HandsOff: Labeled Dataset Generation With No Additional Human Annotations

Austin Xu[*]
Georgia Institute of Technology

Mariya I. Vasileva
Amazon AWS

Achal Dave[†]
Toyota Research Institute

Arjun Seshadri
Amazon Style

Abstract

Recent work leverages the expressive power of generative adversarial networks (GANs) to generate labeled synthetic datasets. These dataset generation methods often require new annotations of synthetic images, which forces practitioners to seek out annotators, curate a set of synthetic images, and ensure the quality of generated labels. We introduce the HandsOff framework, a technique capable of producing an unlimited number of synthetic images and corresponding labels after being trained on less than 50 pre-existing labeled images. Our framework avoids the practical drawbacks of prior work by unifying the field of GAN inversion with dataset generation. We generate datasets with rich pixel-wise labels in multiple challenging domains such as faces, cars, full-body human poses, and urban driving scenes. Our method achieves state-of-the-art performance in semantic segmentation, keypoint detection, and depth estimation compared to prior dataset generation approaches and transfer learning baselines. We additionally showcase its ability to address broad challenges in model development which stem from fixed, hand-annotated datasets, such as the long-tail problem in semantic segmentation. Project page: austinxu87.github.io/handsoff.

1. Introduction

The strong empirical performance of machine learning (ML) models has been enabled, in large part, by vast quantities of labeled data. The traditional machine learning paradigm, where models are trained with large amounts of *human labeled* data, is typically bottlenecked by the significant monetary, time, and infrastructure investments needed to obtain said labels. This problem is further exacerbated when the data itself is difficult to collect. For example, collecting images of urban driving scenes requires physical car infrastructure, human drivers, and compliance with relevant government regulations. Finally, collecting real labeled data can often lead to imbalanced datasets that are unrepresentative of the overall data distribution. For example, in *long-tail settings*, the data used to train a model often does not contain rare, yet crucial edge cases [39].

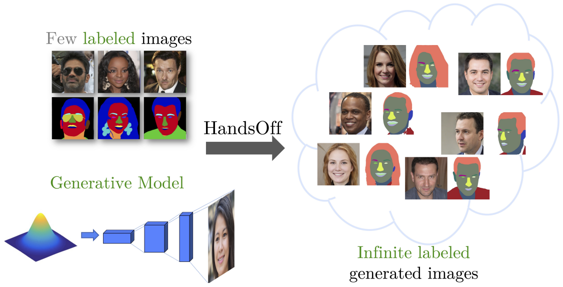


Figure 1. The HandsOff framework uses a small number of existing labeled images and a generative model to produce **infinitely** many labeled images.

These limitations make collecting ever increasing amounts of hand labeled data unsustainable. We advocate for a shift away from the standard paradigm towards a world where training data comes from an *infinite collection* of automatically generated labeled images. Such a dataset generation approach can allow ML practitioners to *synthesize* datasets in a *controlled* manner, unlocking new model development paradigms such as controlling the quality of generated labels and mitigating the long-tail problem.

In this work, we propose HandsOff, a generative adversarial network (GAN) based dataset generation framework. HandsOff is trained on a small number of *existing* labeled images and capable of producing an infinite set of synthetic images with corresponding labels (Fig. 1). To do so, we unify concepts from two disparate fields: dataset generation and GAN inversion. While the former channels the expressive power of GANs to dream new ideas in the form of images, the latter connects those dreams to the knowledge captured in annotations. In this way, our work brings together what it means to dream and what it means to know. Concretely, our paper makes the following contributions:

1. We propose a novel dataset generating framework, called HandsOff, which unifies the fields of dataset

[*] Work done as an intern at Amazon. axu@gatech.edu
[†] Work done while at Amazon

generation and GAN inversion. While prior methods for dataset generation [40] require new human annotations on synthetically generated images, HandsOff uses GAN inversion to train on existing labeled datasets, eliminating the need for human annotations. With ≤ 50 real labeled images, HandsOff is capable of producing high quality image-label pairs (Sec. 3).

2. We demonstrate the HandsOff framework's ability to generate semantic segmentation masks, keypoint heatmaps, and depth maps across several challenging domains (faces, cars, full body fashion poses, and urban driving scenes) by evaluating performance of a downstream task trained on our synthetic data (Sec. 4.2, 4.3, and 4.4).

3. We show that HandsOff is capable of mitigating the effects of the long-tail in semantic segmentation tasks. By modifying the distribution of the training data, HandsOff is capable of producing datasets that, when used to train a downstream task, dramatically improve performance in detecting long-tail parts (Sec. 4.5).

2. Related work

Our work is built on GANs [14], which consist of a generator that synthesizes new images, and a discriminator that discerns between real and generated images. Recent advances in GANs [7, 17–21] have demonstrated an ability to generate highly realistic images in numerous domains. We utilize the popular StyleGAN2 architecture [21], which synthesizes images by passing randomly sampled inputs through a series of *style blocks*. Remarkably, StyleGAN2's $\mathcal{W}$ and $\mathcal{W}+$ latent spaces form rich representations of images in a disentangled manner [1, 2, 33, 37], which can be utilized to edit complex semantic attributes in generated images [4,5,15,26,29,33]. The ability to identify semantically meaningful parts of generated images in the latent representation suggests that it could be used to generate pixel-level labels. This capability, coupled with GANs' ability to generate troves of high quality images, serves as the basis for generating synthetic image *datasets* [3, 24, 25, 40].

