Dataset Enhancement with Instance-Level Augmentations

Orest Kupyn[1,2] and Christian Rupprecht[1]

[1] Visual Geometry Group - University of Oxford
{okupyn, chrisr}@robots.ox.ac.uk
[2] PiñataFarms AI

Fig. 1: Instance-Level Augmentations. We augment images by redrawing individual objects in the scene retaining their original shape. This allows training with the unchanged class label (*e.g.* class, segmentation, detection, *etc.*). The generations are highly diverse and match the scene composition. Guess the original in each row!

Abstract. We present a method for expanding a dataset by incorporating knowledge from the wide distribution of pre-trained latent diffusion models. Data augmentations typically incorporate inductive biases about the image formation process into the training (e.g. translation, scaling, colour changes, etc.). Here, we go beyond simple pixel transformations and introduce the concept of instance-level data augmentation by repainting *parts* of the image at the level of object instances. The method combines a conditional diffusion model with depth and edge maps control conditioning to seamlessly repaint individual objects inside the scene, being applicable to any segmentation or detection dataset. Used as a data augmentation method, it improves the performance and generalization of the state-of-the-art salient object detection, semantic segmentation and object detection models. By redrawing all privacy-sensitive instances (people, license plates, etc.), the method is also applicable for data anonymization. We also release fully synthetic and anonymized expansions for popular datasets: COCO, Pascal VOC and DUTS. The project page is available here.

1 Introduction

Deep learning has transformed computer vision research and applications. Interestingly, methodological progress now advances hand in hand with the avail-

ability of large annotated datasets [5]. This now shifts some of the focus from methods to their training data. Datasets have grown exponentially in size and complexity and are often just as important as the methods themselves. From the one million images of ImageNet [9] to LAION-5B [42], the number of samples has increased by three orders of magnitude. Large-scale datasets proved to be easily applicable to a wide range of computer vision transfer tasks [18, 19, 36]. Yet, growing and curating such datasets at this scale is a cumbersome and costly challenge. Additionally, scraping large image collections from the internet does not adhere to ethical and data privacy standards. Moreover, in many fields, collecting large-scale datasets is impracticable due to the lack of available data and the complexity of the collection and annotation process (medical data, 3D, etc.).

In the wake of larger and larger datasets [23, 42], many older and/or smaller datasets have lost relevance purely due to their size even though they pose unique challenges, applications and benchmarks. One of the main limiting factors of small datasets is their ability to train models that do not overfit. Over the last decade, many ideas have been explored to overcome overfitting problems on small datasets. Data augmentations such as colour shifts, flipping, crops, and rotations are being used to incorporate priors about geometric and lighting variations that we expect in the real world [44]. Their effectiveness is proven by their widespread use to date. Image-level augmentations are geometric transformations or pixel manipulations, such as adding Gaussian blur or changing image contrast. On the level of objects, augmentations are also often simple: for example, copy-pasting objects between images [14].

Simple pixel manipulations only slightly expand a dataset's information and visual diversity. Thus, several advanced techniques to mix images have been proposed [27,59,60] with moderate success. Combining two or more images often creates visual artefacts that models can exploit to recognize artificial samples. As a further improvement, generative models (trained on the to-be-augmented dataset) have been used to generate additional training samples [51,55]. However, as the generative model is trained on the same data, it cannot provide additional information beyond the dataset itself. Additionally, generating whole images also requires generating the accompanying annotations, which can be very challenging to do and verify.

To overcome the problem of annotation generation, we introduce a method that only regenerates a *part* of an image at the level of object instances. This allows us to retain the original data annotations (*e.g.* class labels, segmentation, captions, etc.). To significantly enhance the visual variability of the original data, we incorporate out-of-domain knowledge via a large diffusion-based generative model. As these models have been trained on much larger volumes of data, their generations will naturally go beyond the variety originally contained in the dataset. Another advantage of this approach is that it allows us to naturally combine real and generated data *in the same image*. This exponentially increases the variations of samples that can be constructed from a single image. For example, an image with five objects has 32 variations of real/generated samples if we consider only one generation per object. Finally, for privacy-sensitive

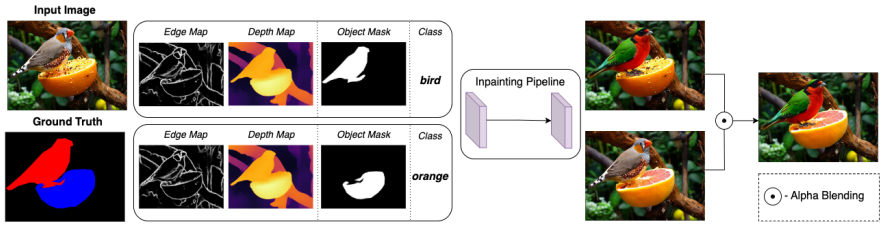


Fig. 2: Overview. Given an image and ground truth (or predicted) segmentation mask, we estimate depth and edge maps at the image level. The annotation is decomposed into the per-object binary masks and class, which together form the conditioning of the inpaining model. We redraw every instance and recombine them into a final image using alpha-blending sorted by depth.

instances such as people, license plates, etc. one can completely replace all real data in the dataset, strongly mitigating privacy concerns.

