performance.

GenCrop-6 makes the fewest violations overall and suggests benefit from training on more diverse data. While GenCrop-H has already shown competitive results on subject-aware cropping problem posed by (Yang et al. 2023) (unconstrained by subject type), GenCrop-6 suggests that one could automatically construct a generated dataset to better match a more generic distribution of categories by enumerating a small set of object classes.

A.3 Ablation of Subject-Awareness

Tab. 2a shows the importance of subject-awareness on the subject-aware benchmarks: the human-centric images in FLMS (Fang et al. 2014) and FCDB (Chen et al. 2017a) used by (Zhang et al. 2022); the SACD (Yang et al. 2023) dataset; and our 1000 annotated human portraits (Portrait1K) from Unsplash (Unsplash 2023). We ablate the subject boundary loss, use of subject information (the mask and bounding box), and subject-focused dataset construction (by training on generic data outpainted from Unsplash images without filtering for humans). These ablations are cumulative; removing subject information also removes the subject boundary loss since its computation depends on the subject mask; using generic images (reflecting the full content distribution of Unsplash) means that the subject and its type, if any, are not known.

We find that removing the subject boundary loss and subject information leads to a drop in IoU and increase in Disp. The result is small (up to 0.02 IoU and 0.007 Disp), compared to the large up-to 0.16 IoU and 0.05 Disp advantage GenCrop has over the prior weakly-supervised method, VFN (Chen et al. 2017b). Training GenCrop on generic images (not restricted to humans and with no estimated subject) leads to a slightly larger loss of performance (up to 0.04 IoU and 0.01 Disp). This shows that our outpainting based method is still effective over VFN, but that selection of relevant data and some subject-awareness (e.g., using subject

[2] Horse images may still contain humans such as riders.

		FLMS+FCDB		SACD		Portrait1K	
	Method	IoU ↑	Disp ↓	IoU ↑	Disp ↓	IoU ↑	Disp ↓
	VFN (for reference)	0.5783	0.1114	0.6555	0.0775	0.6222	0.0948
	GenCrop	0.7334	0.0687	0.7301	0.0632	0.7501	0.0612
(a)	*Ablation of subject-awareness*						
	w/o subject boundary loss	0.7145	0.0724	0.7207	0.0653	0.7458	0.0615
	w/o subject information	0.7202	0.0714	0.7105	0.0695	0.7429	0.0628
	w/ generic images	0.6861	0.0812	0.7037	0.0716	0.7396	0.0643
(b)	*Ablation of data filtering*						
	w/o CNN-based filter	0.7112	0.0734	0.7181	0.0666	0.7409	0.0632
	w/o CNN & heuristic filter	0.6987	0.0765	0.7217	0.0656	0.7398	0.0640
(c)	*Ablation of training dataset size (number of stock images used for outpainting)*						
	w/ 10000 images (19.2 %)	0.7174	0.0719	0.7238	0.0658	0.7640	0.0570
	w/ 1000 images (1.9 %)	0.7007	0.0756	0.7215	0.0658	0.7441	0.0613
	w/ 100 images (0.2 %)	0.6865	0.0780	0.6918	0.0709	0.7143	0.0655
(d)	*Additional inputs to the cropping model*						
	w/ depth (Bhat et al. 2023)	0.7226	0.0708	0.7245	0.0649	0.7505	0.0610
	w/ edges (Canny 1986)	0.7205	0.0715	0.7225	0.0661	0.7477	0.0619
(e)	*Different CNN feature extractors*						
	ResNet-18	0.7181	0.0726	0.7237	0.0648	0.7408	0.0640
	VGG-16	0.7241	0.0706	0.7247	0.0635	0.7457	0.0623
	MobileNet-V2	0.7207	0.0710	0.7284	0.0633	0.7565	0.0592
(f)	*Different model architectures*						
	U-Net (see §C.3 for details)	0.7365	0.0683	0.7195	0.0680	0.7420	0.0652

Table 2: **Ablations and additional experiments.** 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). Portrait1K refers to our annotated human images from Unsplash (§B.2). We provide the results for GenCrop and VFN reported in the main paper for reference. Analysis of these results is provided in the corresponding sections of §A. At a high level, we find that: (a) Knowledge about the subject benefits subject-aware cropping. (b) Removing poorly outpainted images improves cropping model accuracy. (c) Reducing the number of stock images available for outpainting lowers performance, as there is less diversity in the dataset. (d) Additional inputs (such as depth or edge detections) to the cropping model do not provide clear benefits. (e) Different CNN feature extractors provide similar performance as the ResNet-50 used in our main results. (f) GenCrop is not bound to a particular model architecture. The value from GenCrop comes from dataset generation, and even a simpler U-Net baseline (implementation details in §C.3) can realize the benefits of our data generation approach.

Method	Trained on	Weak sup	IoU ↑	Disp ↓
A2RL (Li et al. 2018)	AVA		0.663	0.082
A3RL (Li et al. 2019)	AVA		0.696	0.077
VPN (Wei et al. 2018)	CPC		0.711	0.073
VEN (Wei et al. 2018)	CPC		0.735	0.072
ASM (Tu et al. 2020)	CPC		0.749	0.068
GAIC (Zeng et al. 2019)	GAICD		0.672	0.084
CGS (Li et al. 2020)	GAICD		0.685	0.079
TransView (Pan et al. 2021)	GAICD		0.685	0.080
(Wang et al. 2023)	GAICD		0.686	0.078
VFN (Chen et al. 2017b)	Unsplash	✓	0.450	0.147
GenCrop	Unsplash	✓	0.654	0.090
GenCrop (U-Net)	Unsplash	✓	0.670	0.086

Table 3: **Evaluation on generic images in FCDB** (Chen et al. 2017a). Apart from VFN (Chen et al. 2017b) which we train on Unsplash, the reported numbers are from the original papers and (Wang et al. 2023).

Method	Trained on	Weak sup	$\overline{SRCC}$ ↑	$\overline{Acc_5}$ ↑	$\overline{Acc_{10}}$ ↑
VFN	GAICD		*0.648	*41.3	*60.2
HCIC	GAICD		*0.795	*59.7	*77.0
VFN	Flickr	✓	*0.332	*10.1	*21.1
VFN	Unsplash	✓	0.203	13.3	19.1
GenCrop-R	Unsplash	✓	0.446	23.1	38.5

Table 4: **Comparison on the 50 human-centric test images in GAICD.** * indicates results reported by (Zhang et al. 2022); refer to their paper for the full table of baselines. As noted in supplemental §A.9, the sparse supervision generated by GenCrop is poorly calibrated to the ranking labels in GAICD. Our version of GenCrop (GenCrop-R), modified for ranking performs better than VFN, which also does not have access to score and rank labels (e.g., being trained on Flickr or Unsplash), but results are poor compared to models that directly train on GAICD.

Method	# Params	Inference time (ms)
HCIC (Zhang et al. 2022)	19.47M	*7.8
CACNet (Hong et al. 2021)	18.93M	3.4
GenCrop	24.93M	5.7

Table 5: **Comparison between models: GenCrop and the baselines.** Cropping model complexity and inference time are not a key priority of this paper, but we include these details for completeness. We measured inference time per image on a single NVIDIA RTX A5000 GPU (Nvidia 2021), except for HCIC (*) which is their reported inference time (128 FPS or 7.8 milliseconds), computed on a slightly more powerful RTX 3090. We were unable to reproduce HCIC's performance using the available code. GenCrop is slower than CACNet (Hong et al. 2021) by 1.7×, but faster than HCIC (Zhang et al. 2022) by 0.37× (despite a less powerful GPU). Both GenCrop and CACNet directly regress a crop, while HCIC ranks based on a set of candidates.

masks produced as a by-product of data filtering) is beneficial when the task is to crop images with a defined subject.

A.4 Ablation of Data Filtering

The quality of the outpainted dataset has a small impact on the final cropping performance ranging from 0.03 to 0.01 on IoU and 0.008 to 0.002 on Disp (Tab. 2b). As with subject-awareness, GenCrop outperforms VFN (Chen et al. 2017b) by a large margin even without filtering. The effect of filtering with the CNN classifier, $D_{quality}$, and the additional outpainted subject heuristic are similar; many of the images removed by the subject heuristic are also removed by the CNN classifier, and vice versa. For example, in Fig. 7a (columns 1, 2 and 4), a tiled or composite image is very likely to have additional instances of the subject class.

In the context of this work, the two data filtering steps, as well as our use of the image captioning model (Li et al. 2023), are implementation details for operating the current Stable Diffusion (Rombach et al. 2022) model. We expect that future text-to-image inpainting models will produce fewer composite images or images with redundant subjects and be more faithful to text and image conditioning. Recent works such as (Zhang and Agrawala 2023; Hertz et al. 2022; Sarukkai et al. 2023) have explored additional control for such text-to-image diffusion models and these approaches could eliminate the need for data filtering by preventing undesirable content from being generated.

A.5 Ablation of Training Data Size

We study the impact of the number of stock images needed to train GenCrop. In Tab. 2c, we vary the number of images, fixing the category to humans.

There is a steady drop off in performance on the human-centric images of FLMS + FCDB and on SACD as the number of starting images is reduced to 10K and then 1K. With only 100 images, the performance is similar to removing subject-awareness and training on the unfiltered images in Unsplash – compare to generic images, Tab. 2a. More image diversity is clearly beneficial, possibly due to the domain gap between training on Unsplash images and testing on these datasets.

