Supplementary Materials: 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

1 Visual Examples

We include more visual examples of the method performance for scenes with one main object (*c.f.* Fig. 3) and complex scenes combining real and generated objects (*c.f.* Fig. 4). Additionally, we demonstrate visually diverse samples on different object categories (*c.f.* Fig. 5).

2 Data Anonymization

To measure the efficiency of the method to anonymize sensitive data we check how often a replaced face can be matched with its original appearance using a face identification model. To this end, we use ArcFace [**?**] on the pairs of original and generated images from the COCO dataset, validating whether people can be re-identified after applying our method. As the model does not work on children, manually filter them out from the dataset, From 64115 images with people and 262465 faces in them, only 373 (0.14%) we re-indentified. Inspecting these cases (Fig. 7), we find that they are almost exclusively false positives of the ReID model.

3 Visual Quality

We measure the impact of our method's components on the visual quality of the generated images in Tab. 1. We calculate FID [**?**] and Inception Score [**?**] for every setup to measure both the diversity and realism of the generated samples. We find that constraining the model with edges and depth map has a small impact on the FID score, likely because more constraints inherently limit the diversity of the generated images.

4 Limitations and Future Work

The method's performance is bounded by the performance of the main components: the inpainting and the control net model. For small and occluded objects,

the inpainting model sometimes completely removes the object and inpaints with the background (*c.f.* Fig. 8). Figure 1 shows the impact of prompt engineering for the person category on the inpainting. Recent advancements in image generation [**?**] could help mitigate this as well as further close the gap with the real-world data distribution. The pipeline is invariant to the choice of generative model, so employing more advanced generative models has the potential to further improve performance (*c.f.* Fig. 2). Further, the method is limited to the domain of the large text to mage generative model (real-world scenes) and thus can not be applied to specific domains such as satellite or medical images.

Original "person" "person snowboarding"

Fig. 1: Person Prompts. Providing a more detailed text description for people improves the stability of the method.

Component	**FID↓**	**Inception Score↑**	F_{max}**↑**
full method	1.10	**50.86**	**0.892**
w/o edge and depth control	**0.92**	38.00	0.886
w/o edge control	1.41	37.35	0.890
w/o prompt engineering	1.76	38.31	0.889

Table 1: Visual Quality. Without edge- and depth control the method generates images of high quality. They do not, however, match the original annotations. Removing edge map guidance and prompt engineering significantly drops both quality and diversity.

Model	Dataset	all (excl. people) AP	AP$_{50}$	AP$_{75}$	all (excl. cars) AP	AP$_{50}$	AP$_{75}$
Deformable-DETR	Real	**39.3**	**60.1**	**41.5**	**38.5**	**59.9**	40.4
Deformable-DETR	Anonymized People	38.7	59.1	40.9	-	-	-
Deformable-DETR	Anonymized Cars	-	-	-	38.2	59.5	**40.6**

Table 2: Data Anonymization. Fully replacing a category with synthetic examples has almost no impact on the performance of other classes.