In this paper, we experiment with three tasks (object detection, semantic segmentation and saliency segmentation) and six datasets (MS-COCO [28], PascalVOC [12], DUTS [48], ECSSD [56], DUT-OMRON [57] and HKU-IS [25]). We show consistent improvements using our method to generate data for state-of-the-art methods for each task. Further, we can show that fully replacing people in the training data with synthetic samples does not affect the final performance, which enables retrospectively improving privacy shortcomings of scraped internet datasets. Together with the paper, we are releasing the code and generative variations of DUTS [48], Pascal VOC [12] and COCO [28].

2 Related Work

This section provides an overview of related work on diffusion models, followed by a survey of the methods targeting synthetic data generation.

Image Generative Models. Generative Adversarial Networks first introduced efficient sampling of high-resolution images with good perceptual quality [4, 15, 21]. GAN models can generate visually plausible images but are challenging to optimize and struggle to capture the full data distribution [31]. Recently, diffusion probabilistic models [16, 45] have been introduced to match the underlying data distribution by learning to reverse a sample noising process. The high quality and stability of the training process quickly set diffusion models apart as a frontier in the field of image synthesis [17, 34, 38, 40]. Text-to-image diffusion models condition the generation process by encoding text prompts into the latent vectors utilizing large pre-trained text encoders [37]. Latent diffusion models [39] introduce a more general conditioning method fusing text embeddings, bounding boxes, or inpainting masks through cross-attention layers in a convolutional manner, achieving state-of-the-art results in image inpainting and class-conditional image synthesis performing diffusion steps in the latent image space [11]. ControlNet [61] and T2I-Adapter [33] add spatial conditional control

to diffusion models by encoding a target image representation in the form of edges, depth, segmentation, human pose, *etc*.

Generative Models for Synthetic Datasets. Early studies in synthetic data generation [7, 13] rely on 3D rendering engines to address common 2D vision problems. The process is limited by the domain of 3D models, cannot be generalized to complex real-world scenes and does require modifying the rendering engine for each specific dataset and subtask. For example, Virtual KITTI consists of exclusively street driving scenes for autonomous driving. In contrast, synthetic data generated using generation models using generative adversarial networks [29, 55] are more flexible and generalize better to real-world images. Recently [53] combines diffusion embedding network with the BigGAN [4] to generate synthetic images for the saliency detection. However, those methods primarily focus on learning the real dataset distribution, thus unable to incorporate new information into the dataset. Recently, [3, 41, 43, 47] improve the performance of classification models by generating synthetic data with latent diffusion models [39]. However, the methods are limited to image classification. Diffumask [52] and DatasetDM [51] utilize large diffusion models to simultaneously generate synthetic images and annotations for semantic segmentation or depth prediction tasks. Conditioning on text embeddings from large language models [37] provides a mechanism to encode information outside of the original data distribution. Still, the methods do require finetuning on every specific dataset. In contrast, our method directly employs pre-trained diffusion models without additional finetuning steps, not overfitting to the task-specific datasets and providing a way to efficiently mix real and synthetic annotations for training various vision models. It generalizes existing datasets by incorporating rich priors from the large-scale datasets used for image synthesis by augmenting image samples on an object level.

3 Method

The goal of our method is to take an image $I \in \mathcal{I} = \mathbb{R}^{3\times H\times W}$ with dimensions H and W, and a set of annotations $\mathcal{Y}$, and generate a new image $I^* \in \mathbb{R}^{3\times H'\times W'}$ with a mapping $F(I, \mathcal{Y}) = I^*$ preserving the structure of the scene and ground truth annotations. This way, the sample $(I^*, \mathcal{Y})$ remains a valid training sample. The method should be able to generate high-fidelity image samples from a real-world distribution that are indistinguishable from the input image. Since older datasets are often of low resolution, we allow for an increased image size $H' \geq H, W' \geq W$ for the generated image.

3.1 Generation Pipeline

To generate new samples I^* we base the pipeline on a conditional latent diffusion model (LDM) [39]. Large-scale generative diffusion models have been trained on an extremely large variety of images, providing a strong signal to enhance the data distribution of any dataset.

Since a pure text-to-image generative model [40] is not applicable to the task, we use an LDM trained for image inpainting to operate only on a region of interest. For this purpose, we assume the following annotations to be available within $\mathcal{Y} = \{(M_i, c_i)\}_{1 \leq i \leq N}$. For each of the N objects in the image: a binary mask $M_i \in \mathcal{M} = \{0,1\}^{H \times W}$ defining the segmentation of the object in the image, and a class label c which can be free-form text. If this information is not already available in the dataset, it can be obtained by an off-the-shelf instance segmentation method.

In general, the image inpainting process can be formulated as a function $G : \mathcal{I} \times \mathcal{M} \times \mathcal{T} \to \mathcal{I}$, that takes as input an image, a mask selecting the inpainting region and a text prompt that specifies what to draw. The output is an image with the masked region being redrawn.

A new image I^* is generated by iteratively redrawing each object in the image.

$$I_i^* = G(I_{i-1}^*, M_i, T_i). \tag{1}$$

This iterative process starts with the original image $I_0^* = I$ and ends with every object being redrawn $I_N^* = I^*$. It requires a text-prompt T_i for each object. In the simplest case the class name can be used as the text prompt $T_i = c_i$. We will explore more sophisticated choices in Sec. 3.3.