We build upon DatasetGAN [40], which trains a label generator using representations of an image formed from the GAN latent code. DatasetGAN requires *human annotation of GAN generated images*, which burdens a practitioner to seek out annotations for every new domain of interest. In addition to labeling, users also must actively *curate* images to label to ensure diverse semantic feature coverage and avoid GAN created artifacts. Furthermore, should the labeling scheme change and render the original labels obsolete, then additional annotations are again required. Acquiring additional labels is especially contrived when a large of number of quality human annotated images already exist. A framework that leverages these *real* preexisting labeled images would circumvent all of these drawbacks. EditGAN [26], a follow-on contribution to DatasetGAN, utilizes encoder-based reconstructions to perform image editing. BigDatasetGAN [24] exploits the pre-trained encoder of VQGAN [11] to utilize existing labeled *synthetic* images. In contrast, our approach links latents of labeled *real* images to their labels by employing GAN inversion, the process of mapping a real image to the latent space of a GAN.

The myriad of inversion techniques range from encoder-based approaches [4, 31, 34, 35], which utilize trained encoders to map images directly to the latent space, to optimization-based approaches [1, 2, 9], which directly optimize a similarity loss (e.g., LPIPS [38]) to obtain latents. Some methods modify generator weights to increase image reconstruction quality [2, 5, 32]. Our work exclusively uses inversion methods that do not modify the generator, since the generator must remain unperturbed to generate new images from the original data distribution. We invert images to the $\mathcal{W}+$ space, which is more expressive than the $\mathcal{W}$ space and leads to higher quality reconstructions [37].

3. The HandsOff framework

The HandsOff framework, shown in Fig. 2, consists of three main components: (1) a generator (realized as a GAN), which maps a latent code $w \in \mathcal{W}$ to an image X, (2) an inverter, which maps an image X to a latent code w, and (3) a label generator, which maps a latent code w to a pixel-wise *label* Y, such as a semantic segmentation mask. HandsOff exploits the fact that the generator's latent space forms a rich, disentangled representation of images. Since these latent spaces already encode semantically meaningful concepts from images [1, 2, 33], we aim to train a 'label generator' that maps latents in this space to *labels*.

Unfortunately, training this label generator requires paired data of latents w with labels Y. One approach, espoused by prior work [40], could be to map the latent w to an image X, and ask annotators to manually label the image. However, in many applications, paired data of (X, Y) is readily available, thanks to the careful efforts of dataset collectors. Our key insight is that *existing labeled image datasets can be used to train a label generator on GAN latent spaces,* using techniques from the GAN inversion literature. Below, we describe our specific approach for GAN inversion (Sec. 3.1), our representation of the GAN's latent space (Sec. 3.2), and finally, our label generator (Sec. 3.3).

3.1. GAN inversion

The key step in the HandsOff framework is to connect advances in GAN inversion to dataset generation. GAN inversion allows us to use a small number of pre-existing labeled images to create a dataset of labeled *latents*. Our use of pre-existing labels allows practitioners to re-purpose existing labeled datasets, avoiding the cost of acquiring labels,

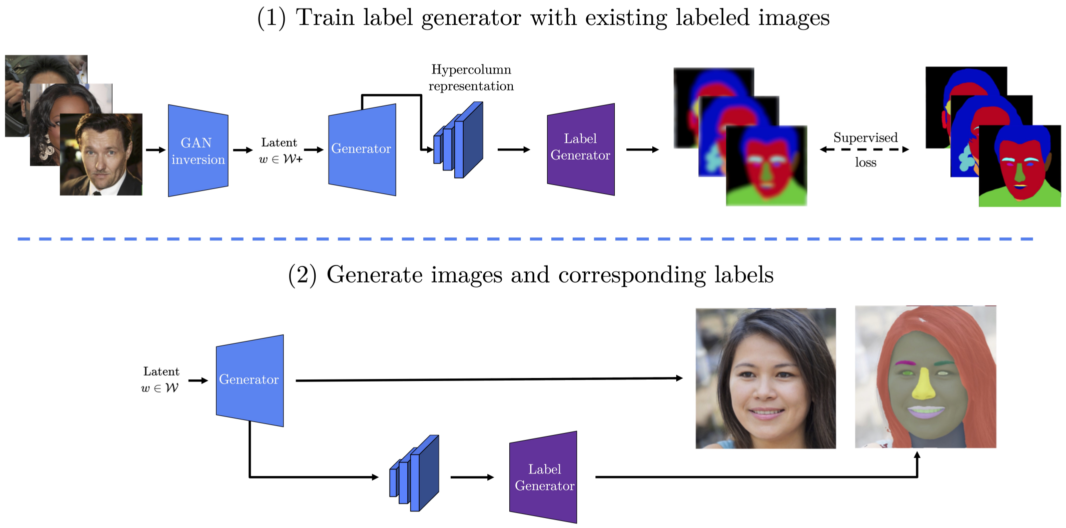


Figure 2. The HandsOff framework. (Top) GAN inversion is used to obtain training image latent codes, which are then used to form hypercolumn representations. The label generator is then trained with the hypercolumn representations and original labels. (Bottom) To generate datasets, the trained label generator is used in conjunction with a StyleGAN2 generator to produce image-label pairs.

including the prerequisite of maintaining annotation workstreams in their machine learning pipelines.