The fall-off is less severe on Portrait1K, our 1000 labeled images from Unsplash. With 10K images, the IoU and Disp actually increase slightly, before falling off at 1000 images and fewer. The initial lift from reducing from 52K to 10K is due to our default hyperparameters and training schedule being more suited to a dataset approximately $\frac{1}{5}$ of the human dataset's size (similar in size to that of the cat, dog, bird, and car data); on Portrait1K, we observe a similar boost with the full 52K images when the training schedule is shortened by $\frac{4}{5}$. For consistency, we keep the same set of hyperparameters across subject classes when training GenCrop.

Training GenCrop on a dataset generated with only 100 images still outperforms VFN (Chen et al. 2017b) (trained with 52K images). This clearly demonstrates the value that dataset generation via outpainting provides.

A.6 Effect of Additional Input Modalities

Photographers take into account factors such as shape and perspective (of which distance is a property) when composing an image. We consider whether additional input representations such estimated depth (Bhat et al. 2023) and Canny detected edges (Canny 1986) that loosely approximate these factors can improve cropping performance. To test this, we concatenate these modalities as an additional input channel to the CNN feature extractor. Tab. 2d shows that directly concatenating these inputs does not provide consistent benefit on IoU and Disp metrics. Note that we do not explore larger modifications to the model architecture beyond concatenation nor do we change the other learning hyperparameters, since these architectural directions are orthogonal to the dataset generation focus of our paper.

Future works may try to incorporate these priors, and our experiment here is to inform that the naive solution does not provide obvious benefit.

A.7 Does Model Architecture Matter?

We test two variations of GenCrop's architecture to evaluate whether the CNN architecture used for feature extraction matters and to test a simpler model architecture also trained on GenCrop's generated data.

Prior works (Hong et al. 2021; Zeng et al. 2019; Zhang et al. 2022; Jia et al. 2022; Pan et al. 2021; Chen et al. 2017b) have used a variety of CNN architectures for feature extraction, including AlexNet (Krizhevsky, Sutskever, and Hinton 2012), VGG-16 (Simonyan and Zisserman 2015), and MobileNet-V2 (Sandler et al. 2018). The choice of CNN for feature extraction has a small impact on the final cropping performance (Tab. 2c), with VGG-16 being slightly worse on the three datasets and MobileNet-V2 being slightly better on Portrait1K, but worse on the others. Reducing the model from ResNet-50 (He et al. 2016) to ResNet-18 lowers performance.

We implemented a simpler baseline GenCrop (U-Net), which uses a U-net predict a binary mask, representing each pixel's inclusion or exclusion in the crop. At inference time, a threshold is applied and the bounding box of the largest connected component is predicted as the crop. (See §C.3 for details.) The GenCrop U-net model performs slightly better than the regular GenCrop on human-centric images in FLMS + FCDB but worse on SACD and Portrait1K, our labeled set of 1000 human images in Unsplash (§B.2).

Overall, in both experiments, the magnitude of variation is small compared to the lift over VFN (Chen et al. 2017b). This suggests that the specific cropping model architecture used for GenCrop is not as important as the generated data.

Note that we do not claim the CACNet (Hong et al. 2021) inspired architecture that we use in the main paper a key contribution of the paper. There is a large space of possible models (such as the U-net) that could have been used in the experimentation, with likely similar effects. However, we do observe that the specific design choices of our main approach are easily amenable to extension with the subject boundary loss (§3.2 in the main text) and conditional control (§5.3 in the main text).

A.8 Comparison on Generic Images

We focused on the subject-aware cropping task in the main text. GenCrop can also be used for 'generic' images, by omitting the subject mask and subject-type specific filtering during dataset generation. In this situation, we train on a pseudo-label distribution that directly reflects Unsplash (Unsplash 2023).

We show results in Tab. 3 for cropping on FCDB (Chen et al. 2017a). GenCrop performs significantly better than VFN (Chen et al. 2017b); approaches the performance of supervised models trained using GAICD (Zeng et al. 2019) and AVA (Murray, Marchesotti, and Perronnin 2012); but lags behind models trained on CPC (Wei et al. 2018). The simpler U-Net variant of GenCrop described in §A.7 performs slightly better on FCDB, possibly because it is less expressive and prone to over-fit to Unsplash.

The distribution mismatch between a stock image dataset and benchmark dataset, such as FCDB (Chen et al. 2017a), is a key confound. For example, artistic images such as textures, where the subject is a pattern, in Unsplash are unlikely to be helpful on a benchmark like FCDB. Likewise, for subsets of FCDB such as landscape images (where there is no clearly segmentable subject), the composition is often subpar compared to stock images even with an 'optimal' crop due to inattention to perspective when the image was captured. For these genres, we believe that helping casual photographers find better perspectives at capture time is a useful and that learning what makes an artistic or dramatic perspective for landscape from stock images is an interesting direction for future work.

A.9 Comparison on Crop Ranking Tasks

Models that are trained on densely annotated crop ranking datasets, e.g., GAICD (Zeng et al. 2019), are often evaluated using ranking metrics. In this formulation of the cropping problem, the training data includes multiple crops per image along with ranks and scores. This is in contrast to FCDB (Chen et al. 2017a) and GenCrop, where supervision is sparse and each image has only a single label.

We find that these sparse datasets are insufficient to reproduce the scores in GAICD. Ranking well involves calibration of various intermediate-quality crops to crop scores. Tab. 4 shows results by GenCrop-R, a naive implementation of crop ranking using our generated datasets (see §C.3 for details).

B Dataset Details

We provide additional details about the Unsplash dataset, our hand-labeled evaluation sets for cropping different subjects, and a comparison to existing image cropping datasets.

B.1 The Unsplash Dataset

The Unsplash dataset (Unsplash 2023) includes images and metadata about each image, including the photographer, popularity, and any collections the image is a part of. For each subject, we filter the dataset by collections; for example, for portraits we select any photo belonging to a collection about portraiture, people, etc.. To remove obscure images, we remove images with fewer than 1,000 views, as of April 2023. Afterwards, we use the object detector (Ultralytics 2023) to remove images that do not contain detectable instances of the subject class (using a score threshold of 0.5). Images that are discarded by this final step may be mistagged or present the subject in a way that confuses object detection (e.g., heavy occlusions, abstract styles). We also discard images that have too many possible subjects (greater than 5), have a subject that is too small (e.g., less than 0.1 of frame height), or too large (more than 0.8 of frame area). The reason for discarding images with very large subjects is because these are often artistic close-ups, like of hands. Because of the large number of human images available, we ran a standard 2D keypoint detector (Xu et al. 2022) on the detections and excluded images where the shoulders are not detected or the head is missing. This also helps to remove extreme close-ups and unusual posing; while helping users compose these types of images is interesting, cropping casual images to extreme close-ups is unlikely to produce pleasing results due to other factors such as perspective that cannot be fixed by cropping. Cars often appear in the background of images, so for the car class, we also discard images with objects of another COCO (Lin et al. 2014) class with larger area than the car.

We split the dataset by photographer ID to avoid similar images (potentially from the same photo-shoot) from crossing the training, validation, and testing splits. When training GenCrop-6 jointly on all six enumerated subject classes, we exclude from the training set any images that are in the test or validation splits of any of the six classes.

Fig. 9d shows example images from Unsplash.

B.2 Subject-Aware Evaluation Sets

For quantitative evaluation following existing protocols and metrics, we annotated crops in 1,900 images from the test splits. 1,000 of these images are for humans, since humans are the most important and ubiquitous subject. For brevity, we refer to this subset as Portrait1K. We annotate 200 images each for cats, dogs, horses, and birds, and 100 images