In practice, two aspects are important to consider with the inpainting process. One is that the order in which the objects are being redrawn is arbitrary in Eq. (1), and does not take into account occlusions or overlapping objects, impacting the final result. The second is that repeatedly passing an image through an inpainting model significantly reduces the image quality. We will deal with both problems next.

Draw order. To ensure a reasonable blending of objects, we sort objects by their estimated relative distance to the camera. We use an off-the-shelf depth-estimation method DepthAnything [58] to compute a (relative) depth map $D \in \mathbb{R}_+^{H \times W}$ and use the masks to compute a depth estimate $d_i = \sum_{u \in \Omega} M_i[u] D[u] \in \mathbb{R}_+$, where $\Omega = \{u, v | 1 \leq u \leq H, 1 \leq v \leq H\}$ is the set of all pixel locations in the image. We can then order the images back-to-front such that $d_i \geq d_{i+1}$ before inpainting.

Iterative Redrawing. Current diffusion inpainting models struggle to reconstruct input images faithfully, even in unmasked regions, due to operating in latent space. This encoding and decoding process introduces noise, degrading image quality, particularly for complex images. In latent diffusion models (LDMs), this issue is evident, as shown in **??**, where repeated encoding and decoding reduce PSNR and SSIM significantly. We address this by modifying the process to always begin with the original image.

$$I_i^* = I_{i-1}^* \odot (\mathbb{1} - M_i) + G(I, M_i, T_i) \odot M_i. \tag{2}$$

This approach has two advantages. First, it does not require repeatedly applying the inpainter G to the image, which would accumulate noise. On the other

hand, this means that each object is drawn independently of the other objects, which allows generating an exponential number of 2^N variations of the image, where each object can be original or redrawn, with only N applications of the inpainter.

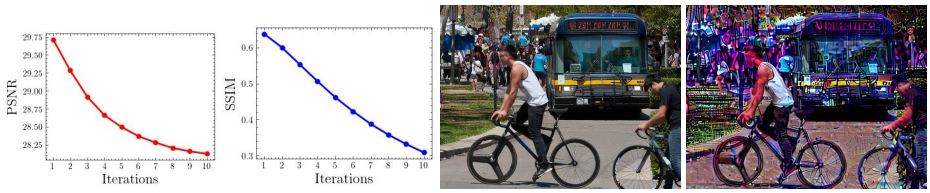


Fig. 3: Noise Accumulation. Images accumulate noise when repeatedly encoding and decoding to and from latent space. PSNR and SSIM compared to the original.

3.2 Better Inpainting

Naively employing an off-the-shelf inpainting model for our task can lead to several shortcomings. Using G as a black box, there is no guarantee that the new object is of the same class as before, nor that it adheres to the original mask M_i. We also observe that the LDM often tends to completely remove small objects, replacing them with background.

We thus make the following modifications to the inpainter G. Several extensions to LDM provide more fine-grained control of the generated output. Specifically, ControlNet [61] injects additional conditions into the decoder blocks of the LDM generator. We can thus condition the generation not only on the mask M_i, but also on the already computed depth map D and an edge map E that significantly decreases the potential discrepancy between the generated and original object contours. We compute the edge map using HED [54].

Interestingly, [61] has been trained on the non-inpainting version of [39], since the inpainting model was finetuned from the same baseline text to image model, they can still be used together [50].

3.3 Prompt Engineering

While we can simply condition the inpainting model on the class information ($T_i = c_i$), several improvements can be made to both increase the quality as well as the control for a diversity of the generated objects.

The diffusion model provides an efficient way to add additional conditioning to the generation by modifying the input text prompt. Trained on the billions of image-text description pairs from the internet, using CLIP [37] as a text encoder, the model can encode the pose, shape, visual appearance, colour, and lightning of the object simultaneously. Our goal is to generate highly diverse samples. We exploit the useful properties of the text encoder to improve the visual quality and the diversity of the generated samples.

Object Description. Following [41], we extend the label c_i with its description. Each class in datasets such as COCO or ImageNet is associated with one or more synsets, *i.e.* entities, in the WordNet [32] graph. We use the synset lemmas corresponding to each class to extend the class name. This additional description of the object provides a more precise textual representation of the object in CLIP space. Empirically, we observe that it stabilizes the performance of the diffusion model and prevents the cases where an incorrect or no object at all is generated due to weak text embedding.

Color and Lighting. We manually select categories that tend to appear in many colour variations in the real world, such as cars, backpacks, snowboards, *etc*. For every such object, we randomly sample from a list of colours (*e.g.*, blue, red, green, pink, *etc*.) and qualifiers (dark, light, natural, *etc*.) and add it as part of the prompt Additionally, we randomly select and add a lighting condition to the prompt (*e.g.*, sunlight, dramatic lighting, soft lighting, *none*, *etc*.). We extensively experiment with these choices in Sec. 5.4 and in the supplement. Overall, our prompt variations increase the diversity across all classes.

Generating People. "Person" is the most visually complex and diverse class in common computer vision benchmarks with 15.5% of DUTS and 30.5% of COCO containing people in different scenes. We observe that for many cases of complex real-world scenes, the above depth and edge maps conditioning and simple prompting strategy are insufficient to generate a person that adheres to the scene's structure. Thus, to increase the descriptiveness of the prompt, for people, we use a visual question-answering model (BLIP-VQA [26]) to predict the action of the person and add it as a part of the prompt.