		ECCSD				DUTS-TE				DUT-OMRON				HKU-IS				HRSOD			
Model	**Data**	$F_{max}\uparrow$	**MAE**$\downarrow$	$S_m\uparrow$	$F_{avg}\uparrow$	$F_{max}\uparrow$	**MAE**$\downarrow$	$S_m\uparrow$	$F_{avg}\uparrow$	$F_{max}\uparrow$	**MAE**$\downarrow$	$S_m\uparrow$	$F_{avg}\uparrow$	$F_{max}\uparrow$	**MAE**$\downarrow$	$S_m\uparrow$	$F_{avg}\uparrow$	$F_{max}\uparrow$	**MAE**$\downarrow$	$S_m\uparrow$	$F_{avg}\uparrow$
U2Net	orig.	0.944	0.052	0.900	0.882	0.863	0.066	0.836	0.766	0.835	0.075	0.819	0.734	0.930	0.043	0.895	0.872	0.895	0.063	0.862	**0.816**
U2Net	ours	**0.948**	**0.047**	**0.908**	**0.894**	**0.874**	**0.061**	**0.844**	**0.781**	**0.848**	**0.069**	**0.827**	**0.746**	**0.935**	**0.040**	**0.900**	**0.881**	**0.901**	**0.062**	**0.863**	**0.816**
F3-Net	orig.	0.955	0.035	0.924	0.917	0.899	0.038	0.887	0.841	0.831	**0.054**	0.835	0.752	0.942	**0.029**	0.918	0.904	0.916	0.038	0.904	0.876
F3-Net	ours	**0.962**	**0.033**	**0.929**	**0.924**	**0.907**	**0.037**	**0.890**	**0.846**	**0.850**	0.055	**0.843**	**0.765**	**0.947**	**0.029**	**0.922**	**0.909**	**0.926**	**0.037**	**0.907**	**0.879**
TRACER-4	orig.	0.956	0.027	0.929	0.931	0.911	0.029	0.896	0.867	0.847	0.048	0.848	0.786	0.944	0.024	0.921	0.916	0.933	0.025	0.918	0.905
TRACER-4	ours	**0.960**	**0.026**	**0.933**	**0.938**	**0.918**	**0.026**	**0.905**	**0.883**	**0.848**	**0.045**	**0.853**	**0.794**	**0.948**	**0.022**	**0.928**	**0.926**	**0.936**	**0.024**	**0.923**	**0.911**

Table 3: Full Salient Object Detection Evaluation. Comparison of performance on five benchmark datasets and four metrics. Our method generates samples in a higher than native resolution, improving performance on High-Resolution SOD benachmarks [?].

Original Stable Diffusion 1.5 Stable Diffusion XL

Fig. 2: Stronger Generative Model - Better Performance. Recent advancements in image generation models open new possibilities to utilize synthetic data for vision tasks. In our pipeline, more advanced diffusion models generate more realistic and detailed objects in higher resolution.