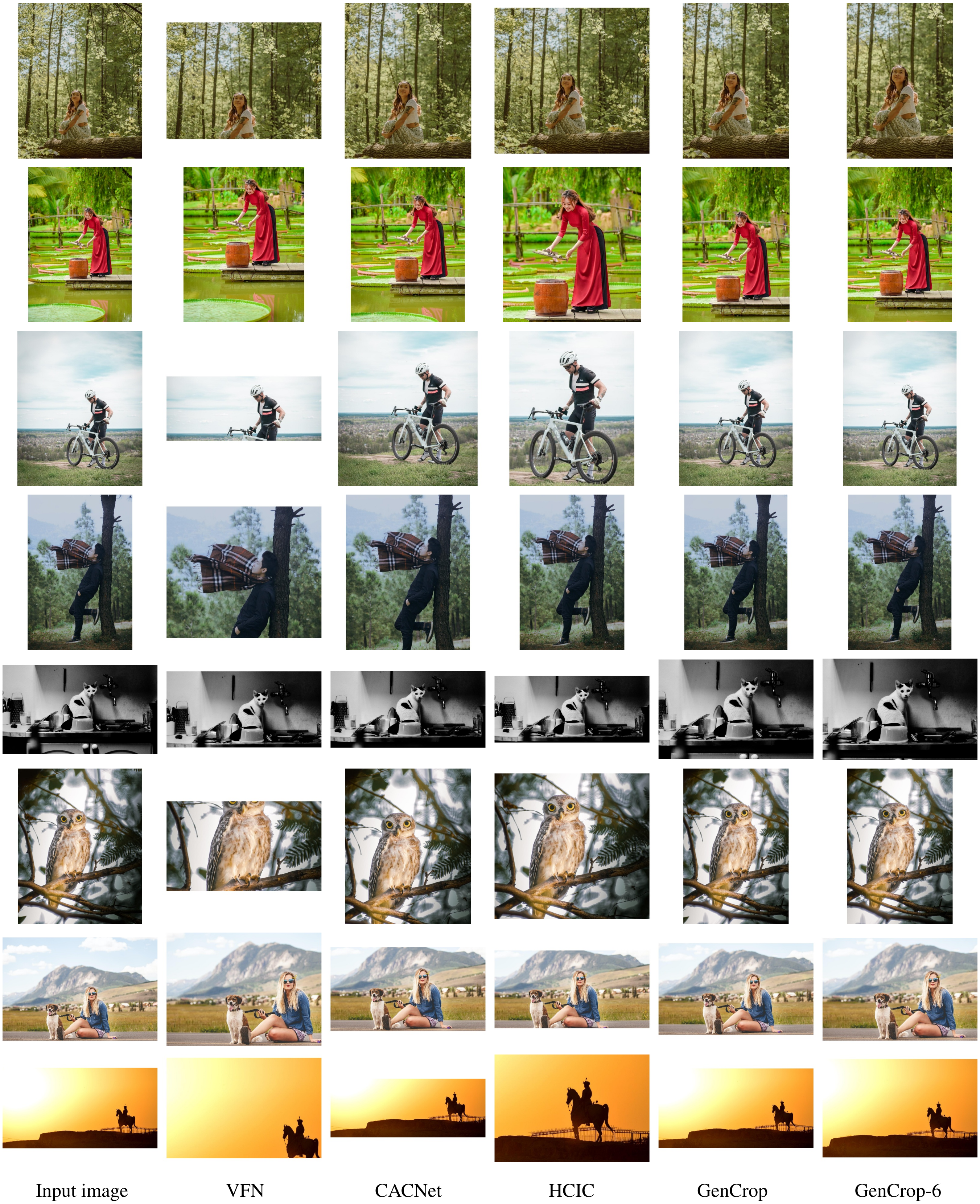


Figure 1: **Example crops on Unsplash.** (Unsplash 2023) GenCrop and GenCrop-6 are trained on generated images of the subject category and images of all six studied categories (listed in §4 of the main text), respectively. GenCrop produces similar quality results to (supervised) HCIC while VFN performs poorly.

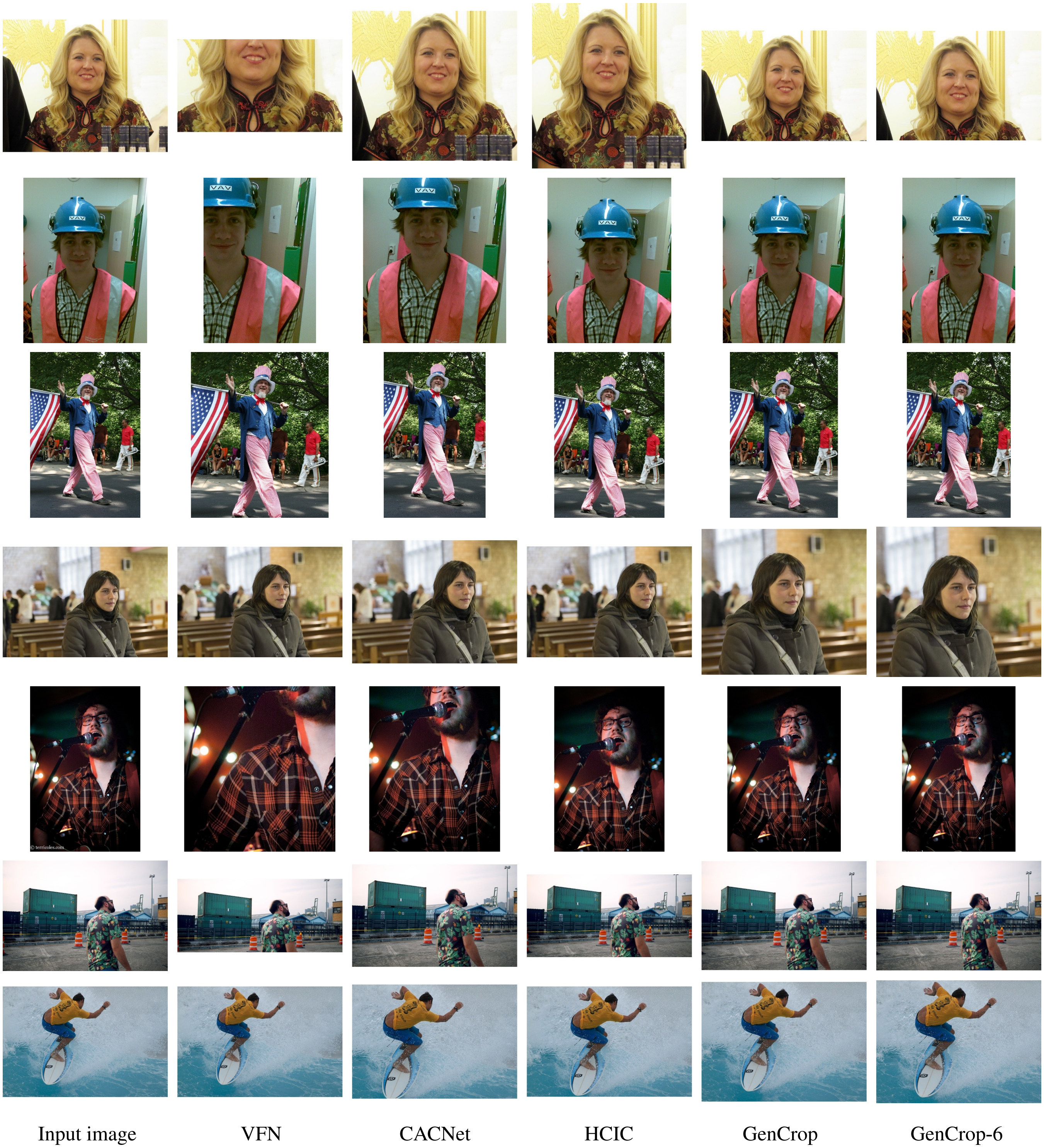
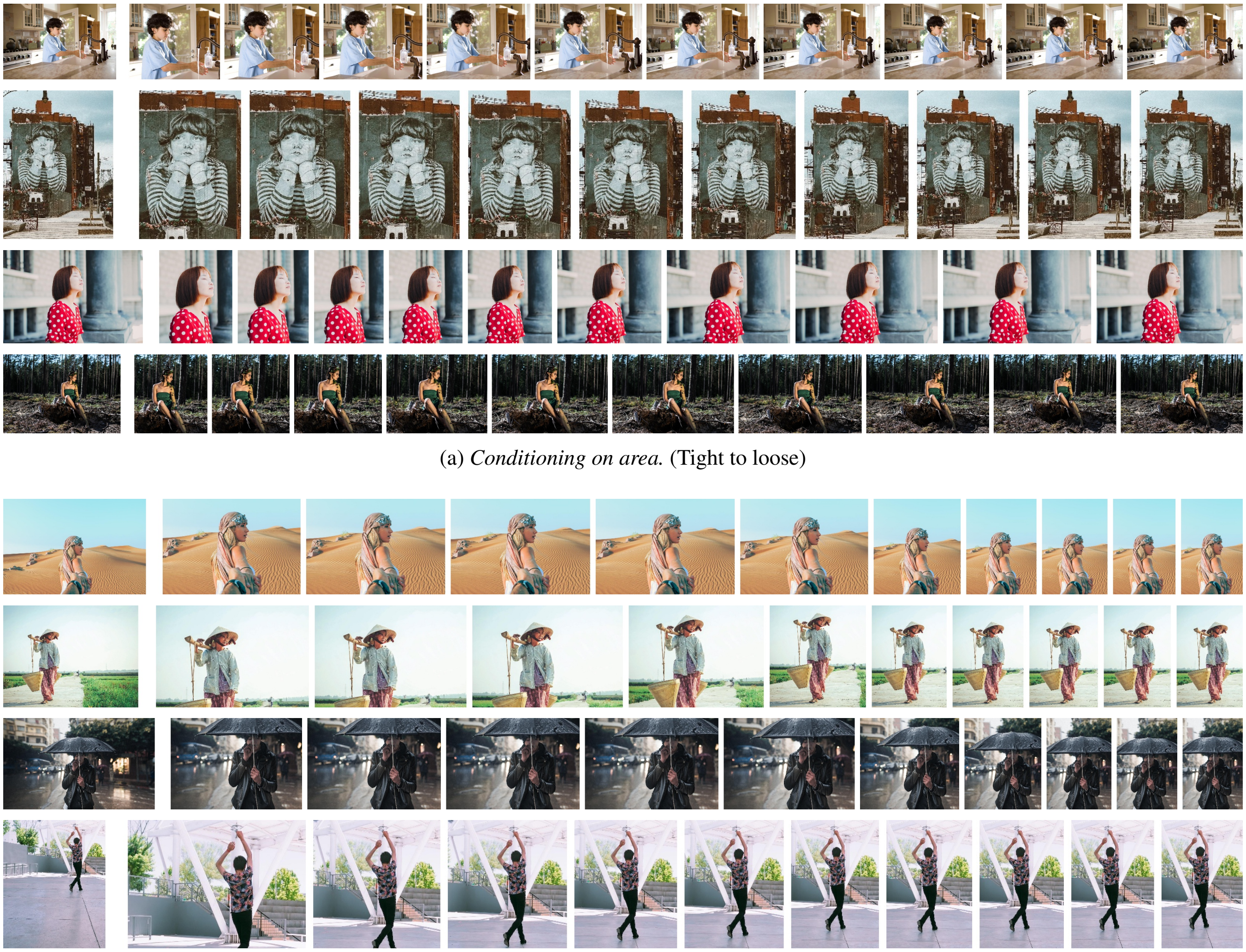