4 Datasets

The method can generate images for an arbitrary dataset labelled with bounding boxes or segmentation masks, making it applicable for a wide range of tasks, including Object Detection, Semantic and Instance Segmentation, Panoptic Segmentation, *etc*. For our experiments, we generate synthetic versions for three popular datasets: DUTS [48], Pascal VOC [12] and MS-COCO [28]. We run our generation pipeline up to three times for each object to obtain three synthetic replacements of the original instance. This drastically increases the set of potential variations depending on the number of objects per image. Naturally, we only generate novel data for the training set and leave the validation and test set untouched so as not to affect benchmark results.

For additional safety, we keep regenerating objects that do not pass an NSFW safety filter [1] with a strict threshold minimising the false negative rate.The filter also detects NSFW content in the original datasets (which should not have been included in the dataset in the first place). Our redrawing method rectifies this. We will release all datasets and the code to generate them.

DUTS. [48] is the largest benchmark dataset for salient object detection. Yet it still only contains 10,553 training and 5,019 test images of various real-world scenes and types of objects. The annotations consist of a single saliency map per

image and do not contain class labels or names. To construct the text prompt for the pipeline, we crop the image by the bounding rectangle of the binarized saliency map and use the BLIP-VQA [26] model to predict the object name in an open vocabulary setting. The dataset is of relatively low resolution, which we also improve with our method.

Since salient object segmentation highly depends on predicting accurate object boundaries, we add an optional mask refinement stage to preserve the sharpness and high quality of the masks. To this end, we crop every generated object from the images in the train set using its corresponding bounding rectangle of the saliency map. We use TRACER-7 [24] as an off-the-shelf segmentation model to obtain high-resolution, tight object crop saliency maps. This removes the potential mismatch between the generated object and its original annotations.

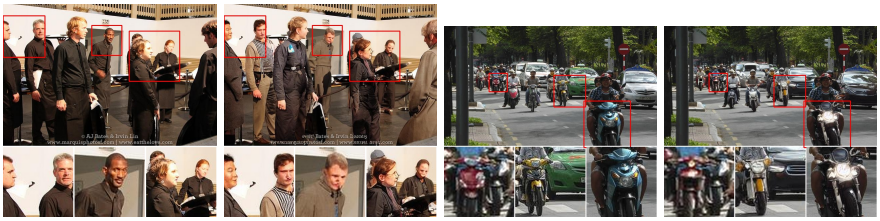
(a) DUTS **(b)** VOC **(c)** COCO

Fig. 4: Qualitative Evaluation. The examples of method performance on three different datasets. The method generalizes well to complex scenes and datasets with no ground truth instance labels.

MS-COCO. The MS-COCO dataset [28] is a standard benchmark for evaluating object detection, instance segmentation, and other tasks. It includes complex everyday scenes containing common objects in their natural context.

Pascal VOC. The PASCAL Visual Object Classes (VOC) dataset [12] contains images with pixel-level segmentation annotations, bounding box annotations, and object class annotations. This dataset has been widely used as a benchmark for object detection, semantic segmentation, and classification tasks.

Dataset Anonymization. In addition to the generic version of the datasets we generate, we use our method to improve the privacy aspect of both datasets. For DUTS, we manually select all images containing people in the dataset and repaint them with "virtual" people. We thus generate an anonymized version of the dataset, which can be used by itself with all original images and annotations (except people) or in conjunction with the generated version.



(a) Original **(b)** Generated **(c)** Original **(d)** Generated

Fig. 5: Data Anonymization. The method efficiently repaints each annotation in the complex scenes, strongly mitigating privacy concerns for sensitive instances such as people or cars.

For the COCO dataset, we generate two additional versions: anonymizing people and cars (for personal information such as license plates). For this task, to ensure every object (and not only the ten largest ones) in the dataset is anonymized, we process every object, including small instances. For objects with an area lower than $32 * 32$px the image is first cropped and upsampled to 512px resolution to match the minimum resolution the generator model is trained on. See Fig. 5 for visual examples of repainting people and vehicles.

5 Experimental Evaluation

We evaluate the efficiency of our method (i) as a data augmentation technique, (ii) as a pipeline for data anonymization, and (iii) its impact on the model generalization. To address (i) we train the state-of-the-art saliency segmentation, semantic segmentation and object detection models on the combined original and generated by our methods versions of DUTS, VOC and COCO datasets, respectively, and report the findings in the 5. To test the effectiveness of data anonymization (ii) we train the object detection models on the generated version of COCO datasets with all cars and people entirely replaced with the synthetic objects and compare it with the original model 6. To test generalization (iii) we analyze the performance of the models trained on combined datasets on the established benchmarks for salient object detection, detailed in 4.

5.1 Data Augmentation

Object Detection. The effect of the instance augmentations is tested on a large-scale complex COCO dataset. We evaluate the performance on three different object detection architectures, transformer-based: Deformable-DETR (single scale) [62], RT-DETR (with ResNet-50) backbone [30] and an anchor-free YOLOv5m [20]. On each training iteration, every instance in an image is repainted with 30% probability. This ensures a high diversity of training samples and reduces the overfitting. The results (Tab. 1) show that adding augmented data consistently improves the performance of object detection models.