Our GAN inversion is inspired by popular approaches in the image-editing community [26, 41]. Given a pre-trained generator G, we first train an encoder to predict a latent $w^{(e)}$ from an input image X. In practice, this feed-forward encoder results in a good initial inversion of an image to a latent input. To refine this initial estimate further, we solve the following regularized optimization problem:

$$\min_{w:\|w-w^{(e)}\|_2^2\leq c_{reg}} \mathcal{L}_{LPIPS}(X, G(w))+\lambda_{\ell_2}\|X-G(w)\|_2^2$$

where $\mathcal{L}_{LPIPS}$ is the Learned Perceptual Image Patch Similarity (LPIPS) loss [38]. Although this problem is highly non-convex, in practice we find that using a fixed number of gradient descent iterations significantly refines the latent code. This refinement step requires additional inference time, but this additional cost is incurred only once on a small number of training images. In our experiments, we utilize ReStyle [4] as the encoder, but we emphasize that our framework is amenable to *any* GAN inversion procedure that does not modify the generator weights. Note that common approaches for GAN inversion fine-tune the *generator* in order to achieve a better *inversion* for a specific image [2, 5, 32]. To ensure our generator can produce *new* images from the task domain, we keep the generator parameters frozen throughout the inversion process.

3.2. Hypercolumn representation

GAN inversion allows us to map images X to latent codes w. We could use these latent codes directly to train a label generator that maps latent codes w to labels Y. However, this discards the rich representations encoded by the intermediate layers within the generator. Rather than training on w directly, we construct a hypercolumn representation $S^{\uparrow}$ from the generator's intermediate layers. Specifically, we use a StyleGAN2 generator, where the latent code w is used to modulate convolution weights in intermediate style blocks, which progressively grow an input to the final output image. For a 1024 × 1024 resolution image, there are $L = 18$ style blocks. We utilize the approach of [40] and take the intermediate output of these style blocks, upsample them channel-wise to the resolution of the full image, then concatentate each upsampled intermediate output channel-wise to obtain pixel-wise hypercolumns. Our final hypercolumn representation is denoted by $S^{\uparrow}$, with each pixel j now having a hypercolumn $S^{\uparrow}[j]$ of dimension C. Due to the high dimensionality of the hypercolumns ($C = 6080$ for 1024×1024 images), we cap the generated image resolution to 512 × 512, and downsample intermediate outputs from higher resolutions.

3.3. Label generator

The label generator exploits the semantically rich latent space of the generator to efficiently produce high quality labels for generated images. Because the latent codes already map to semantically meaningful parts of generated images, simple, efficient models suffice for generating labels. Specifically, like in [40], we utilize an ensemble of M multilayer perceptrons (MLPs). The MLPs operate on a pixel-level, mapping a pixel's hypercolumn to a label. To generate a label for a synthetic image, we pass the hyper-