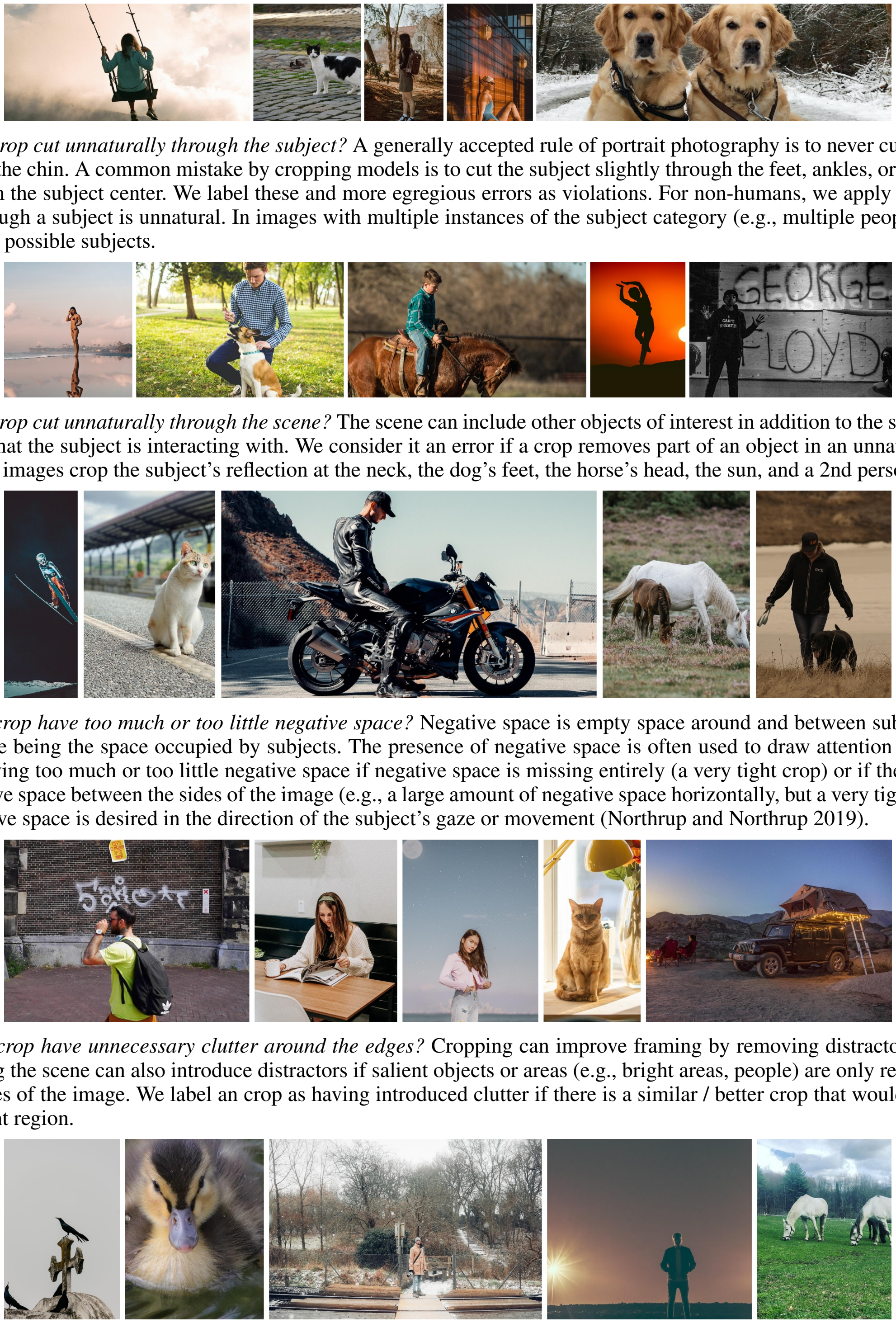
Figure 2: **Example crops for human-centric images in FCDB.** (Chen et al. 2017a) GenCrop and GenCrop-6 are trained on generated images of humans and images of all six studied categories (listed in §4 of the main text), respectively. GenCrop produces similar quality results to (supervised) HCIC while VFN performs poorly.



(a) *Conditioning on area.* (Tight to loose)

(b) *Conditioning on aspect ratio.* (Wide to tall)

Figure 3: **Additional examples of cropping with conditioning (GenCrop-C).** Original image on the left. By sweeping the value of the conditioning signal, (a) from 0.1 to 1 for area and (b) from 16:9 to 9:16 for aspect ratio, we are able to sample crops of different tightness. Note that extreme values can cause unnatural crops such as in the last row of (b), where the portrait orientation of the starting image provides limited room for a wide, landscape crop. GenCrop-C also does not enforce exact adherence to the conditioning at inference time; the conditioning is a continuous input signal akin to a hint to the model that is considered jointly with the image content.



(a) **V1:** *Does the crop cut unnaturally through the subject?* A generally accepted rule of portrait photography is to never cut a person through a joint or through the chin. A common mistake by cropping models is to cut the subject slightly through the feet, ankles, or hands as these are often furthest from the subject center. We label these and more egregious errors as violations. For non-humans, we apply similar criteria for whether a cut through a subject is unnatural. In images with multiple instances of the subject category (e.g., multiple people) we assess cuts through any of the possible subjects.

(b) **V2:** *Does the crop cut unnaturally through the scene?* The scene can include other objects of interest in addition to the subject. Sometimes these are objects that the subject is interacting with. We consider it an error if a crop removes part of an object in an unnatural or distracting way. The example images crop the subject's reflection at the neck, the dog's feet, the horse's head, the sun, and a 2nd person in bottom right.

(c) **V3:** *Does the crop have too much or too little negative space?* Negative space is empty space around and between subjects in an image, with positive space being the space occupied by subjects. The presence of negative space is often used to draw attention to the subject. We label a crop as having too much or too little negative space if negative space is missing entirely (a very tight crop) or if there are unbalanced amounts of negative space between the sides of the image (e.g., a large amount of negative space horizontally, but a very tight crop vertically). Often more negative space is desired in the direction of the subject's gaze or movement (Northrup and Northrup 2019).

(d) **V4:** *Does the crop have unnecessary clutter around the edges?* Cropping can improve framing by removing distractors from an image. However, cropping the scene can also introduce distractors if salient objects or areas (e.g., bright areas, people) are only removed partially or placed on the edges of the image. We label an crop as having introduced clutter if there is a similar / better crop that would have removed or included the salient region.

(e) **V5:** *Does the crop lack balance?* Positioning of areas of interest in the frame is an important part of composition. The Rule-of-Thirds is often promoted as an alternative to placing a subject directly centered. Other common techniques include looking for symmetries or lines that lead into the image or arranging regions of interest to balance the sides of the image. For example, a framing might be nearly symmetric but slightly off from the axis of symmetry, or the framing may place salient content too far toward the edges. In the example images, there are more balanced compositions that can be achieved by slightly adjusting the crop.

Figure 4: **Criteria for qualitative evaluation and example failures.** (The example crops are taken from the baselines and GenCrop.) We use these guidelines to control the subjectivity of qualitative evaluation and to provide insight on the types of cropping errors made by the various models. Note that the criteria are not mutually exclusive.