Model	Data	AP	AP_{50}	AP_{75}
Def.-DETR	orig.	39.3	60.0	42.0
Def.-DETR	ours	**40.5**	**60.2**	**43.4**
RT-DETR	orig.	51.4	69.6	55.4
RT-DETR	ours	**52.4**	**69.7**	**56.5**
YOLOv5m	orig.	44.1	63.4	47.8
YOLOv5m	ours	**45.7**	**64.0**	**49.7**

Table 1: Object Detection. Our instance augmentations improve all object detectors on MS-COCO [28].

Data	10%	25%	50%	75%	100%
orig.	25.8	34.2	39.1	41.2	44.1
ours	**27.9**	**36.1**	**40.3**	**42.5**	**45.7**

Table 2: Data-Sparse Object Detection. Our data improves the performance of YOLOv5 detector consistently, even with limited amount of training data available.

Additionally, we evaluate the effectiveness of augmented data in a data-spare setting, with only a subset of the training data available. We randomly select subsample 10%, 25%, 50%, 75% and 100% of the COCO training set, covering all categories. Table 2 shows the consistent boost of *+2-3 AP* for each setting.

Semantic Segmentation. To demonstrate the impact of augmentations when no ground truth instance masks are available we evaluate the performance of semantic segmentation models on Pascal VOC augmented set. We utilize the pseudo-ground-truth instances from [2]. To generate the augmented images. Following [52] we first train Mask2Former [8] with ResNet-50 backbone on the synthetic dataset and then finetune the model on the subset of original VOC data. 3 show significant improvement of **15.5 mIoU** with even larger improvements on challenging categories (**42.4** vs **14.7** mIoU for chair). The main limitation of generating fully synthetic samples is what motivates our approach: diffusion models do not perform well in generating complex scenes with multiple foreground/background objects, even with additional guidance. Further, the model finetuned on the original data achieves superior performance on almost all categories comparing both to the baseline and the Diffumask method. While the Diffumask targets different application and use only generated data it is, however, one of the closest works in terms of generating synthetic data with diffusion models. Thus the comparison still provides insights on the effectivness of data augmentations in cases with no ground truth instance masks.

Salient Object Detection. The performance of the salient object detection models is evaluated on four popular datasets, namely ECSSD [56] with 1000 images of relatively complex backgrounds, DUT-OMRON [57] with 5168 images that include one or more salient objects with rather complex backgrounds, HKU-IS [25] with 4,447 images, that include two or more objects with various backgrounds and DUTS [48] with 15,572 images which is the largest available dataset for training divided into 10,553 training images (DUTS-TR) and 5,019 testing images (DUTS-TE). All datasets are labelled with pixel-wise ground truth.

Real	Syn.														mIoU
VOC	✗	87.5	**94.4**	70.6	**95.5**	87.7	92.2	**44.0**	85.4	89.1	82.1	**89.2**	80.6	53.6	77.3
✗	[52]	80.7	86.7	56.9	81.2	74.2	79.3	14.7	63.4	65.1	64.6	71.0	64.7	27.8	57.4
✗	Ours	86.5	89.1	71.7	85.9	80.7	92.5	42.2	66.0	87.2	74.8	86.3	72.0	43.5	72.9
VOC	[52]	85.4	92.8	**74.1**	92.9	83.7	91.7	38.4	**86.5**	86.2	82.5	87.5	81.2	39.8	77.6
VOC	Ours	**89.2**	89.4	69.0	92.3	**87.8**	**93.6**	40.5	79.8	**89.6**	**86.7**	88.9	**85.2**	**61.5**	**78.2**

Table 3: Semantic Segmentation Evaluation. We evaluate the performance of Mask2Former [8] with ResNet-50 backbone on VOC [12] dataset. Training on augmented data improves the results by a wide margin compared to [52], while finetuning on a part of the original dataset further improves model performance.

		ECCSD			DUTS-TE			DUT-OMRON			HKU-IS		
Model	**Data**	$F_{max}\uparrow$	**MAE**$\downarrow$	$S_m\uparrow$	$F_{max}\uparrow$	**MAE**$\downarrow$	$S_m\uparrow$	$F_{max}\uparrow$	**MAE**$\downarrow$	$S_m\uparrow$	$F_{max}\uparrow$	**MAE**$\downarrow$	$S_m\uparrow$
U2Net	orig.	0.944	0.052	0.900	0.863	0.066	0.836	0.835	0.075	0.819	0.930	0.043	0.895
U2Net	ours	**0.948**	**0.047**	**0.908**	**0.874**	**0.061**	**0.844**	**0.848**	**0.069**	**0.827**	**0.935**	**0.040**	**0.900**
F3-Net	orig.	0.955	0.035	0.924	0.899	0.038	0.887	0.831	**0.054**	0.835	0.942	**0.029**	0.918
F3-Net	ours	**0.962**	**0.033**	**0.929**	**0.907**	**0.037**	**0.890**	**0.850**	0.055	**0.843**	**0.947**	**0.029**	**0.922**
TRACER-4	orig.	0.956	0.027	0.929	0.911	0.029	0.896	0.847	0.048	0.848	0.944	0.024	0.921
TRACER-4	ours	**0.960**	**0.026**	**0.933**	**0.918**	**0.026**	**0.905**	**0.848**	**0.045**	**0.853**	**0.948**	**0.022**	**0.928**

Table 4: Salient Object Detection Evaluation. Comparison of performance with five existing methods on five benchmark datasets. The best results per method pair are highlighted. Our method improves the performance in 34/36 metrics.