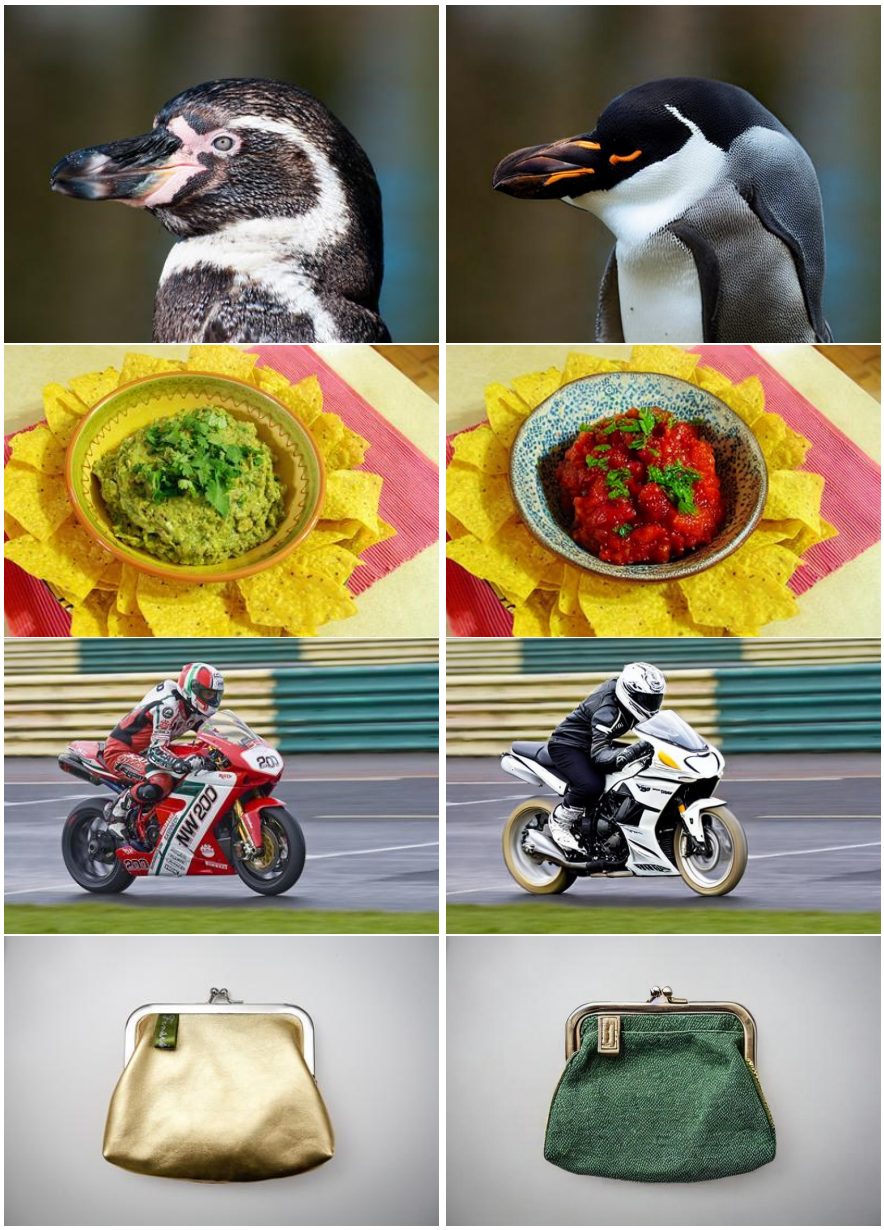