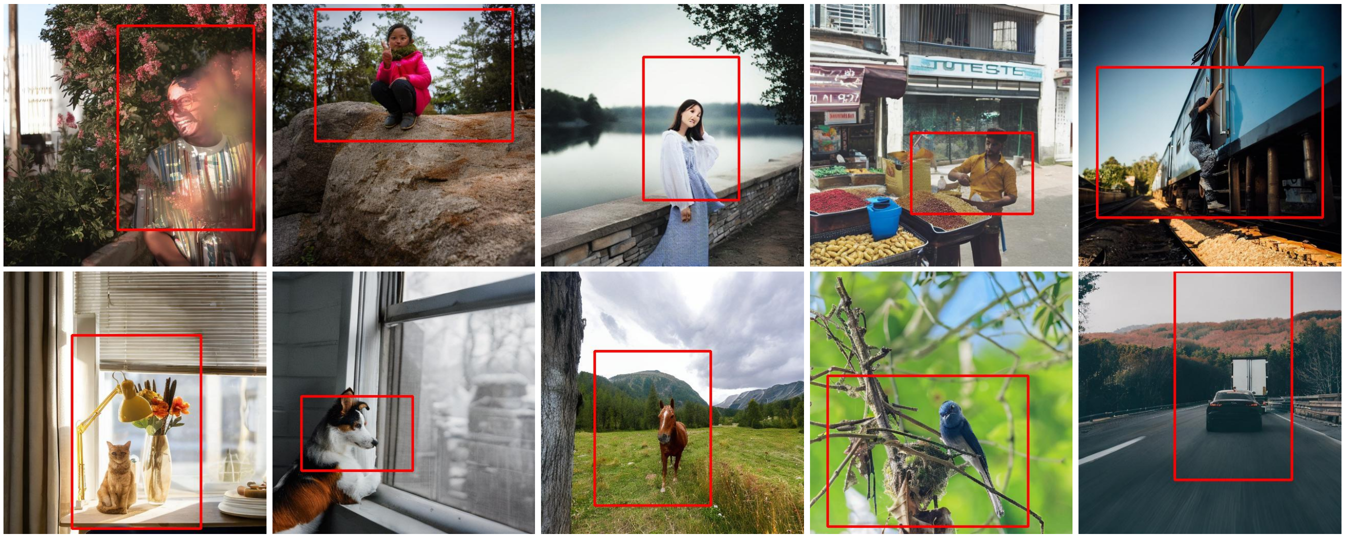
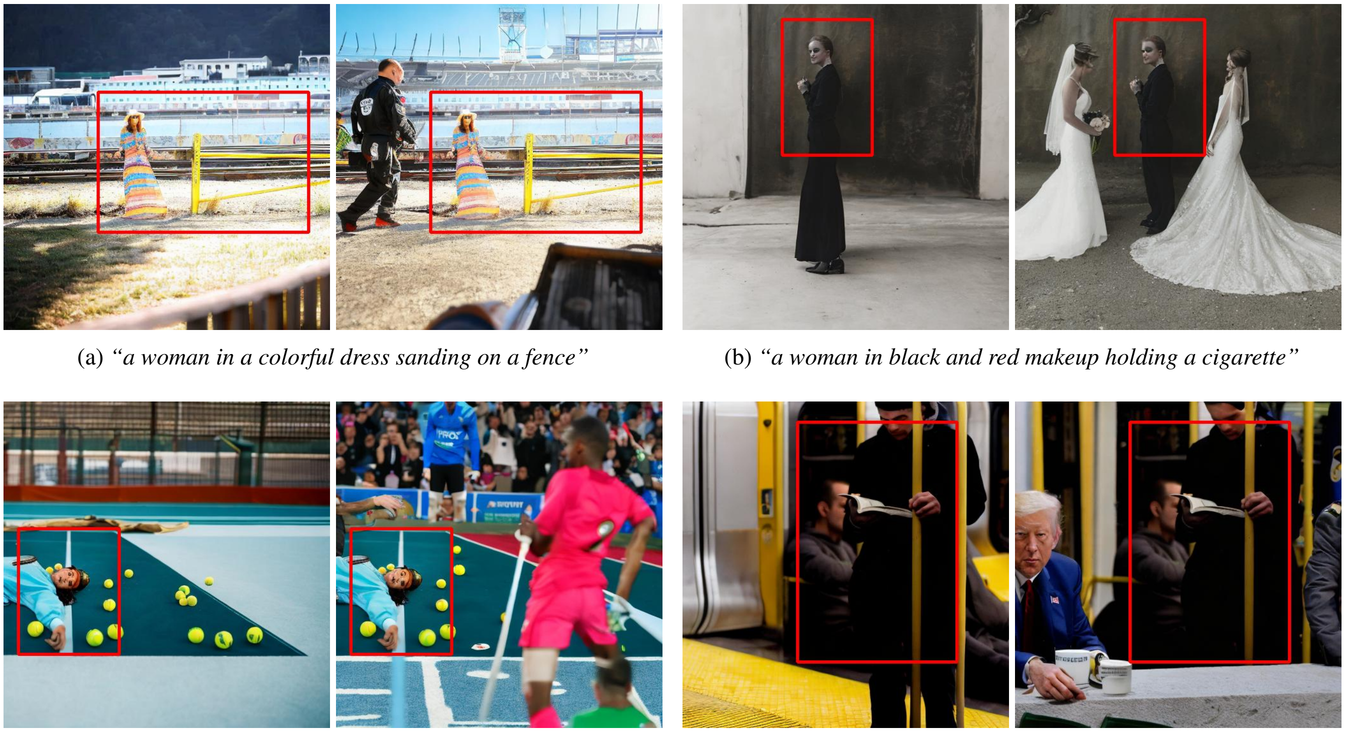


Figure 5: **Outpainted images used to train GenCrop.** A random sample. Red shows the original image. Stable Diffusion (Rombach et al. 2022) is able to produce a plausible, uncropped scene that a photographer might have seen. Contrast this with Fig. 7, which shows images that were discarded by our filtering and not used for training.



(a) *"a woman in a colorful dress sanding on a fence"*

(b) *"a woman in black and red makeup holding a cigarette"*

(c) *"a woman laying on a tennis court with tennis balls"*

(d) *"a man reading a book on a subway"*

Figure 6: **Examples of outpainting with (left) and without (right) text conditioning.** Conditioning with an estimated caption from BLIP-2 (Li et al. 2023) improves the quality of the synthesized region. Without text conditioning, Stable Diffusion (Rombach et al. 2022) may hallucinate arbitrary other objects and people into the outpainted regions. The text can be approximate since its purpose is to prevent arbitrary content in the hallucinated region and the subject region is already known.

for cars. The selected images have at least one foreground subject and have room to crop while preserving a good composition. While the images in Unsplash are already high-quality images, our goal is to test cropping models' abilities to find alternative framings, while controlling the other variables such as proper exposure, good subject selection, and subject posing that contribute to composition. Fig. 9e shows examples of our annotations.

B.3 Comparison to Existing Evaluation Datasets

Prior works are limited by the amount of data available to evaluate subject-aware cropping. This is a problem even for portraits: (Zhang et al. 2022) uses 215 images from FCDB (Chen et al. 2017a) and FLMS (Fang et al. 2014), and 50 images from GAICD (Zeng et al. 2019) which they identify as human-centric. The number of images for cats, dogs, etc. is even more limited.

There is a large domain gap between professional images from Unsplash and the annotations in FCDB (Chen et al. 2017a), FLMS (Fang et al. 2014), CPC (Wei et al. 2018), GAICD (Zeng et al. 2019), and SACD (Yang et al. 2023). This can be seen for example in the distribution of aspect ratios in Fig. 8. GenCrop trained on Unsplash is likely to produce crops that more closely reflect the Unsplash modes of 3:2 and 2:3 (41% and 25%). By contrast, a large number of annotations in the prior benchmarks are 16:9 or wider (e.g., 36% in FCDB compared to 6% in Unsplash). We also performed a close examination of these datasets (Fang et al. 2014; Chen et al. 2017a; Yang et al. 2023) and we found examples of poor-quality ground-truth annotations, which violate composition rules found in the photography literature. For example, while the annotators of these datasets may apply basic knowledge such as placing the subject according to the rule-of-thirds (Popular Photography 2016; Northrup and Northrup 2019), they sometimes do so in a way that is inattentive to the wider context of the image, by cropping through other subjects or salient content in the image. Fig. 10 shows examples of these awkward labels.

C Implementation Details

C.1 Dataset Generation

Sampling an outpainting region and mask. We use outpainting in our pipeline to create an un-cropped image given an input image. The outpainting mask, defining the region to be inpainted by the text-to-image diffusion model, needs to be automatically generated for GenCrop to scale.

There are many possible ways to sample a region to be outpainted (i.e., paste the input image into a square canvas). We using the following approach. Given an input image, we uniformly sample a desired area between 0.1 and 0.5 that the input image should occupy in the outpainted result. We downscale the image accordingly using bilinear interpolation and paste the input image randomly in the 512×512 canvas. This approach is unreliable for input images with very long or tall aspect ratios (e.g., 1:3), since they may exceed the canvas bounds even when resized. For these images, we fall back to resizing them such that their longest side fits in the canvas.

