NEURAL INFORMATION PROCESSING SYSTEMS

Where's Waldo: Diffusion Features for Personalized Segmentation and Retrieval

Dvir Samuel[1,2], Rami Ben-Ari[2], Matan Levy[3], Nir Darshan[2], Gal Chechik[2,4]

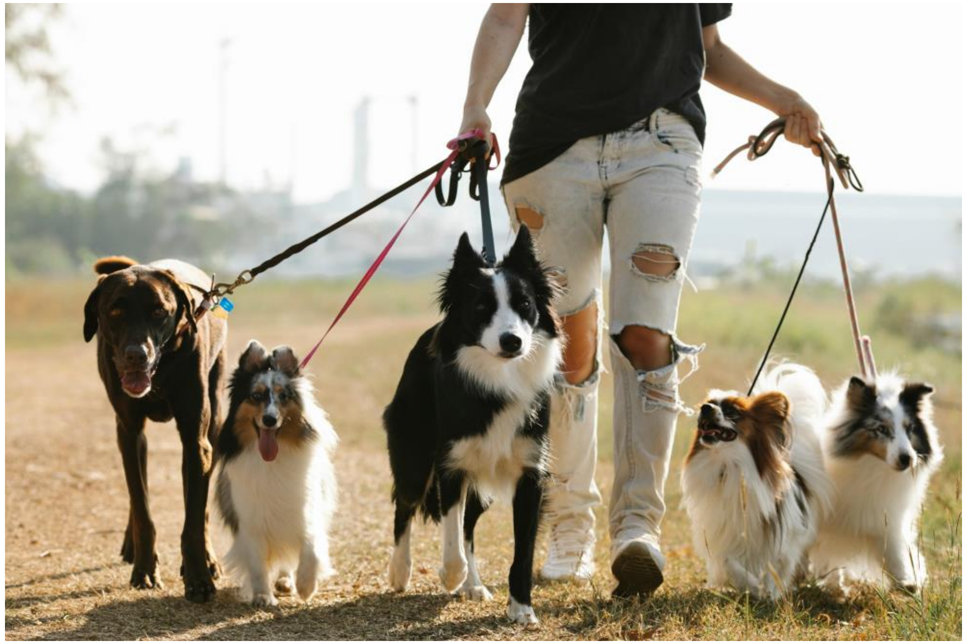
[1]Bar-Ilan University, Ramat-Gan, Israel
[2]OriginAI, Tel-Aviv, Israel
[3]The Hebrew University of Jerusalem, Jerusalem, Israel
[4]NVIDIA Research, Tel-Aviv, Israel

OriginAI NVIDIA

Personalized Segmentation and Retrieval

Personalized retrieval and segmentation focus on identifying and segmenting specific instances within a dataset.