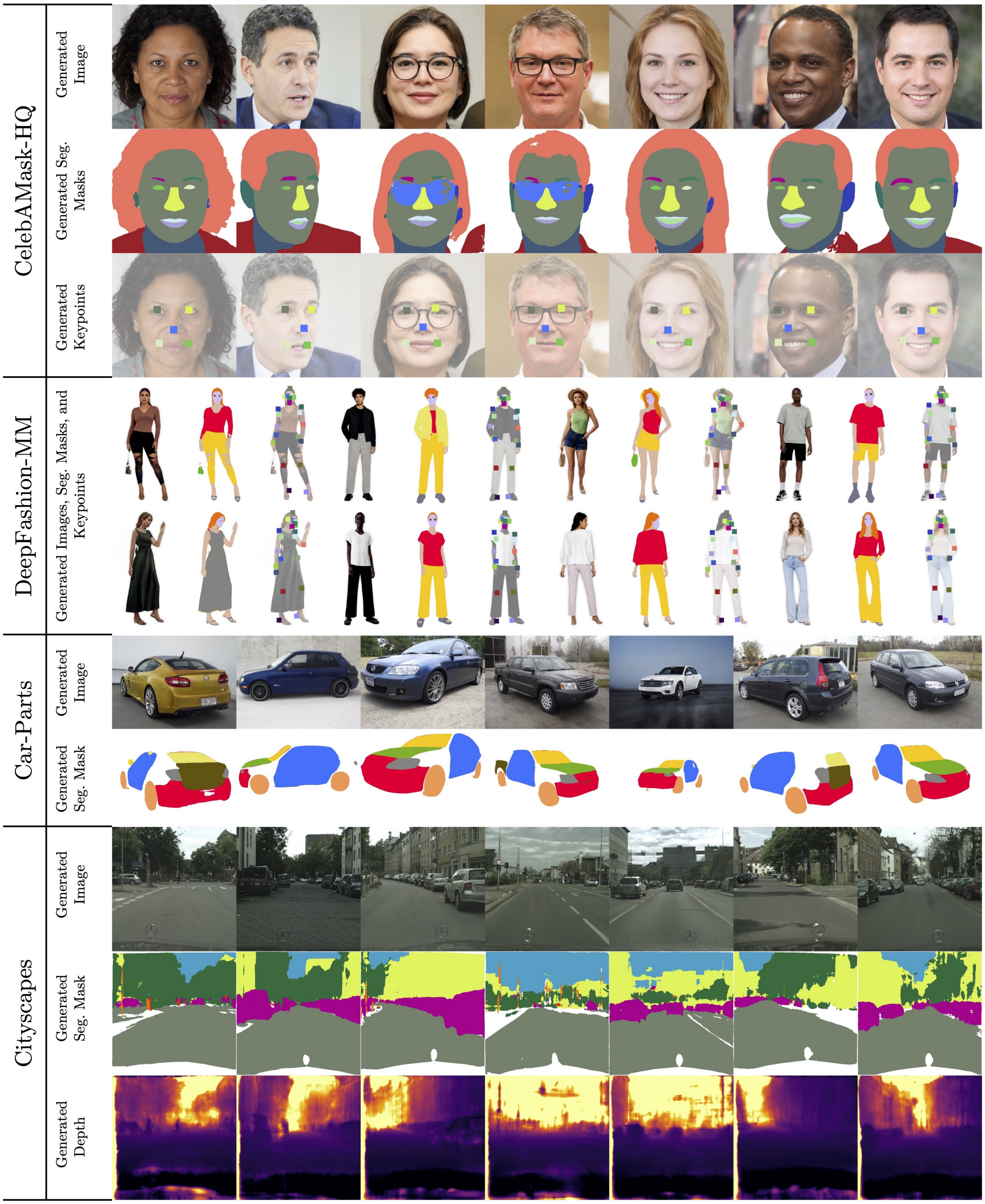


Figure 3. Examples of HandsOff generated labels (segmentation masks, keypoints, and depth) across four different domains. Generated labels capture fine details across various object orientations (CelebAMask-HQ, Car-Parts), object poses (DeepFashion-MM), and lighting conditions (Cityscapes). Note that HandsOff correctly assigns the label "skin" to the visible parts of the leg in the ripped areas of jeans (DeepFashion-MM, first row, first human) and correctly assigns the labels "jacket" and "shirt", despite the fact that the jacket and shirt are almost indistinguishable color-wise (DeepFashion-MM, first row, second human). Furthermore, generated keypoints are accurate despite partial occlusion, such as eyes behind glasses (CelebAMask-HQ, third and fourth image) or feet covered by long pants (DeepFashion-MM, second row, last human). HandsOff is also capable of identifying spatially small objects, such as street signs (Cityscapes, first, third, and fourth image).

column formed by latent code w through the M MLPs, and aggregate the outputs (via majority vote or averaging) to produce a label. The M MLPs are trained using a small number ($\sim$50) of pre-existing labeled images with a cross-entropy loss for generating discrete labels (e.g., segmentation masks) and mean-squared error loss for generating continuous labels (e.g., keypoint heatmaps).

Our use of an ensemble of MLPs naturally provides a way to filter out potentially poor labels by using the prediction uncertainty as a proxy for label quality. For discrete labels, we can utilize Jensen-Shannon divergence [6,22,28,40] across the M MLPs to produce pixel-wise uncertainty maps. For predicting continuous labels, we compute the pixel-wise variance across the MLP outputs. In both cases, the overall image uncertainty is computed by summing across all pixels.

4. Experimental results

We extensively evaluate HandsOff in generating both discrete (segmentation masks) and continuous (keypoint heatmaps and depth) labels across four challenging domains: Faces, Cars, Full-Body Human Poses, and Urban Driving Scenes. We utilize various pre-trained StyleGAN2 generators [12, 13, 21] and ReStyle inverters [4]. To train the label generator, we utilize existing labels from CelebAMask-HQ [23], Car-Parts [30], DeepFashion-MultiModal [16,27], and Cityscapes [10]. The key assumption of HandsOff is that GAN inverted image reconstructions align well with the original labels. We present visualizations of reconstructed image alignment in Appendix D.1. Our label generator architecture across all domains and tasks is an $M = 10$ ensemble of 2-hidden layer MLPs. This simple architecture is a distinct strength of the HandsOff framework: intensive parameter and architecture finetuning are not necessary to achieve state-of-the-art empirical performance. Please see appendix for additional results, ablations, and details about models, training, and datasets.

4.1. Experimental set-up

Downstream network In all domains and tasks, we utilize DeepLabV3 with a ResNet151 backbone as our downstream network. We generate 10,000 synthetic images and labels, filter out the top 10% most uncertain images (see Sec. 3.3), and train our downstream network for 20 epochs with the 9,000 remaining images. For segmentation, we have DeepLabV3 output a probability distribution over all of the parts for each pixel, whereas for keypoints or depth, we have DeepLabV3 output continuous values. Due to the dynamic nature of elements in the Cityscapes dataset, slight imperfections in the reconstructions uniquely affect segmentation mask alignment. To mitigate this, we perform an extra fine tuning step with the original 16 or 50 labeled examples used to train the label generator while training for semantic segmentation.

