Supplementary Material

A Softmax Temperature Ablation

In Fig. 1, we ablate different values for τ for $\sigma_\tau(\cdot)$. Higher values result in smoother distributions, while lower values result in sharper distributions. We show that the sweet-spot for this parameter is at 0.01.

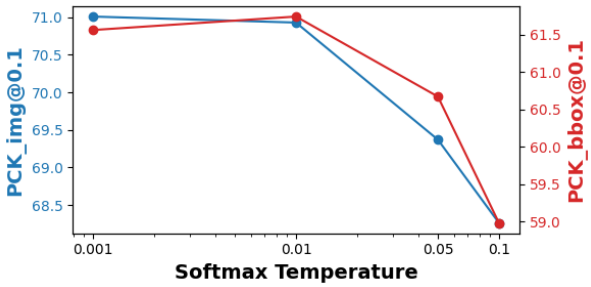

Figure 1: **Ablation of the softmax temperature parameter τ, evaluated on SPair-71k.** Trained for 10 epochs on COCO with retrieval sampling.

B 3D Threshold Ablation

Fig. 2 shows the effect of the threshold parameter ϵ on our 3D data augmentation method. Smaller values tend to exclude more points that are on the visible surface, whereas larger values tend to include too many points that are not on the visible surface. We set this parameter to 0.01 as it shows a good balance between those to extremes.

C Model Analysis

In Tab. 1 and Tab. 2 we ablate different Diffusion- and ViT-based models to find the best combination. We show that SDXL Turbo and DINOv2 (vitb14) with registers are the best performing models in our evaluation. In Tab. 3 we ablate different combinations of models to find the best performing setting. We show that the combination of the best performing models in the individual ablation, are also the best performing combination in general. Increasing the input resolution and adding additional layers further boosts the performance. In Table 4 we ablate the use of an additional strong teacher model, namely CLIP. However, we did not find any improvement in adding this model to the teacher ensemble.

Method	PCK_{img}@0.1	PCK_{bbox}@0.1
DINOv2 + SD	**71.77**	**63.29**
DINOv2 + SD + CLIP	<u>68.87</u>	<u>60.17</u>

Table 4: **Performance on SPair-71k for different teachers.** Adding CLIP to the teacher ensemble does not improve performance.

D Foreground Segmentation

We assess our model on other downstream tasks, including zero-shot foreground/ background segmentation. The examples in Figure 3 show a marginal improvement in mask quality. The masks generated with our model are slightly less noisy compared to the baseline DINOv2 model.