Identifying a specific product in a catalog

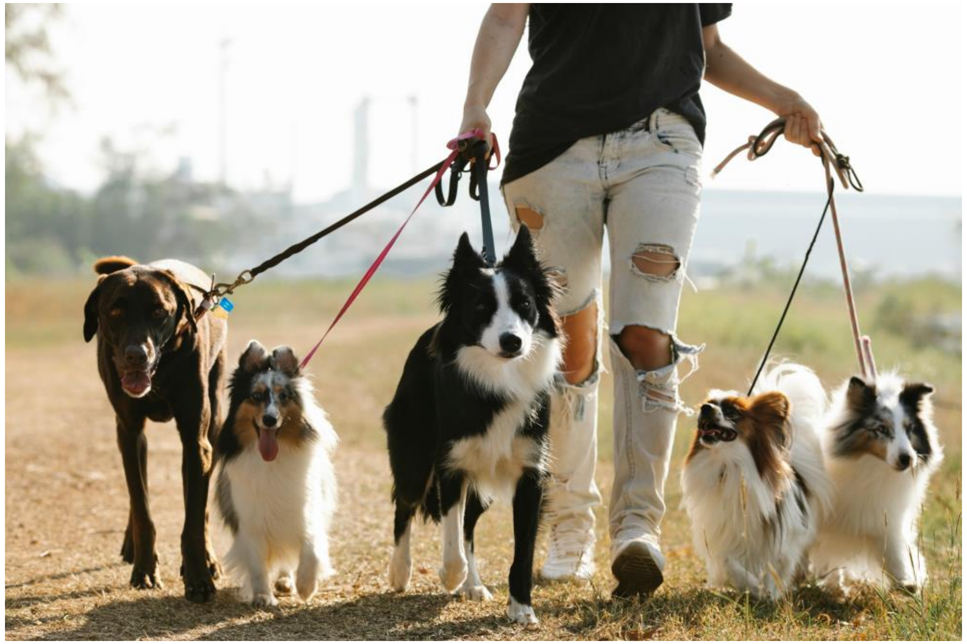


Tracking your beloved dog in images that contain multiple similar dogs

Personalized Segmentation and Retrieval

Personalized retrieval and segmentation focus on identifying and segmenting specific instances within a dataset.

Identifying a specific product in a catalog



Tracking your beloved dog in images that contain multiple similar dogs

Personalized Segmentation and Retrieval

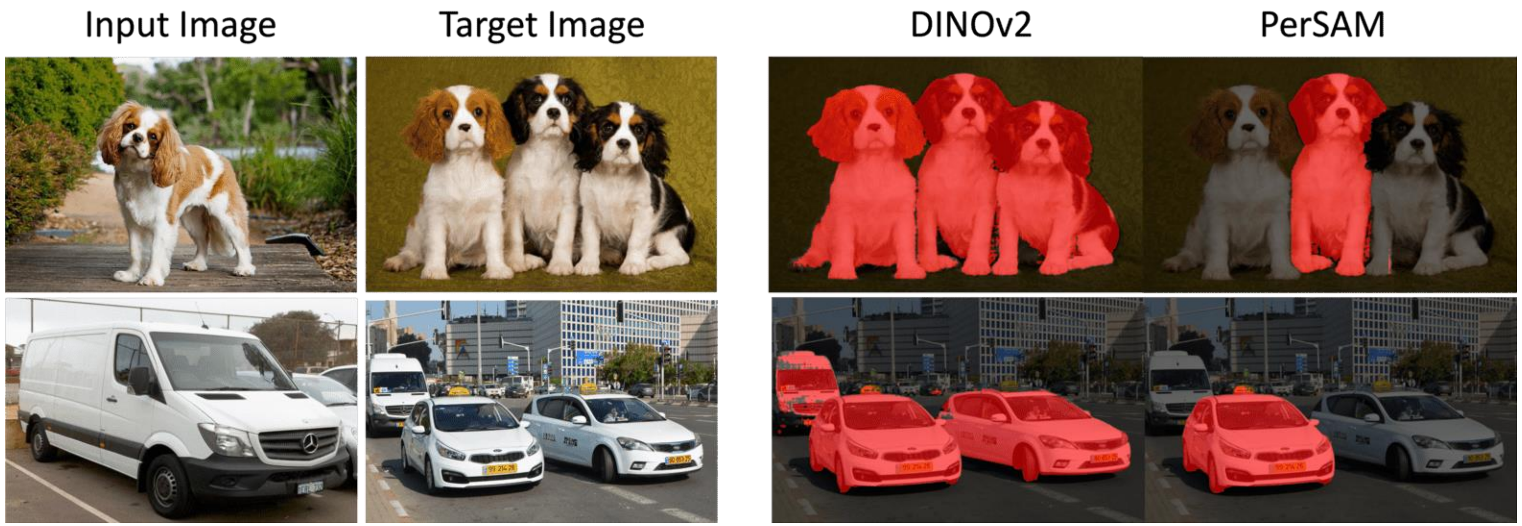
Personalized retrieval and segmentation focus on identifying and segmenting specific instances within a dataset.