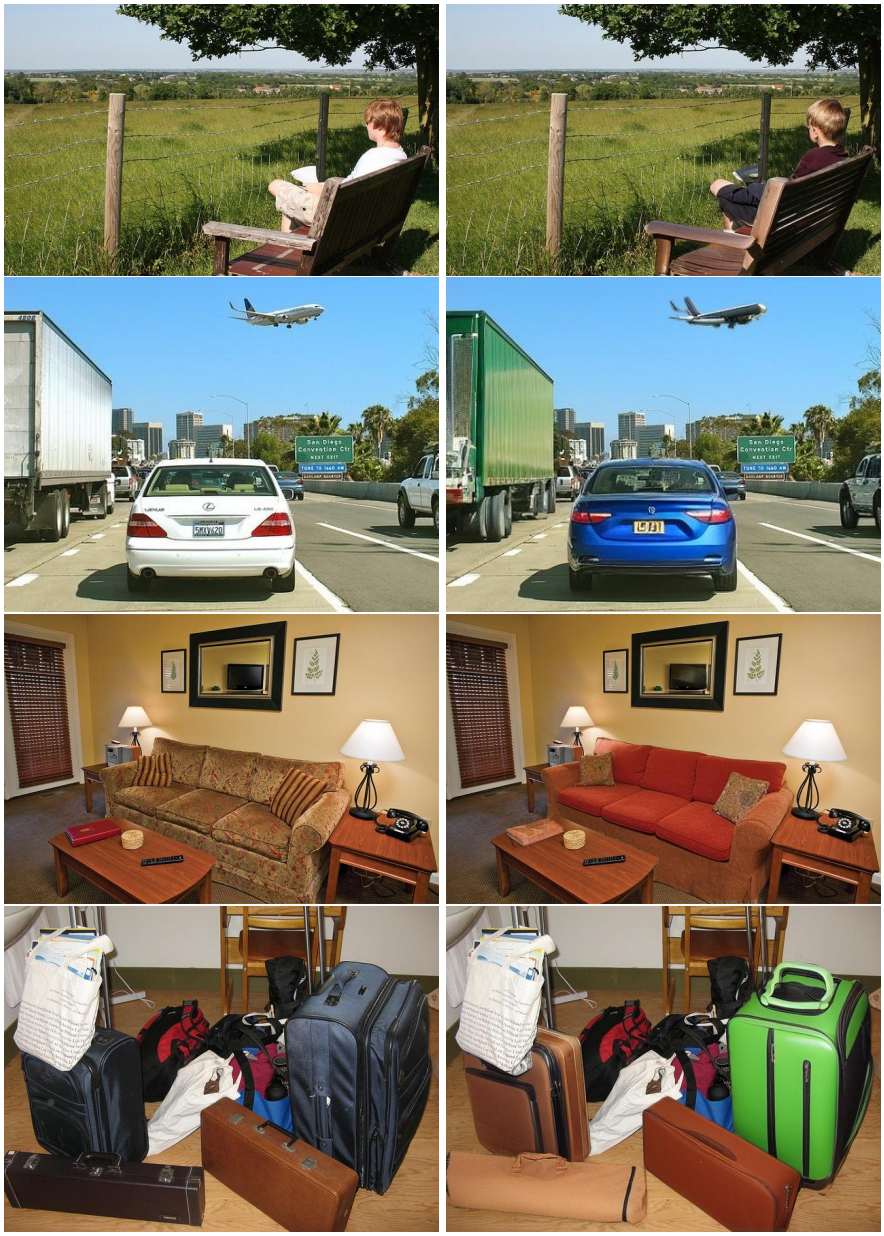