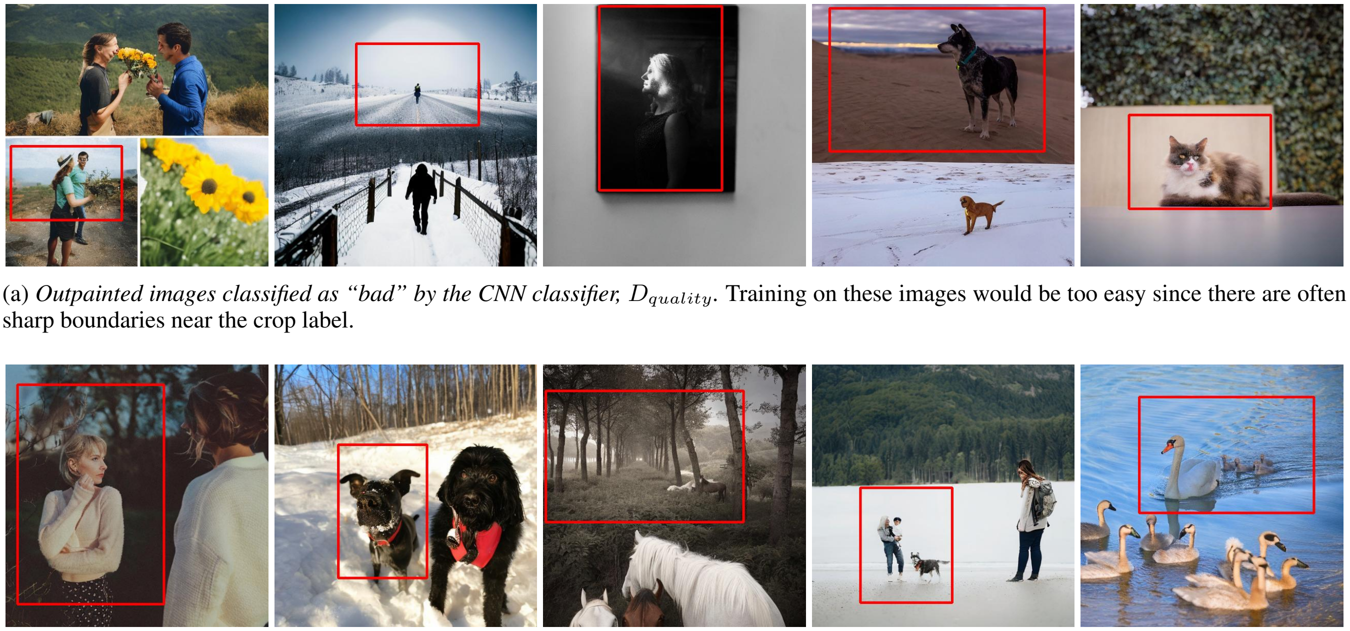
Stable Diffusion configuration. We use the Stable Diffusion V2 (Rombach et al. 2022) inpainting model, with guidance scale 4 and 50 denoising steps. The resolution is 512×512. In addition to the image caption from BLIP-2, we apply the following negative prompt: *"unrealistic, unnatural, collage, multiple images, ugly, deformed, disfigured, watermark, signature, picture-frame, image border, photo album, photo gallery"*. Despite aspects of the negative prompt referring to diffusion artifacts (such as faces and limbs) and the tiled/composite images, we find that such artifacts and behavior are not avoidable using negative prompting in the current Stable Diffusion models.

CNN-based quality filter. We anticipate that future pre-trained text-to-image models will produce fewer 'bad' images (tiled, composite, or bordered) images. Classifying the images (shown in Fig. 7a) is not a challenging vision task, however, since the patterns are visually distinctive. Our CNN-based quality filter, $D_{quality}$, serves as an example of how to do so using well-known computer vision techniques.

$D_{quality}$ is a standard ResNet-50 (He et al. 2016) trained for binary prediction. It is trained on 3,048 outpainted images from Unsplash, with the starting images sampled at random. Due to the low visual complexity of the task, a single annotator was able to label these images in under 2 hours ($<$ 1.6 seconds per image). 500 additional images are used for testing, to report accuracy statistics. Unlike the datasets used for training GenCrop, the data used to train $D_{quality}$ are generic since issues such as tiling and borders are unrelated to having a defined subject.

Low resolution is sufficient to detect composite, tiled, and bordered images. The input dimension to the CNN is $128 \times 128 \times 3$. We initialize the CNN with ImageNet (Deng et al. 2009) pre-trained weights and train for 100 epochs using AdamW (Loshchilov and Hutter 2019), with a base learning rate of 0.0001 and cosine learning rate annealing (Loshchilov and Hutter 2017). Batch size is 64. After training, the model has 93% accuracy overall and 0.79 precision and 0.74 recall for the 'bad' image category. The training time is 5 seconds per epoch on an NVIDIA RTX A5000 (Nvidia 2021).

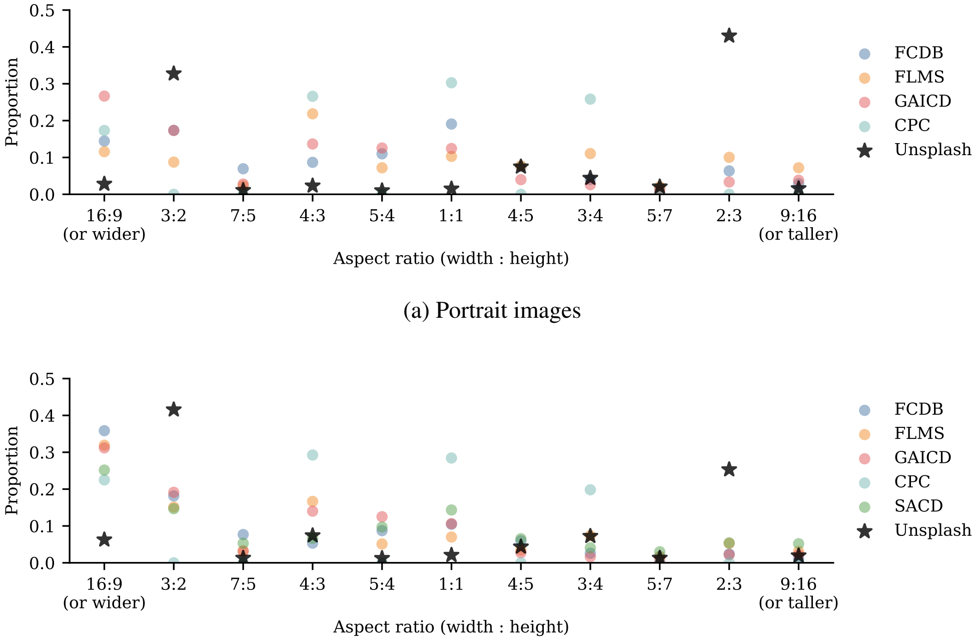
Sampling training pairs. The outpainted images produced by Stable Diffusion V2 are 512×512 squares. Since we wish for our cropping model to generalize to cropping images of other input dimensions that a user may supply, we obtain training pairs by sampling an enclosing view within the outpainted images. This requires that we sample an enclosing aspect ratio, a scale, and an (x, y) position. The aspect ratio is sampled from between 1:1 to 16:9 (long:short). With 20% probability, we choose an orientation different from the label; i.e., enclosing a landscape crop within a portrait image or vice versa. We sample a scale between 1× and 2×, multiplied against the longest side of the crop label. The (x, y) coordinates of the enclosing view are sampled using a piece-wise function: with 25% probability, we choose an edge of the crop label, and, with the remaining 75%, we sample uniformly. This is to prevent the model from learning that the edges of an image should always be removed, since it is unlikely that uniformly sampling an enclosing



(a) *Outpainted images classified as "bad" by the CNN classifier, $D_{quality}$.* Training on these images would be too easy since there are often sharp boundaries near the crop label.

(b) *Outpainted images removed by the subject heuristic.* Training on these images would be noisy since there is potentially another strong subject to compete with the original for interest.

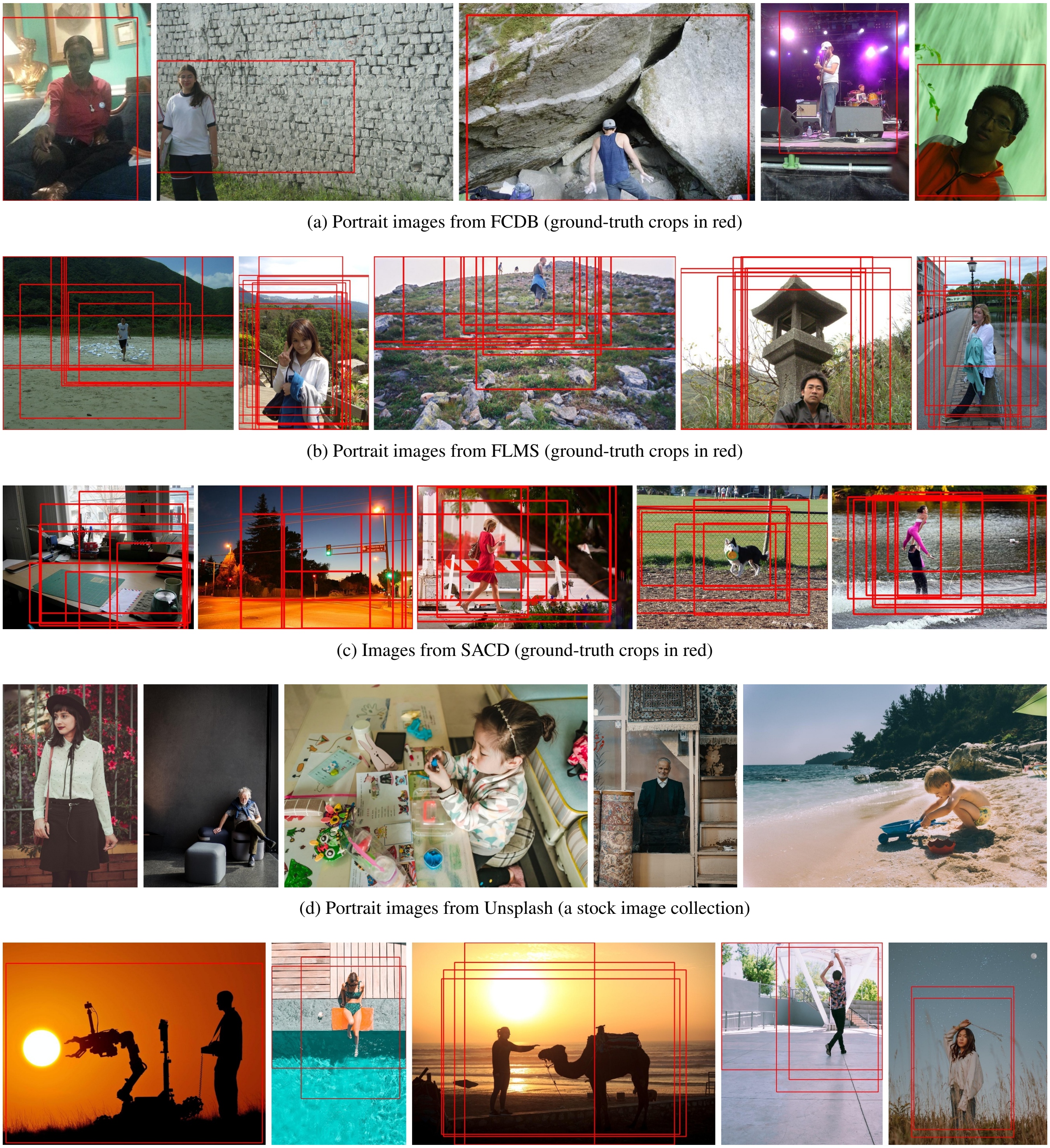
Figure 7: **Examples of rejected outpainted images.** Red shows the original image. (a) and (b) show images that are removed by the CNN quality model and subject heuristics, respectively. Detecting the most common failures to produce a seamless and realistic scene is not a challenging vision task; a combination of a binary CNN-based classifier and heuristics is effective.