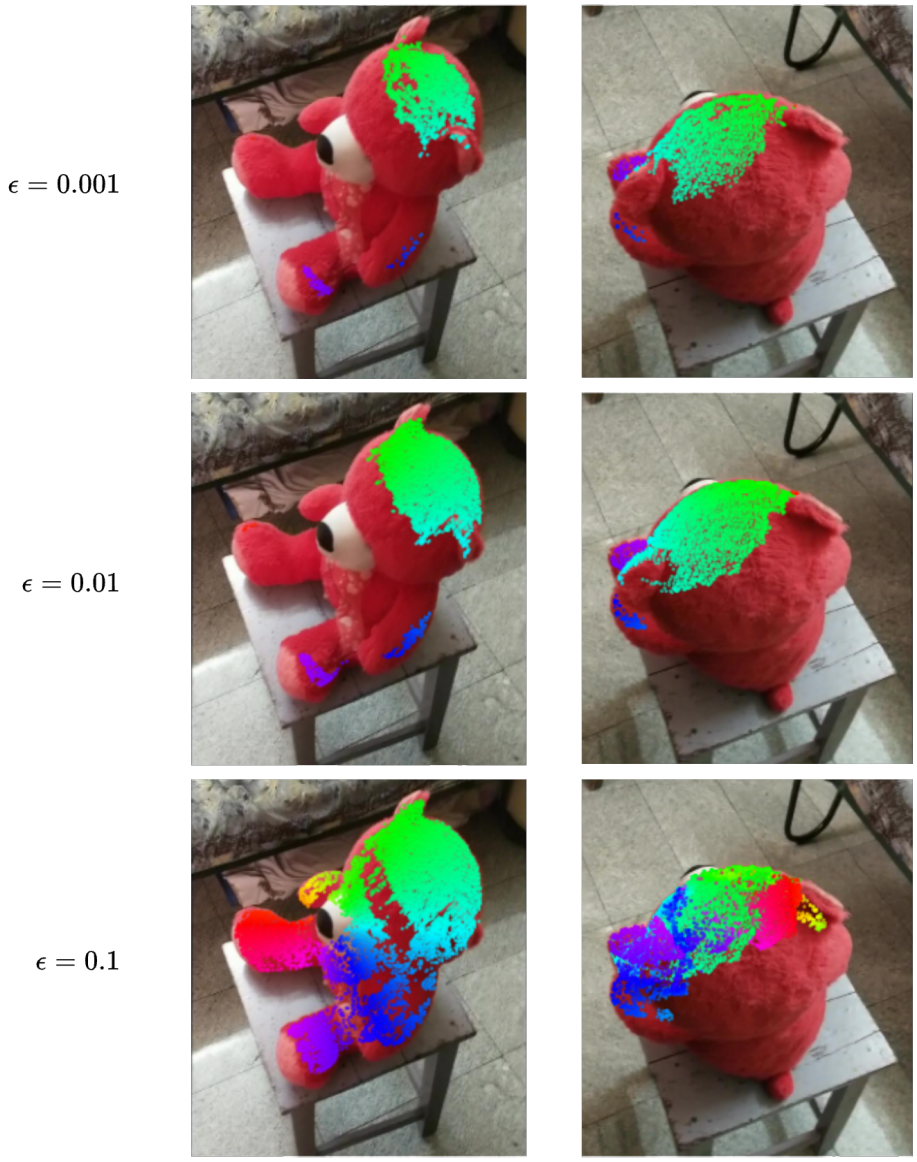

Figure 2: **The effect of the 3D data threshold parameter ϵ.**

Model	SPair-71K	PF-WILLOW	CUB-200	S	T	L
SD1.5	66.11/56.24	86.58/73.60	90.58/79.18	768^2	201	5
SD2.1	65.29/57.87	87.18/74.83	88.63/78.23	768^2	261	8
SDXL Base	64.02/55.52	88.37/76.49	92.39/84.20	768^2	101	1
SDXL Base	65.64/57.87	88.58/76.30	92.41/84.20	1024^2	201	1
LCM-XL	62.9/54.5	86.52/73.81	92.59/84.40	768^2	64	1
SDXL Turbo	67.26/58.54	**89.59/77.76**	93.54/85.57	768^2	101	1
SDXL Turbo	**67.40/59.50**	88.48/76.44	**93.35/85.72**	1024^2	101	1

Table 1: **The performance of different diffusion-based models evaluated on different datasets.** Values are measured in PCK@0.1 (img/bbox), per keypoint and averaged over all keypoints. S: Size of the input image, T: Timestep, L: Layer. Prompt for all models: "a photo of a [category]".

Model	SPair-71K	PF-WILLOW	CUB-200	R	L
CLIP (ViT-L-14)	47.05/37.05	73.51/57.67	82.31/67.86	336^2	11
MAE (ViT-L-14)	33.26/23.99	73.04/56.54	64.25/45.04	224^2	26
ZoeDepth	12.80/6.63	38.47/25.93	22.90/9.75	512×384	10 (BeiT)
I-JEPA (ViT-H-16 448)	51.88/44.78	−/−	−/−	448^2	31
DINOv1 (ViT-S-8)	46.69/35.92	61.66/47.99	84.06/70.09	224^2	9
DINOv2 (ViT-B-14)	67.45/57.69	**84.14/68.78**	94.54/85.90	840^2	11
DINOv2R (ViT-B-14)	**69.10/58.83**	83.07/67.38	**94.61/85.90**	840^2	11

Table 2: **The performance of different ViT-based models evaluated on different datasets.** Values are measured in PCK@0.1 (img/bbox), per keypoint and averaged over all keypoints. S: Size of the input image, L: Layer.

Model	SPair-71K	PF-WILLOW	CUB-200	S	T	L
SD1.5 + DINOv2	71.57/62.03	89.02/75.94	94.43/85.27	840^2	201	5 + 11
SD1.5 + DINOv2	71.38/62.08	88.84/75.70	94.24/85.69	840^2	201	3, 7, 11 + 11
SD1.5 + DINOv2	71.67/63.08	88.43/74.84	94.55/86.25	960^2	100	3, 7, 11 + 11
SDXL Turbo + DINOv2	70.90/61.88	**89.77/76.62**	94.89/86.45	840^2	101	1 + 11
SDXL Turbo + DINOv2	71.21/62.79	88.03/74.76	94.22/85.81	840^2	101	1, 4, 7 + 11
SDXL Turbo + DINOv2	**71.77/63.29**	89.36/75.98	**94.83/87.43**	980^2	101	1 + 11

Table 3: **The performance of different combinations of models and layers evaluated on different datasets.** Values are measured in PCK@0.1 (img/bbox), per keypoint and averaged over all keypoints. With DINOv2, we mean DINOv2 (ViT-B-14) with registers. S: Size of the input image, T: Timestep, L: Layer. Prompt for all models: "a photo of a [category]".