Baselines We compare HandsOff against three baselines: DatasetGAN, EditGAN, and Transfer Learning. We are only able to evaluate DatasetGAN in the face domain, as DatasetGAN is unable to accommodate the change in labeling scheme from their custom labeled car dataset to the larger Car-Parts-Segmentation dataset, thus highlighting another drawback of requiring GAN labeled images. For EditGAN, we adopt the image editing framework to synthesize labels for images. However, we are unable to test in the full-body human poses and urban driving scene domains, as EditGAN has only released checkpoints for the face and car domains. For the Transfer Learning baseline, we initialize DeepLabV3 with pretrained weights on ImageNet, then finetune the classification head of the model on the 16 or 50 labeled images used to train HandsOff until convergence. This baseline is used to benchmark our method, which is trained on up to 50 labeled images, against a model that is trained on 100,000+ *labeled* out-of-domain images in addition to the 16 or 50 labeled in-domain images.

Datasets For faces, we split CelebAMask-HQ into a set of 50 training, 450 validation, and 29,500 testing images. We collapse the 19 original segmentation classes into 8 and scale the keypoint locations in the low resolution version of images found in CelebA to the full resolution images. For cars, we retain the original 400 image train set, split the test set into a set of 20 images for validation and 80 images for testing, and collapse the 19 original classes into 10. For full-body human poses, we split DeepFashion-MultiModal into a set of 200 training, 500 validation, and 12,000 testing images. We collapse the 24 original segmentation classes into 8 and 10 classes and retain the original 21 labeled keypoint locations. For Cityscapes, because the ground truth test labels are not released, we split 300 and 1275 images from the original train set for validation and test, respectively. We utilize the eight groups (e.g., human, vehicle, etc) as our class labels. Note that while our train sets may contain more than 50 images, we use *at most* 50 labeled images from the train sets to train HandsOff in each domain.

4.2. HandsOff generated datasets

We visualize the generated image-label pairs from HandsOff in Fig. 3. HandsOff is capable of generating very high quality labels across all domains. In the face domain, HandsOff is capable of producing segmentation masks that can correctly distinguish left/right features like eyes or ears and identify rare occurring classes such as glasses. Furthermore, it produces extremely accurate keypoint locations even when such locations may be partially occluded.

	# labeled	CelebAMask-HQ	Car-Parts	DeepFashion-MM	DeepFashion-MM	Cityscapes
	images	8 classes	10 train	8 classes	10 classes	8 classes
DatasetGAN	16	0.7013	×	×	×	×
EditGAN	16	0.7244	0.6023	×	×	×
Transfer Learning	16	0.4575	0.3232	0.5192	0.4564	0.4954
HandsOff (Ours)	16	**0.7814**	**0.6222**	**0.6094**	**0.4989**	**0.5510**
Transfer Learning	50	0.6197	0.4802	0.6213	0.5559	0.5745
HandsOff (Ours)	50	**0.7859**	**0.6679**	**0.6840**	**0.5565**	**0.6047**

Table 1. Downstream task performance for semantic segmentation tasks across various domains, reported in mIOU (↑). HandsOff outperforms all baselines across all domains with both 16 and 50 labeled training images. × indicates a method that could not be run for a particular domain due to methodological shortcomings, such as requiring additional hand-labeled data.

	# labeled	CelebAMask-HQ			DeepFashion-MM			Cityscapes-Depth		
	images	PCK-0.1 ↑	PCK-0.05 ↑	PCK-0.02 ↑	PCK-0.1 ↑	PCK-0.05 ↑	PCK-0.02 ↑	mNMSE ↓	RMSE ↓	RMSE-log ↓
Transfer Learning	16	78.96	42.06	7.32	91.24	83.52	48.21	0.4022	18.12	2.75
HandsOff (Ours)	16	**97.19**	**76.36**	**17.44**	**94.19**	**88.48**	**70.22**	**0.2553**	**14.52**	**1.64**
Transfer Learning	50	90.88	61.75	12.30	91.24	83.52	48.20	0.2525	15.07	3.01
HandsOff (Ours)	50	**97.71**	**79.99**	**19.10**	**95.41**	**90.89**	**74.02**	**0.1967**	**13.01**	**1.58**

Table 2. Downstream task performance for keypoint detection and depth estimation. HandsOff outperforms all other methods when trained on 16 or 50 labeled images, demonstrating an impressive ability in generating *continuous*-valued keypoint heatmaps and depth maps.

Within the full-body human pose domain, HandsOff produces finely detailed segmentation masks, best illustrated by the segmentation mask for the first human in the top row of Fig. 3, who is wearing a pair of ripped jeans and the second human in the top row who is wearing the same colored jacket and shirt (see caption for more details). Generated labels are consistently high quality across a diverse array of object orientations, as seen in the various face rotations, human poses, or car orientations of Fig. 3. Finally, in extremely complex scenes, such as Cityscapes, HandsOff produces labels for visually minuscule classes, such as street lamps or traffic signs.