(a) Portrait images

(b) All images

Figure 8: **Distribution of ground-truth crop aspect ratios in the existing cropping datasets and Unsplash.** For CPC and GAICD, which have multiple crops per image with scores, we include up to the 75th percentile of crops. Professional images in Unsplash (Unsplash 2023) (black stars) have a very different distribution than the ground-truth crops in FCDB (Chen et al. 2017a), FLMS (Fang et al. 2014), GAICD (Zeng et al. 2019), CPC (Wei et al. 2018), and SACD (Yang et al. 2023). The top aspect ratios in Unsplash are 3:2 and 2:3, while the other datasets have a much more varied distribution.



(a) Portrait images from FCDB (ground-truth crops in red)

(b) Portrait images from FLMS (ground-truth crops in red)

(c) Images from SACD (ground-truth crops in red)

(d) Portrait images from Unsplash (a stock image collection)

(e) Images and annotations (red) from our Portrait1K evaluation set.

Figure 9: **Example images from FCDB, FLMS, SACD, and Unsplash.** (a, b, c) Images from FCDB (Chen et al. 2017a), FLMS (Fang et al. 2014), and SACD (Yang et al. 2023), respectively. (d) Our goal is to learn properties of professional image framing from Unsplash (Unsplash 2023). (e) For quantitative evaluation, we annotated images from Unsplash that have room for cropping while retaining good composition (e.g., adhering to the quality criteria in Fig. 4).

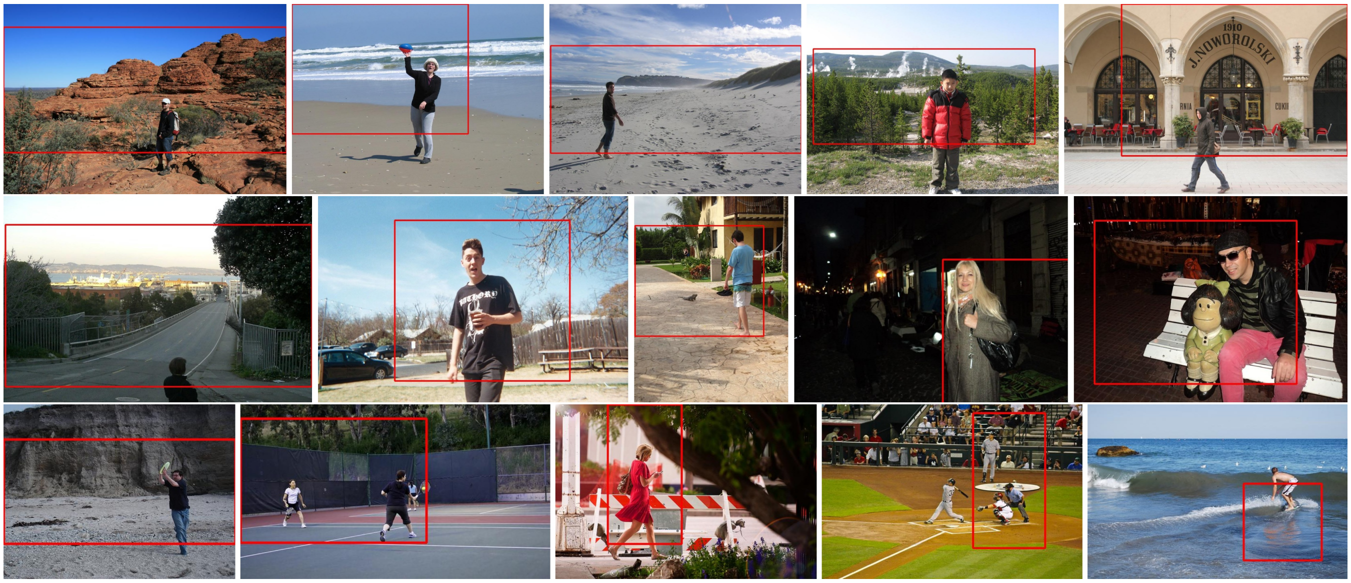


Figure 10: **Examples of awkward ground-truth crop labels in prior datasets.** The three rows show images from FLMS (Fang et al. 2014), FCDB (Chen et al. 2017a), and SACD (Yang et al. 2023), respectively. Red is a ground-truth label. While some of these crops follow basic composition rules such as the rule-of-thirds (Northrup and Northrup 2019), they have unclear aesthetic quality (often due to the low quality of the starting image) and avoidable mistakes on other (more subtle) technical criteria such as cropping people through their knees, ankles, and hands; having inappropriate amounts of negative space (i.e. space around the subject); and handling the scene in an awkward manner (third row: tennis players; baseball players; and surfer).

view alone will include the edges in the label. We implement the sampling steps listed above using rejection sampling.

C.2 Cropping Model and Training

Our default model architecture is based loosely on CAC-Net (Hong et al. 2021), retaining a similar two branch structure — 'cropping branch' and 'composition branch' — following CNN feature extraction. However, we modify the internal components and losses in order to utilize the training data produced by our pipeline. Since the focus of our work is on dataset generation, we leverage relevant modules from prior works (Hong et al. 2021; Zeng et al. 2019; He et al. 2017) and do not propose any novel architectural components.

Inputs. The inputs to our model are $256 \times 256 \times 4$ in dimension, representing the three RGB color channels and the subject mask ($\mathbf{x}_o$ concatenated with $\mathbf{m}_o$). If the image is not a square, then we pad the image using zeros to make it square. The RGB channels are normalized with the ImageNet (Deng et al. 2009) mean and standard deviation. The subject mask is a binary 0 or 1. During training we augment the RGB channels using random color jitter, gaussian blur, and grayscale conversion. We also apply random elastic distortion and horizontal flips to both the RGB and subject mask. A small amount of jitter is also introduced to the subject bounding box.

CNN feature extractor. Similar to prior work (Hong et al. 2021; Zhang et al. 2022), we extract multi-scale features from the input using a CNN. We use a ResNet-50 (He et al. 2016) with ImageNet (Deng et al. 2009) pre-trained weights, from the Pytorch Image Models (Wightman 2019) library. ResNet-50 has 5 stages; we take the features from the final three stages. Given 256×256 spatial dimension input, the features produced by the last three stages are $32 \times 32 \times 512$, $16 \times 16 \times 1024$, and $8 \times 8 \times 2048$. We downsample each of these features using a learned 1×1 convolution followed by bilinear interpolation to $16 \times 16 \times 256$. Then we sum the features and downsample them to $16 \times 16 \times 32$ using another learned 1×1 convolution followed by a ReLU activation.

Cropping branch. The cropping branch is a transformer-encoder (Vaswani et al. 2023). The 256 input tokens are the flattened $16 \times 16 \times 32$ grid of features from the CNN feature extractor. We use positional encoding to encode the spatial location of each token. The transformer-encoder has 8 attention heads and 2 layers. We apply a final linear layer to the transformer-encoder output to produce 256 crop proposals, which represent the offsets of a crop (similar to CAC-Net (Hong et al. 2021)).

Composition branch. The task of the 'composition branch' is to predict the relative weights of the crop proposals. In CACNet (Hong et al. 2021), it is trained on KUPCP (Lee et al. 2018) a composition classification dataset, hence the name.