Identifying a specific product in a catalog

Tracking your beloved dog in images that contain multiple similar dogs

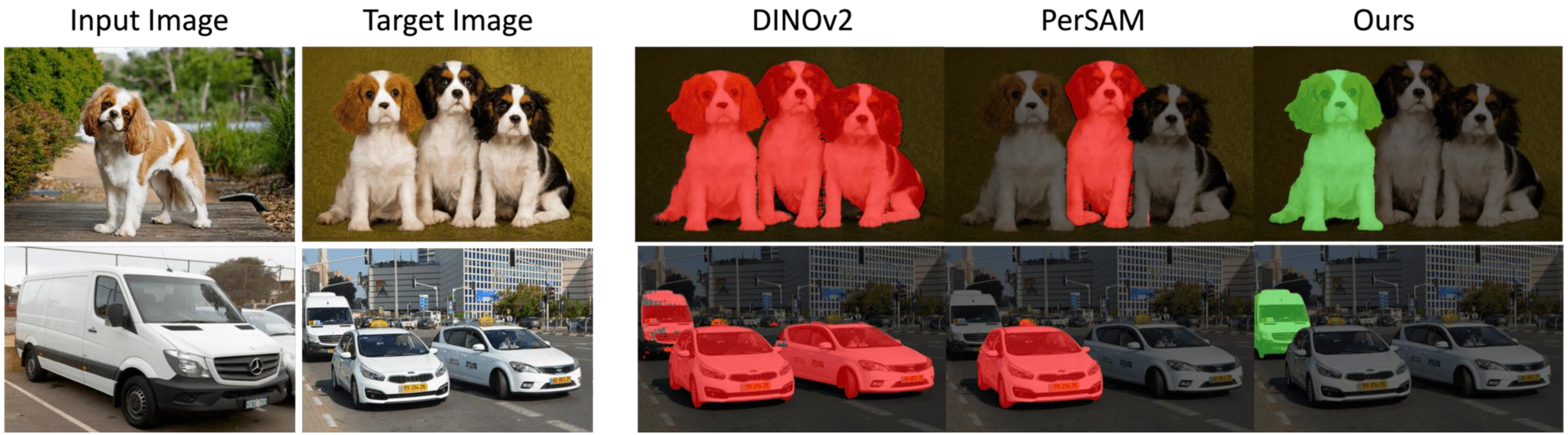
Current Methods

- Traditional supervised approaches are accurate but often rely on large amounts of domain-specific labeled data.
- Self-supervised methods (DINOv2 and PerSAM) have good discriminative performance between categories, however they struggle with multiple similar instances.



Current Methods

- Traditional supervised approaches are accurate but often rely on large amounts of domain-specific labeled data.
- Self-supervised methods (DINOv2 and PerSAM) have good discriminative performance between categories, however they struggle with multiple similar instances.