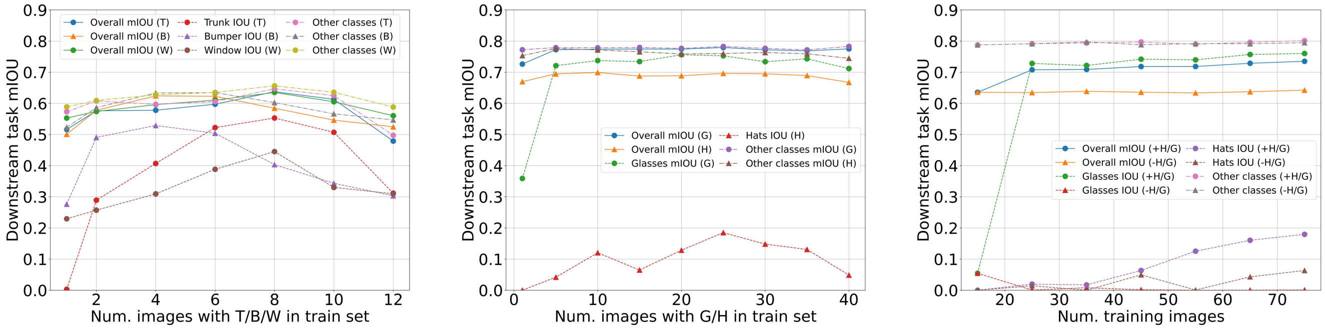
4.3. Segmentation results

As seen in Tab. 1, we achieve **state-of-the-art performance** on synthetic data trained semantic segmentation in all four domains, as measured in mean Intersection-over-Union (mIOU). Specifically, HandsOff outperforms DatasetGAN by 11.4% and EditGAN by 7.9% in the face domain when trained with *the same number of labeled images*. Increasing the number of labeled training images for HandsOff results in further performance gains, with 12.1% and 8.5% improvements over DatasetGAN and EditGAN, respectively. Unlike DatasetGAN, we are able to increase the number of labeled training images without incurring the associated costs of collecting new human annotated images. We emphasize again that with new domains, such as full-body human poses or urban driving scenes, it is not possible to train DatasetGAN-based frameworks as they rely on manual labels for GAN generated images. Therefore, we benchmark against the transfer learning baseline in these domains. Notably, HandsOff outperforms the transfer learning baseline by 17.4% (full-body human poses) and 11.2% (urban driving scenes) when both methods are trained on 16 labeled images; and 10.1% (full-body human poses) and 5.3% (urban driving scenes) when trained on 50 images.

4.4. Keypoint and depth results

We utilize HandsOff to generate *continuous* valued labels for keypoints and depth tasks. As seen in Tab. 2, we demonstrate strong empirical performance in generating both keypoints and depth maps. For keypoints, we report the Percentage of Correct Keypoints (PCK) for different threshold values α, denoted PCK-α. For a keypoint to be predicted correctly, the estimate must be no further from the true keypoint than $\alpha \cdot \max\{h, w\}$, where h and w are the height and width of the minimum size bounding box that contains all of the keypoints. We note that even for small α (i.e., $\alpha = 0.02$), HandsOff is able to correctly predict 2.4× and $1.5\times$ more keypoints than the transfer learning baseline in the face and full-body human pose domains, respectively. This implies that HandsOff is able to predict keypoints up to an extremely tight radius of the original keypoint location compared to other methods.

For depth, we report masked normalized mean-squared error (mNMSE), root mean-squared error (RMSE), and root mean-squared error of the log-depth values (RMSE-log). Because Cityscapes depth maps contain corrupted depth



(a)

(b)

(c)

Figure 4. Substitution experiments for various long-tail parts; (a) in cars - trunk (T), back bumper (B), back window (W); (b) in faces - glasses (G), hats (H). As the proportion of images containing the long-tail part increases in the training set, the performance of the long-tail class improves until it enters the *overfitting* regime. Non-long-tail mIOU tracks closely with overall IOU, implying dramatic gains in long-tail IOU do not come at the expense of other parts. (c) Addition experiments for face long-tail parts. (+H/G) indicates that images containing hats and images containing glasses are added to a base set, while (-H/G) indicates images containing neither hats nor glasses are added. The long-tail IOU of both parts *simultaneously* increase as images containing hats and images containing glasses are added to the base training set, with no negative impact on the performance of other classes.

values, we train HandsOff only non-corrupted pixels. Furthermore, to compute mNMSE, we compute the normalized mean-squared error only on the non-corrupted pixels. That is, let $\widehat{\boldsymbol{y}}$ and $\boldsymbol{y}$ are the predicted and true depth maps, respectively, and $\boldsymbol{M}$ be a mask indicating the non-corrupted elements of $\boldsymbol{y}$. mNMSE is computed as $\frac{\|\widehat{\boldsymbol{y}}_{\boldsymbol{M}}-\boldsymbol{y}_{\boldsymbol{M}}\|_2^2}{\|\boldsymbol{y}_{\boldsymbol{M}}\|_2^2}$, where $\boldsymbol{a}_{\boldsymbol{M}}$ denotes the depth map $\boldsymbol{a}$ at non-corrupted locations. When reporting RMSE and RMSE-log, we adopt the standard practice [8, 36] in depth estimation of cropping the middle 50% of the image and clamping predicted depth values to be within 0.001 and 80 before computing RMSE and RMSE-log values. As shown in Tab. 2, HandsOff is able to achieve a sizable advantage in all three metrics, outperforming transfer learning, resulting in 36.5%, 19.9%, and 40.27% decreases in mNMSE, RMSE, and RMSE-log when trained on 16 labeled images and 22.1%, 13.6%, and 47.6% decreases when trained on 50 labeled images.