Fig. 3: Qualitative Examples. Single Object Augmentation on DUTS.



Original Augmented

Fig. 4: Complex Scenes. Mixing real and generated objects.

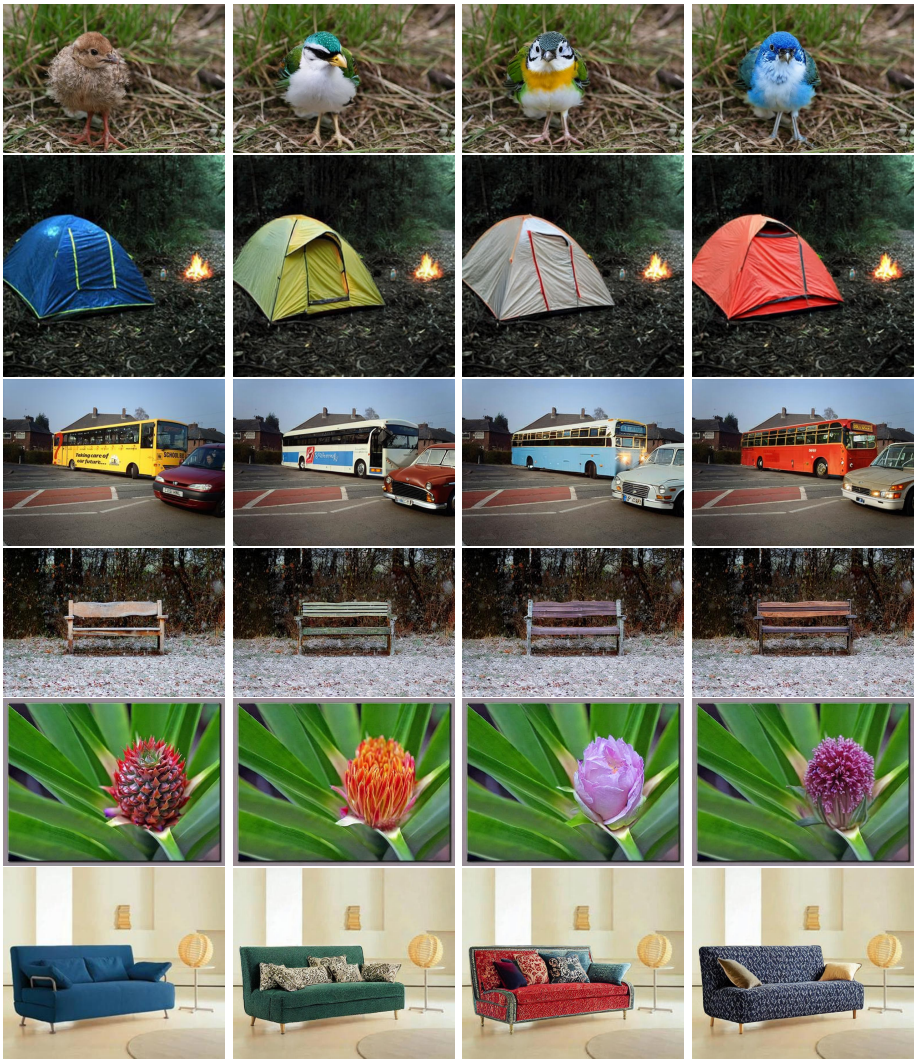
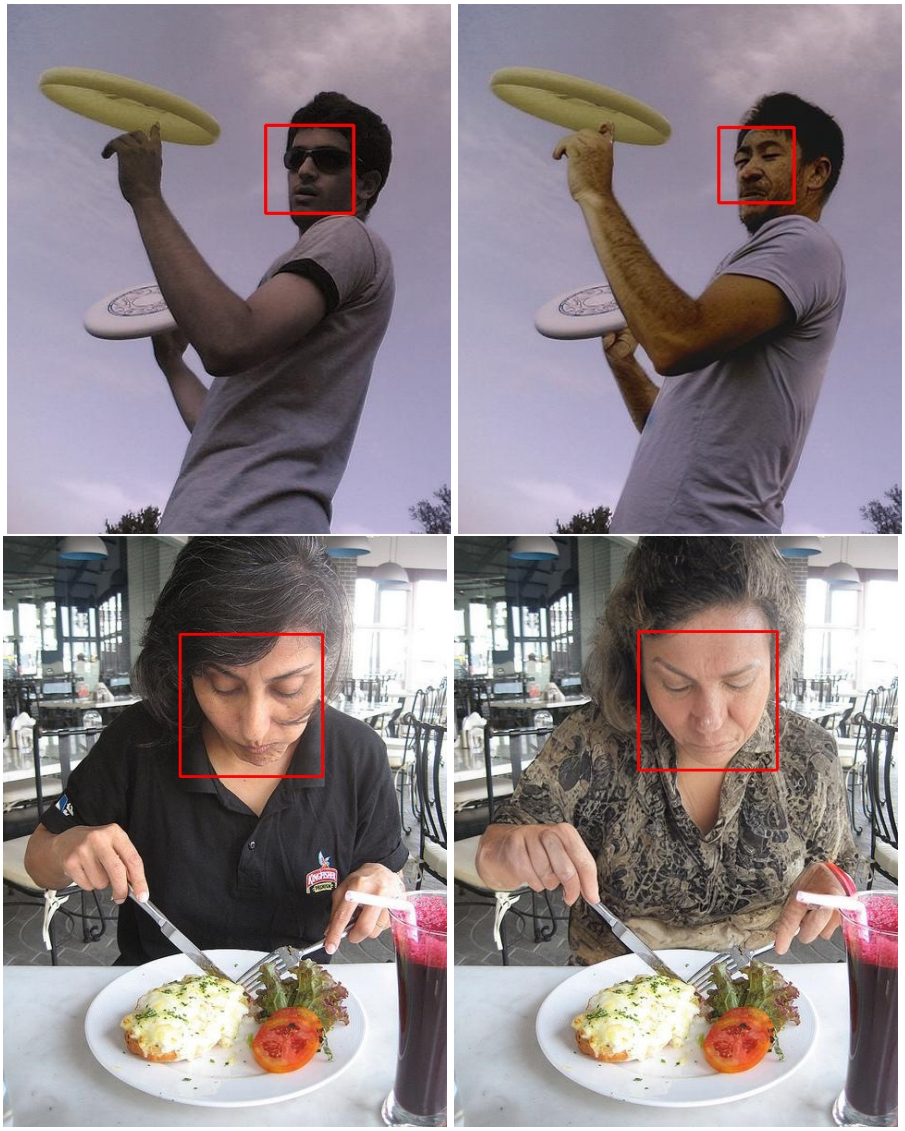


Fig. 5: Diversity. Showing three generated examples from the original (left) on DUTS.



Original Anonymized

Fig. 7: ReIDentified Faces. Red squares indicate faces that could be matched between the original and generated images. These are likely false positives of the ReID model.

Fig. 8: Failure Cases. Examples of removed objects and lower-quality generations. Red ellipses mark objects that were not repainted correctly.