We use three saliency segmentation models: U2Net [35], F3-Net [49] and TRACER [24]. Specifically, we choose the U2Net-lite version as a simpler architecture and the more complex multi-head TRACER model with an EfficientNet-4 [46] backbone to show the impact of the data augmentation on different model sizes and architectures.

Table 4 shows saliency segmentation results of the three models on the original and our version of each dataset. Using our data pipeline over the original images improves the performance in almost all metrics across five datasets and three models with up to 8% in error reduction. We note that all models are trained on DUTS (either the original data or ours). Thus, this benchmark also measures generalization, which is one of the strengths of our method.

We also investigate how traditional data augmentations (including flips, rotations, adding noise, blur, brightness, contrast, *etc.*) as implemented in [6] affect the results. In Tab. 5, we analyse the influence of using augmentations, real data and or synthetic data. Additionally, we compare our synthetic data generation pipeline to the synthetic dataset of [53]. Classical image augmentations are complementary to our method and can be used in conjunction. Further, our synthetic data performs better than [53], especially for generalization.

5.2 Data Anonymization

The effectiveness of the introduced method for data anonymization is evaluated on COCO using our anonymized datasets, where all cars and people in the

Model	Aug	Real	Syn.	ECCSD $F_{max}\uparrow$	ECCSD MAE$\downarrow$	ECCSD $S_m\uparrow$	DUTS-TE $F_{max}\uparrow$	DUTS-TE MAE$\downarrow$	DUTS-TE $S_m\uparrow$	DUT-OMRON $F_{max}\uparrow$	DUT-OMRON MAE$\downarrow$	DUT-OMRON $S_m\uparrow$	HKU-IS $F_{max}\uparrow$	HKU-IS MAE$\downarrow$	HKU-IS $S_m\uparrow$
F3-Net	✗	✗	[53]	0.938	0.039	0.913	0.856	0.053	0.851	0.793	0.077	0.802	0.932	0.032	0.909
F3-Net	✗	✗	ours	**0.959**	**0.036**	**0.923**	**0.903**	**0.040**	**0.880**	**0.841**	**0.058**	**0.829**	**0.949**	**0.029**	**0.919**
TRACER-0	✗	✓	✗	0.943	0.035	0.912	0.884	0.035	0.874	0.816	0.050	0.825	0.929	0.031	0.904
TRACER-0	✓	✓	✗	**0.949**	0.032	0.919	0.889	0.035	0.880	**0.833**	0.051	0.836	0.933	0.028	0.912
TRACER-0	✗	✓	ours	**0.949**	0.033	0.919	0.891	0.034	0.879	0.826	**0.048**	0.834	0.932	0.029	0.909
TRACER-0	✓	✓	ours	**0.949**	**0.031**	**0.923**	**0.892**	**0.033**	**0.885**	0.829	0.049	**0.839**	**0.937**	**0.027**	**0.916**

Table 5: Data Augmentation. Comparison of the influence of standard augmentations (Aug) and the influence of including real samples (Real) as well as a comparison with the augmentation strategy in [53].

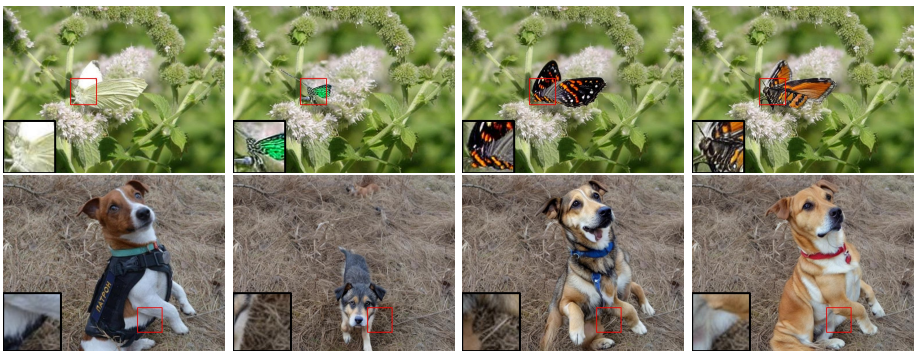
Model	Dataset	all classes AP	all classes AP_{50}	all classes AP_{75}	people only AP	people only AP_{50}	people only AP_{75}	cars only AP	cars only AP_{50}	cars only AP_{75}
Deformable-DETR	Real	39.3	60.0	42.0	49.7	78.5	52.7	39.5	68.5	39.0
Deformable-DETR	Anonymized People	38.3	58.7	40.6	45.6	75.2	47.1	-	-	-
Deformable-DETR	Anonymized Cars	38.7	59.5	41.5	-	-	-	37.2	65.6	36.7

Table 6: Data Anonymization. Replacing all real data of people or cars only has a marginal impact on the overall performance. This is important as one might not even be interested in a detector for people, yet people are contained in the images. The replaced category slightly decreases in performance which is acceptable given the strong increase in privacy.