4.5. Long-tail semantic segmentation

The HandsOff framework's ability to generate high quality synthetic datasets unlocks new degrees of freedom for model development previously unachievable with fixed, hand-annotated datasets. We now explore one example: mitigating the effects of the long-tail common in semantic segmentation datasets. For CelebAMask-HQ, images with hats and glasses make up less than 5% of the 30,000 labeled images, and a similar situation exists with trunks, back bumpers, and back windows in the Car-Parts dataset. These examples form the long-tail classes of their respective datasets, and their rare occurrence during training results in poor model performance at evaluation time.

The HandsOff framework altogether sidesteps this limitation of traditional datasets: by generating labeled synthetic images, we can control the occurrence of rare classes in our training data and significantly mitigate the effects of the long-tail. Because training the label generator requires less than 50 annotated images, we only require 5-10 occurrences of long-tail classes in order to generate an unlimited number of those occurrences in our synthetic dataset. Our experiments precisely quantify the small number of annotated examples of rare classes required to significantly improve downstream task performance on those classes. They fall into two categories: **Substitution** experiments, that fix a total number of training images and vary the proportion of rare class occurrences, and **Addition** experiments, that grow the size of the training set by adding images with rare classes. The substitution experiments ensure that any gains in the performance of identifying the long-tail class are not a by-product of increasing training set size. We perform substitution experiments considering only one long-tail part at a time. On the other hand, the addition setting is indicative of how a practitioner would deploy HandsOff: starting with a base set of labeled training images and further augmenting it with images containing rare classes deemed crucial to identify. To mirror what often happens in practice, we perform addition experiments by adding images containing multiple long-tail classes at a time.

Substitution. We begin with an initial set of 16 (cars) or 50 (faces) labeled images containing one image of the rare part, and then vary the proportion of the rare part. As seen in Fig. 4a and 4b, a small proportion of rare classes results in poor class identification performance, but as the proportion of images with long-tail classes increases, the long-tail part IOU increases by as much as 0.55 for car trunks and 0.40 for face glasses before eventually plateauing. We note that hats are a particularly challenging part to generate labels for due

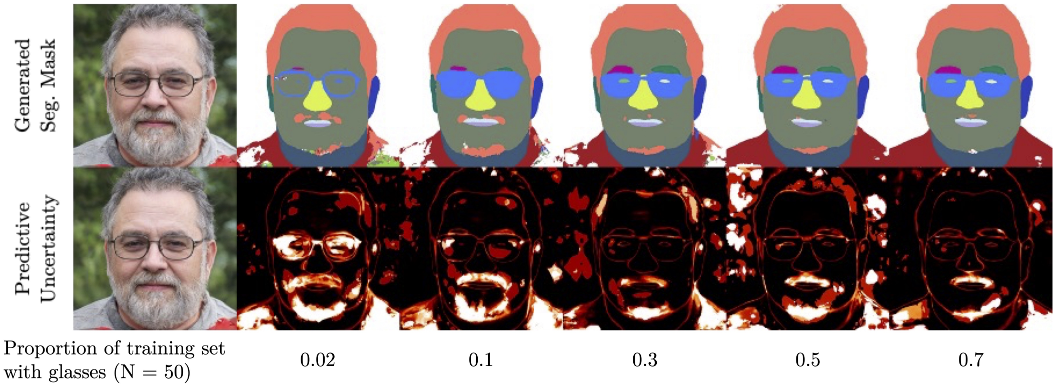


Figure 5. Visualization of generated segmentation mask (top row) and pixel-wise label generator uncertainty (bottom row) as the proportion of the training set containing the glasses increases. Not only do we see qualitative improvement in the generated label for glasses, we also see that the classifier is less *uncertain* when generating the correct label.

to the diversity of their size, shape, color, and orientation. Nevertheless, we still see a sizable increase of 0.2 IOU. We additionally plot the overall mIOU and the mIOU of non-long-tail parts to demonstrate that modifying the composition of the training set does not hurt performance on non-long-tail parts. In other words, shifting the training set part distribution to an extent has negligible impacts on the performance of non-long-tail parts, while resulting in large gains in long-tail class detection. Beyond proportions of $\sim$0.7, further increasing the proportion of the training set eventually causes drops in both long-tail part IOU and the mIOU of non-long-tail parts, owing to the label generator hallucinating long-tail classes where they do not belong. The impacts of substituting images with long-tail classes are best illustrated in Fig. 5. As the proportion of images with glasses grows, the generated mask captures glasses with increasing accuracy, eventually even distinguishing eyes that are visible through the glasses. Underneath the segmentation masks, we showcase the pixel-wise label generator uncertainty measured by Jensen-Shannon divergence (See Sec. 3.3). Not only does the generated label improve qualitatively, the label generator is *less uncertain* about the region of the image corresponding to the glasses.

Addition. We augment a small training set of 15 images with additional images containing hats or glasses. Fig. 4c demonstrates significant IOU increases (+0.71) in long-tail classes. The figure further highlights that these increases are not simply due to additional examples: targeted additions outperform the scenario where we add the same number of images, but the added images do **not** contain hats or glasses. These improvements in long-tail classes do not come at the expense of performance in other classes, as demonstrated by the overall mIOU and mIOU of non-long-tail classes. Unlike the substitution experiments, these performance improvements do not eventually drop, since the number of training examples continues to increase.