To remain weakly-supervised, we do not use KUPCP for training. Instead, we train the branch to weight the crop proposals using only the regression losses described in §3.2 of the main text. To obtain a weight for a crop proposal, we use RoIAlign (He et al. 2017) and RoDAlign (Zeng et al. 2019)

— commonly used in the cropping literature (Zeng et al. 2019; Zhang et al. 2022; Yang et al. 2023). RoIAlign pools the features *inside* the crop proposal region (box), while RoDAlign pools the features *outside* the crop proposal region (box). The input to these two layers are the $16 \times 16 \times 32$ grid of features produced by the CNN backbone. RoIAlign and RoDAlign each produce $5 \times 5 \times 32$ features, which we concatenate and pool to a 128 dimensional vector using a single learned $5{\times}5$ convolution, without padding, followed by a ReLU activation. This is fed to a shallow feed-forward network with 128 hidden units and a ReLU activation to produce a scalar weight.

Only weights for anchor points that are within the subject bounding box are considered. For anchor points that are outside the subject bounding box, we set their weight to 0. (This also excludes anchor points that lie in padded regions.) Lastly, we compute a softmax over the weights to produce a distribution over the crop proposals, and the weighted sum of the proposals is the final crop prediction: $\hat{\mathbf{y}}$.

Hyperparameters and training. We train all parts of the network (CNN feature extractor, 'cropping branch', and 'composition branch') end-to-end. As mentioned in the main text, we use AdamW (Loshchilov and Hutter 2019) with a learning rate of 0.0001, cosine annealing (Loshchilov and Hutter 2017), and batch size of 32. The first 500 steps are warm-up. The network is trained for 50 epochs, with an epoch defined as one pass over the quantity of original Unsplash images — i.e., even if we generated five outpaintings per image, we sample only one outpainting per image per epoch. At the end of every epoch, we compute the loss on the validation set.

Inference. At inference time, we process the inputs in the same way as during training, with all data augmentations disabled. The output of our model is a 4 dimensional vector, $\hat{\mathbf{y}}$, representing the coordinates of the crop (i.e., x_1, y_1, x_2, y_2).

Model complexity and performance. Our model has 24.9 million parameters. During training, the model takes 0.43 seconds per batch on a single NVIDIA RTX A5000 GPU (Nvidia 2021). The sizes of the datasets used for training vary (see Tab. 1 in the main text). An epoch on the human (49K images after quality filtering) dataset takes approximately 11 minutes, while an epoch on the horse dataset (2.1K) takes 26 seconds. For consistency, we use the same hyper-parameters (learning rate, epochs, etc.) for all datasets. Inference with the model takes approximately 5.7 milliseconds per image on a single NVIDIA RTX A5000 GPU. Tab. 5 compares our model to the baselines.

C.3 Cropping Model Variants

In addition to the GenCrop model described in §C.2, we describe three architectural variations that are also trained on the same outpainted datasets. These are the conditional cropping model, GenCrop-C, used in the §5.3 of the main text; the U-net baseline, GenCrop (U-Net), used in §A.7; and the naive crop ranking method, GenCrop-R, used in §A.9.

Conditional model: GenCrop-C. GenCrop-C has the same architecture as GenCrop except that the transformer-encoder (Vaswani et al. 2023) is replaced with a transformer-decoder. The transformer-decoder receives as additional input the encoded conditioning signal. The encoding is performed by a 2-layer feed-forward network with 32 hidden units, ReLU activations, and dropout. At training time, the conditioning signal is the area and/or aspect ratio of the ground truth crop. We feed area as a single scalar value between 0 and 1. Aspect ratio is encoded as a 2-dimensional vector, where the first dimension is the aspect ratio (i.e., height divided by width) and the second is the inverse aspect ratio.

The area conditioned model receives only the area as additional input. Meanwhile, the aspect ratio conditioned model receives both the area and aspect ratio. We find that the aspect ratio conditioned model also requires area conditioning in order to produce a larger range of aspect ratios; for example, finding a wide 3:2 landscape crop in a tall portrait-oriented image.

At test time, the ground truth area and aspect ratio are not known. For the example images in Fig. 3, we sweep between 0.1 and 1 for area and 16:9 to 9:16 (holding area conditioning constant at 0.34).

One limitation of GenCrop-C is that the conditioning signal is not directly enforced. I.e., applying area conditioning of 0.1 will not produce to a crop with exact area of 0.1. However, as Fig. 3 shows, shrinking the area condition does lead to smaller crops. The model successfully learns a correlation between the conditioning and the outputs. We delegate further architectural improvements to future work.

U-Net Baseline: GenCrop (U-Net). As a simple baseline using GenCrop's dataset, we train a U-Net (Ronneberger, Fischer, and Brox 2015) to directly predict the crop mask. The architecture is based on ResNet-50 (He et al. 2016) and initialized with ImageNet (Deng et al. 2009) pre-trained weights. The model receives as input a $224{\times}224$ RGB image and subject mask. The output is a $224{\times}224$ mask prediction, where each pixel is a score for whether it is in the crop or not. The model is trained with a binary cross-entropy loss on every pixel. We train the model for 10 epochs, using AdamW (Loshchilov and Hutter 2019) with a learning rate of 0.0001, cosine annealing (Loshchilov and Hutter 2017), and batch size of 32.

At inference time, we apply a threshold of 0.5 to the predicted mask and select the largest connected component. We then crop the image to the bounding box of the connected component. Compared to GenCrop, the U-Net baseline is simpler, but lacks control over the crop boundary.

Ranking model: GenCrop-R. GenCrop-R is used only for the ranking-methods experiment in §A.9. We use a standard ResNet-50 (He et al. 2016), tasked with classifying whether a crop is real or a randomly sampled from the image. Like other variations of GenCrop, GenCrop-R's inputs are a $224{\times}224$ RGB image and subject mask. We initialize the model with ImageNet (Deng et al. 2009) pre-trained weights and optimize with binary cross-entropy loss. We train the model for 10 epochs, using AdamW (Loshchilov and Hutter 2019) with a learning rate of 0.0001, cosine annealing (Loshchilov and Hutter 2017), and batch size of 32. At inference time, we generate a grid of crop candidates (Zeng et al. 2019) and compute the binary class prediction scores.

The pseudo-labels used to train GenCrop-R are binary and therefore lack ranking information, such as for intermediate quality crops. As a result, performance on the GAICD (Zeng et al. 2019) test set is poor compared to alternatives that train with direct supervision from GAICD (see Tab. 4).

C.4 Baseline Cropping Methods

We compare GenCrop primarily to two supervised methods, HCIC (Zhang et al. 2022) and CACNet (Hong et al. 2021). While numerous supervised image cropping methods exist, few have released code and models. We train HCIC using their public code, on both GAICD (Zeng et al. 2019) and CPC (Wei et al. 2018), using their default hyper-parameters. We use the epochs with the best test SRCC for GAICD trained models and best test IoU on FCDB (Chen et al. 2017a) for CPC trained models. While HCIC's main contribution is the human-centric image cropping task, its supplemental materials show that it is also at or near state-of-the-art on generic images (Zhang et al. 2022), due to training on all of CPC. For CACNet (Hong et al. 2021), we use an unofficial implementation and model weights on GitHub since that is the only one available.

We implemented VFN (Chen et al. 2017b) following the example in its official repository. Since GenCrop receives subject information in the form of a concatenated mask, we also provide the subject mask to VFN. In our training, we follow the ranking-pair mining procedure described by the VFN paper. The VFN paper refers to their approach as unsupervised, but the approach can more accurately be described as weakly-supervised since, like GenCrop, the goal of their approach is to learn from professional, high-quality images. We note that HCIC (Zhang et al. 2022) also trains VFN on CPC and GAICD.

References

Bhat, S. F.; Birkl, R.; Wofk, D.; Wonka, P.; and Müller, M. 2023. ZoeDepth: Zero-shot Transfer by Combining Relative and Metric Depth. arXiv:2302.12288.

Canny, J. 1986. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6): 679–698.

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)*.

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)*.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-Prompt Image Editing with Cross Attention Control. arXiv:2208.01626.

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)*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Proceedings of Advances in Neural Information Processing Systems*.

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

Li, D.; Wu, H.; Zhang, J.; and Huang, K. 2018. A2-RL: Aesthetics Aware Reinforcement Learning for Image Cropping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, D.; Wu, H.; Zhang, J.; and Huang, K. 2019. Fast A3RL: Aesthetics-Aware Adversarial Reinforcement Learning for Image Cropping. *IEEE Transactions on Image Processing*, 28(10): 5105–5120.

Li, D.; Zhang, J.; Huang, K.; and Yang, M.-H. 2020. Composing Good Shots by Exploiting Mutual Relations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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)*.

Murray, N.; Marchesotti, L.; and Perronnin, F. 2012. AVA: A large-scale database for aesthetic visual analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Nvidia. 2017. Nvidia Tesla V100 GPU Accelerator.

Nvidia. 2021. Nvidia RTX A5000 Data Sheet.

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.

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)*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sarukkai, V.; Li, L.; Ma, A.; Ré, C.; and Fatahalian, K. 2023. Collage Diffusion. arXiv:2303.00262.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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

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)*.

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)*.

Wightman, R. 2019. PyTorch Image Models. https://github.com/rwightman/pytorch-image-models.

Xu, Y.; Zhang, J.; Zhang, Q.; and Tao, D. 2022. ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*.

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

Zeng, H.; Li, L.; Cao, Z.; and Zhang, L. 2019. Reliable and Efficient Image Cropping: A Grid Anchor Based Approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (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, L.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. arXiv:2302.05543.