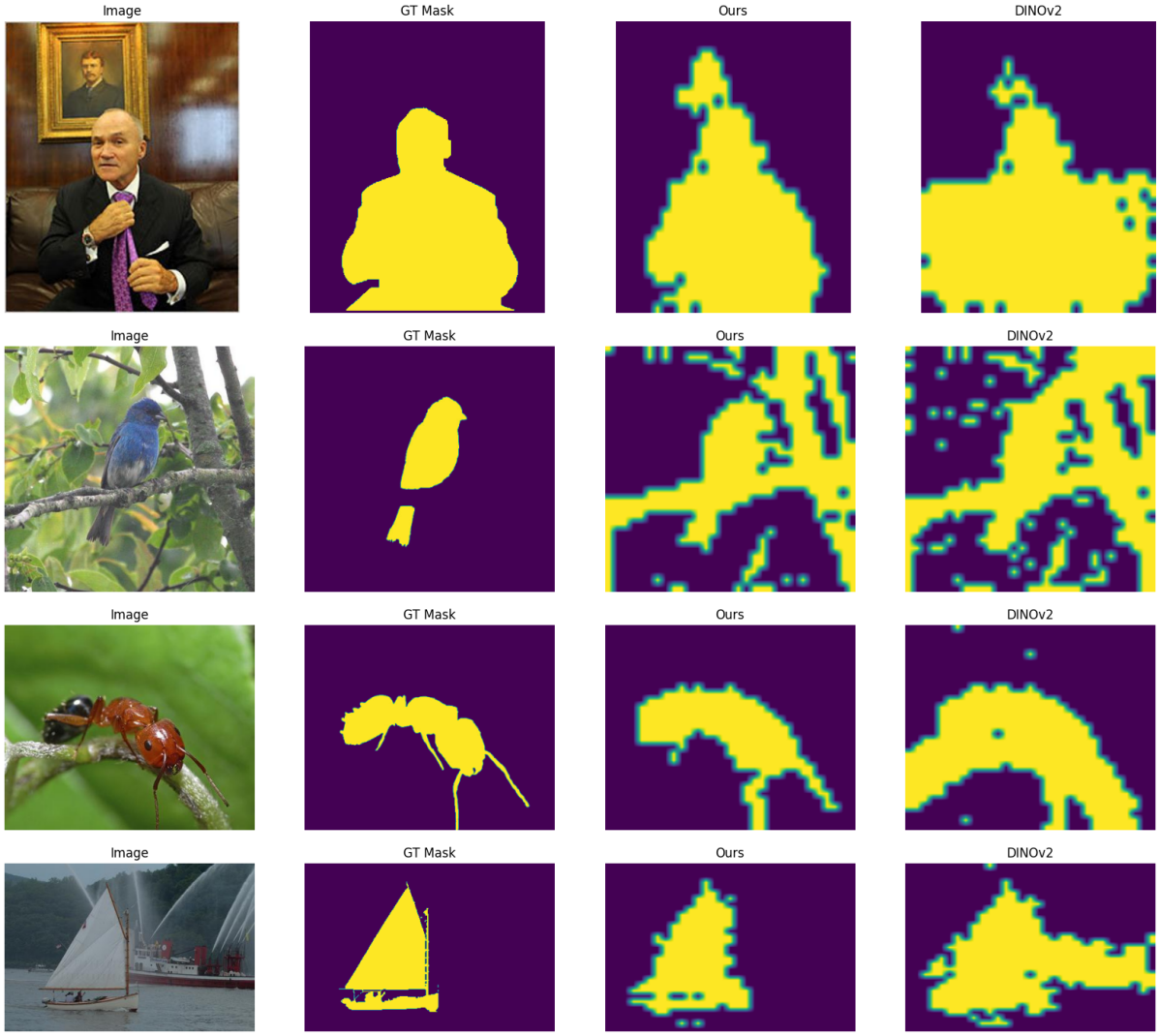

Figure 3: **Examples of the improved foreground/background segmentation masks with our model.**