dataset were substituted with the synthetic versions. We train a Deformable-DETR [62] model for 50 epochs on the original MS-COCO [28] and the two anonymized variants and compute mean average precision and thresholded average precision metrics for all objects and people and cars separately. The model trained on purely synthetic subcategories without real people/car objects achieves comparable performance to the model trained on the original dataset within the category and on a complete validation set, proving the method can efficiently be employed to anonymize potentially sensitive data. While the performance of the synthetic categories decreases slightly, the overall performance is almost unaffected (1 AP drop with 30.5% of all instances repainted). As many applications are object-centric and do not depend on detecting humans accurately, the boost in privacy comes at no cost in performance. Additionally, we measure the anonymization strength comparingly to the state-of-the-art face anonymization method [22] by calculating how often a replaced face can be matched with its original appearance using a face identification model. To this end, we use ArcFace [10] on the pairs of original and generated images from the COCO dataset, validating whether people can be re-identified after applying our method. From 64115 images with people and 262465 faces, only 373 (0.14%) of faces were re-identified vs. 2.8% faces anonymized by LDFA [22]. This shows the importance of extending the method beyond face anonymization because it can replace all pixels covering a given individual. It is important as making only faces unidentifiable can be considered to be only pseudonymization, i.e., not a complete removal of personal information from an image.

Model	Dataset	$F_{max}\uparrow$	MAE$\downarrow$	$S_m\uparrow$	$F_{avg}\uparrow$
TRACER-0	Real	0.889	0.035	0.880	0.847
TRACER-0	Anonymized	0.886	0.036	0.876	0.846

Table 7: Anonymizing People in DUTS. Training with only synthetic humans achieves comparable performance to the model trained on real images of humans.



a) Original b) Inpainting c) Depth Map d) Depth + Edge

Fig. 6: Inpainting Conditioning. Inpainting without conditioning (b) does not preserve the structure of the object. Depth conditioning alone (c) fails to generate finer details and sharp edges. The full model (d) accurately preserves original annotations while producing high-quality samples.

Additionally, we test the effect on the DUTS dataset by anonymizing all people in Tab. 7. Similarly, the drop in performance is negligible, demonstrating high potential for tasks where preserving data privacy is crucial.

5.3 Ablation Study

In this section, we verify the pipeline components' importance. We start with the full pipeline and analyze its performance on DUTS by subtracting individual components. We use TRACER [24] with EfficientNet-0 [46] for all experiments and report the results in Tab. 8.

Our method conditions the inpainting network on the predicted depth and edge map. Removing it from the pipeline results in a performance drop across all metrics. We also train a model without the additional edge map input. This also results in reduced performance, indicating that both depth and edge map help preserve both high and low-level information. Figure 6 compares the original image to naive inpainting and control with depth, mask, and both. Additionally, we train a model on the version of the dataset with full image generated from only the depth and edges condition on the input. The main limitation of this method is still failing to generate scenes with a complex structure.

Component	$F_{max}\uparrow$	MAE$\downarrow$	$S_m\uparrow$	$F_{avg}\uparrow$
full method	**0.892**	**0.033**	**0.885**	**0.853**
w/o instances (full image)	0.889	0.035	0.881	0.848
w/o edge and depth control	0.886	0.036	0.881	0.846
w/o edge control	0.890	0.034	0.880	0.850
w/o prompt engineering	0.889	0.034	0.880	0.849
w/o mask refinement	0.888	0.035	0.879	0.844

Table 8: Ablation study on DUTS. Removing any component from the generation pipeline decreases the downstream task performance. The most important components are additional conditioning of the inpaiting model (ControlNet) and mask refinement.

Fig. 7: Prompt Engineering. Samples generated without (left) and with (right) color/lighting prompts. Extra prompts result in more diversity and variations.

5.4 Prompt Engineering

We introduce several improvements over simply using the class name as a prompt. We verify this choice in Tab. 8 where we show that the down-stream task performance decreases without "prompt engineering". Further, we show visual examples in Fig. 7. The class label prompt limits the diversity of the generated cups, whereas including color and light information shows much more diverse results. Finally, we observe that without including more detailed descriptions for people, the inpainter tends to sometimes fully remove (inpainting the background) complicated objects instead of redrawing them.

Mask Refinement. The pipeline performance is bounded by the performance of the underlying methods. In complex scenes, the error level of the depth or edge map model might cause the divergence between the generated image and the original annotations. The mask refinement module fixes disconnection, yet again improving the metrics and ensuring the stability of the generation pipeline.

6 Conclusions

In this paper, we introduce a method for object-level data augmentation. We combine a large pre-trained diffusion model with the low-level object representation to sample high-quality, diverse samples outside of the original dataset distribution. We demonstrate the method's efficiency as a data augmentation and anonymization technique. Additionally, we release the synthetic and anonymized versions of the standard detection benchmarks, COCO, VOC and DUTS.

References

1. Machine vision & learning group lmu. safety checker model card. https://huggingface.co/CompVis/stable-diffusion-safety-checker, accessed: 2023-11-16 7
2. Ahn, J., Cho, S., Kwak, S.: Weakly supervised learning of instance segmentation with inter-pixel relations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2209–2218 (2019) 10
3. Azizi, S., Kornblith, S., Saharia, C., Norouzi, M., Fleet, D.J.: Synthetic data from diffusion models improves imagenet classification. arXiv preprint arXiv:2304.08466 (2023) 4
4. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018) 3, 4
5. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020) 2
6. Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A.: Albumentations: fast and flexible image augmentations. Information **11**(2), 125 (2020) 11
7. Cabon, Y., Murray, N., Humenberger, M.: Virtual kitti 2. arXiv preprint arXiv:2001.10773 (2020) 4
8. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022) 10, 11
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009) 2
10. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019) 12
11. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12873–12883 (2021) 3
12. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal of Computer Vision **88**(2), 303–338 (Jun 2010) 3, 7, 8, 11
13. Gaidon, A., Wang, Q., Cabon, Y., Vig, E.: Virtual worlds as proxy for multi-object tracking analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4340–4349 (2016) 4
14. Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.Y., Cubuk, E.D., Le, Q.V., Zoph, B.: Simple copy-paste is a strong data augmentation method for instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2918–2928 (2021) 2
15. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014) 3
16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 3

17. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. The Journal of Machine Learning Research **23**(1), 2249–2281 (2022) 3
18. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., et al.: Openclip, july 2021. If you use this software, please cite it as below **2**(4), 5 2
19. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021) 2
20. Jocher, G.: Ultralytics yolov5 (2020). https://doi.org/10.5281/zenodo.3908559, https://github.com/ultralytics/yolov5 9
21. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8110–8119 (2020) 3
22. Klemp, M., Rösch, K., Wagner, R., Quehl, J., Lauer, M.: Ldfa: Latent diffusion face anonymization for self-driving applications. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3198–3204 (2023) 12
23. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International Journal of Computer Vision **128**(7), 1956–1981 (2020) 2
24. Lee, M.S., Shin, W., Han, S.W.: Tracer: Extreme attention guided salient object tracing network (student abstract). In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 12993–12994 (2022) 8, 11, 13
25. Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5455–5463 (2015) 3, 10
26. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022) 7, 8
27. Li, S., Wang, Z., Liu, Z., Wu, D., Tan, C., Li, D.W.S.Z.: Openmixup: A comprehensive mixup benchmark for visual classification. ArXiv **abs/2209.04851** (2022) 2
28. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014) 3, 7, 8, 10, 12
29. Ling, H., Kreis, K., Li, D., Kim, S.W., Torralba, A., Fidler, S.: Editgan: High-precision semantic image editing. Advances in Neural Information Processing Systems **34**, 16331–16345 (2021) 4
30. Lv, W., Xu, S., Zhao, Y., Wang, G., Wei, J., Cui, C., Du, Y., Dang, Q., Liu, Y.: Detrs beat yolos on real-time object detection. arXiv preprint arXiv:2304.08069 (2023) 9
31. Metz, L., Poole, B., Pfau, D., Sohl-Dickstein, J.: Unrolled generative adversarial networks. arXiv preprint arXiv:1611.02163 (2016) 3
32. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM **38**(11), 39–41 (1995) 7

33. Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453 (2023) 3
34. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023) 3
35. Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O.R., Jagersand, M.: U2-net: Going deeper with nested u-structure for salient object detection. Pattern recognition **106**, 107404 (2020) 11
36. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 2
37. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 3, 4, 6
38. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022) 3
39. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 3, 4, 6
40. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022) 3, 5
41. Sarıyıldız, M.B., Alahari, K., Larlus, D., Kalantidis, Y.: Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8011–8021 (2023) 4, 7
42. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022) 2
43. Shipard, J., Wiliem, A., Thanh, K.N., Xiang, W., Fookes, C.: Diversity is definitely needed: Improving model-agnostic zero-shot classification via stable diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 769–778 (2023) 4
44. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. Journal of big data **6**(1), 1–48 (2019) 2
45. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015) 3
46. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019) 11, 13
47. Tian, Y., Fan, L., Isola, P., Chang, H., Krishnan, D.: Stablerep: Synthetic images from text-to-image models make strong visual representation learners. arXiv preprint arXiv:2306.00984 (2023) 4

48. Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X.: Learning to detect salient objects with image-level supervision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 136–145 (2017) 3, 7, 10
49. Wei, J., Wang, S., Huang, Q.: F^3net: fusion, feedback and focus for salient object detection. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 12321–12328 (2020) 11
50. Wortsman, M., Ilharco, G., Kim, J.W., Li, M., Kornblith, S., Roelofs, R., Lopes, R.G., Hajishirzi, H., Farhadi, A., Namkoong, H., et al.: Robust fine-tuning of zero-shot models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7959–7971 (2022) 6
51. Wu, W., Zhao, Y., Chen, H., Gu, Y., Zhao, R., He, Y., Zhou, H., Shou, M.Z., Shen, C.: Datasetdm: Synthesizing data with perception annotations using diffusion models. arXiv preprint arXiv:2308.06160 (2023) 2, 4
52. Wu, W., Zhao, Y., Shou, M.Z., Zhou, H., Shen, C.: Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. arXiv preprint arXiv:2303.11681 (2023) 4, 10, 11
53. Wu, Z., Wang, L., Wang, W., Shi, T., Chen, C., Hao, A., Li, S.: Synthetic data supervised salient object detection. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 5557–5565 (2022) 4, 11, 12
54. Xie, S., Tu, Z.: Holistically-nested edge detection. In: Proceedings of the IEEE international conference on computer vision. pp. 1395–1403 (2015) 6
55. Xu, A., Vasileva, M.I., Dave, A., Seshadri, A.: Handsoff: Labeled dataset generation with no additional human annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7991–8000 (2023) 2, 4
56. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1155–1162 (2013) 3, 10
57. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3166–3173 (2013) 3, 10
58. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. arXiv preprint arXiv:2401.10891 (2024) 5
59. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6023–6032 (2019) 2
60. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017) 2
61. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023) 3, 6
62. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020) 9, 12