Our experiments showcase the power of the HandsOff framework to mitigate the long-tail problem. By explicitly including images with the long-tail class in our label generator training data, we are able to bridge the gap between performance in rare and common classes. The number of images with long-tail classes necessary to generate high quality labels of the long-tail is even smaller than the already small number of images needed to train HandsOff, meaning that the gains in long-tail class performance essentially come for free. If the long-tail class has been deemed crucial to identify, then it is likely that a practitioner has access to $\sim$20 labeled images containing the long-tail class. The performance gains in long-tail performance achieved by HandsOff are not practically replicable in DatasetGAN, where human supervision is needed to both identify generated images containing the long-tail class and provide precise pixel-level annotations.

5. Discussion

We present the HandsOff framework, which produces high quality labeled synthetic datasets without requiring further annotation of images for a multitude of tasks across various challenging domains. HandsOff achieves state-of-the-art performance over several recent baselines when training a downstream network with our synthetically generated data. Furthermore, HandsOff enables user control of the training data composition, leading to dramatic performance gains in long-tail semantic segmentation. This suggests that HandsOff can play a vital role in curtailing the effects of the long-tail. While synthetic datasets have the potential to supplant human annotations, they can also complement them. We leave as future work to investigate the collaborative power of having a human-in-the-loop refine synthetically generated annotations, and bring about the best of both worlds.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2StyleGAN: How to Embed Images Into the StyleGAN Latent Space? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2StyleGAN++: How to Edit the Embedded Images? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3

[3] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Labels4Free: Unsupervised segmentation using StyleGAN. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[4] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A Residual-Based StyleGAN Encoder via Iterative Refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 5

[5] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit Bermano. HyperStyle: StyleGAN Inversion with HyperNetworks for Real Image Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3

[6] William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The Power of Ensembles for Active Learning in Image Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5

[7] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[8] Vincent Casser, Soeren Pirk, Reza Mahjourian, and Anelia Angelova. Unsupervised Monocular Depth and Ego-motion Learning with Structure and Semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 7

[9] Edo Collins, Raja Bala, Bob Price, and Sabine Susstrunk. Editing in Style: Uncovering the Local Semantics of GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[11] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming Transformers for High-Resolution Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[12] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen-Change Loy, Wayne Wu, and Ziwei Liu. StyleGAN-Human: A Data-Centric Odyssey of Human Generation. 2022. 5

[13] Raghudeep Gadde, Qianli Feng, and Aleix M Martinez. Detail Me More: Improving GAN's photo-realism of complex scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 5

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. 2014. 2

[15] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. GANSpace: Discovering Interpretable GAN Controls. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[16] Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2Human: Text-Driven Controllable Human Image Generation. *ACM Transactions on Graphics (TOG)*, 2022. 5

[17] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[18] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training Generative Adversarial Networks with Limited Data. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[19] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-Free Generative Adversarial Networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[20] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[21] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and Improving the Image Quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 5

[22] Weicheng Kuo, Christian Häne, Esther Yuh, Pratik Mukherjee, and Jitendra Malik. Cost-Sensitive Active Learning for Intracranial Hemorrhage Detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2018. 5

[23] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. MaskGAN: Towards Diverse and Interactive Facial Image Manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5

[24] Daiqing Li, Huan Ling, Seung Wook Kim, Karsten Kreis, Sanja Fidler, and Antonio Torralba. BigDatasetGAN: Synthesizing ImageNet with Pixel-wise Annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[25] Daiqing Li, Junlin Yang, Karsten Kreis, Antonio Torralba, and Sanja Fidler. Semantic Segmentation with Generative Models: Semi-Supervised Learning and Strong Out-of-Domain Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[26] Huan Ling, Karsten Kreis, Daiqing Li, Seung Wook Kim, Antonio Torralba, and Sanja Fidler. EditGAN: High-Precision Semantic Image Editing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2, 3

[27] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 5

[28] Prem Melville, Stewart M Yang, Maytal Saar-Tsechansky, and Raymond Mooney. Active Learning for Probability Estimation Using Jensen-Shannon Divergence. In *Proceedings of the European Conference on Machine Learning*, 2005. 5

[29] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[30] Kitsuchart Pasupa, Phongsathorn Kittiworapanya, Napasin Hongngern, and Kuntpong Woraratpanya. Evaluation of deep learning algorithms for semantic segmentation of car parts. *Complex & Intelligent Systems*, 2021. 5

[31] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in Style: a StyleGAN Encoder for Image-to-Image Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[32] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal Tuning for Latent-based Editing of Real Images. *ACM Transactions on Graphics (TOG)*, 2022. 2, 3

[33] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the Latent Space of GANs for Semantic Face Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[34] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an Encoder for StyleGAN Image Manipulation. *ACM Transactions on Graphics (TOG)*, 2021. 2

[35] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-Fidelity GAN Inversion for Image Attribute Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[36] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The Temporal Opportunist: Self-Supervised Multi-Frame Monocular Depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 7

[37] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. GAN Inversion: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[38] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3

[39] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep Long-Tailed Learning: A Survey. *arXiv preprint arXiv:2110.04596*, 2021. 1

[40] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. DatasetGAN: Efficient Labeled Data Factory with Minimal Human Effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 5

[41] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative Visual Manipulation on the Natural Image Manifold. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 3