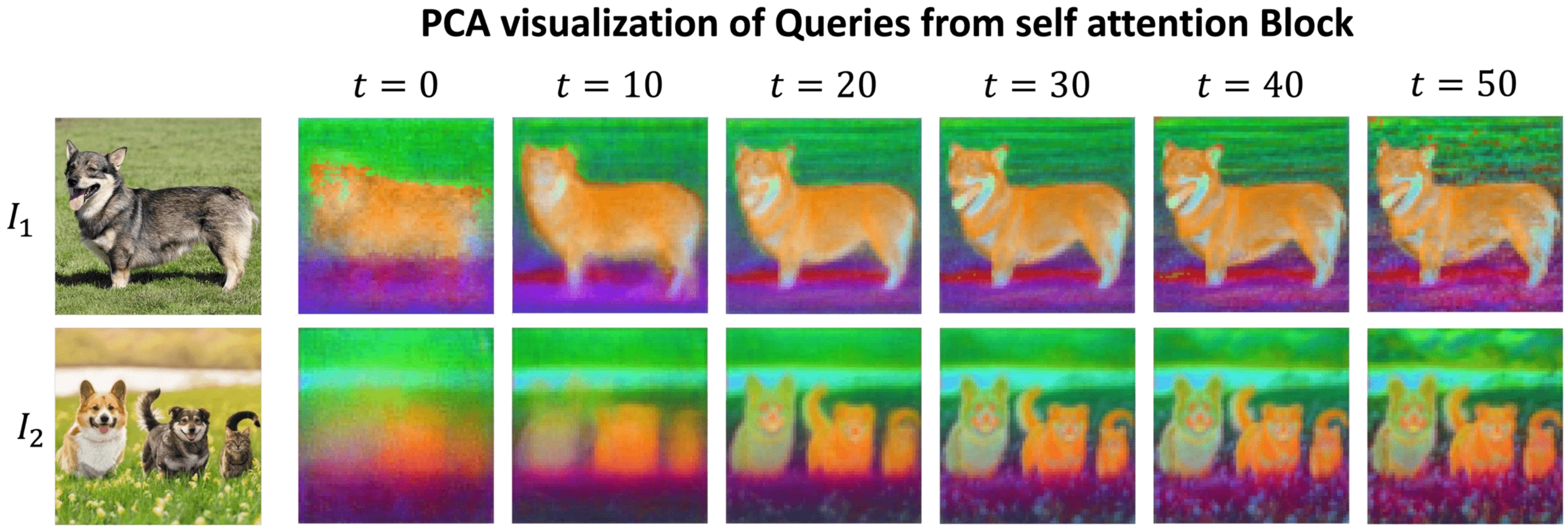


Motivation

- Text-to-image foundation models have achieved remarkable success in generating new and unique images from text prompts.
- We hypothesize that properties of generated objects are encoded within the intermediate features of the diffusion model during generation

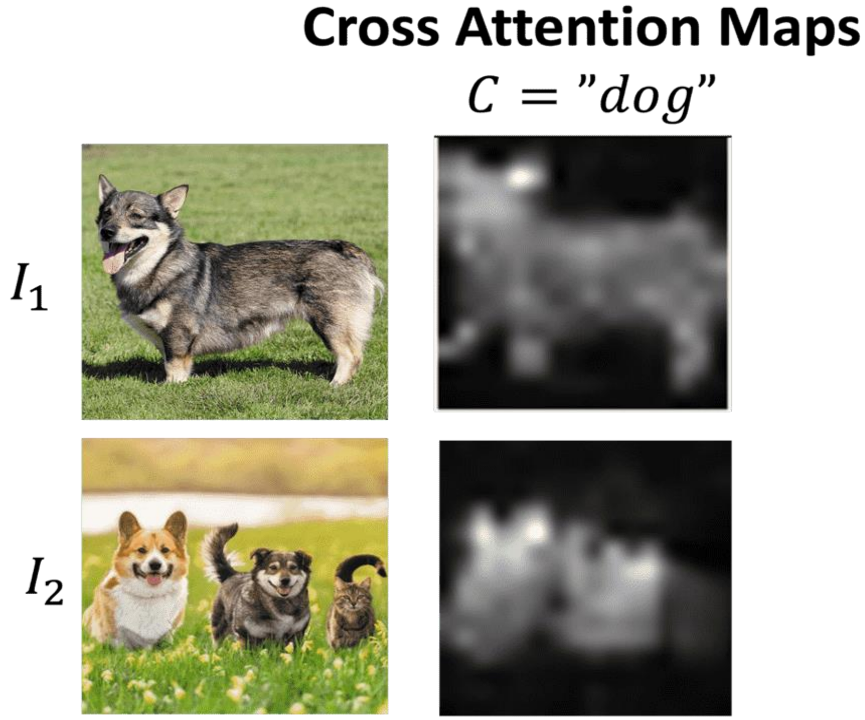
Are instance features encoded in diffusion models?

Appearance Features: We found that instance appearance features are encoded in the queries and keys matrices of the self-attention block.



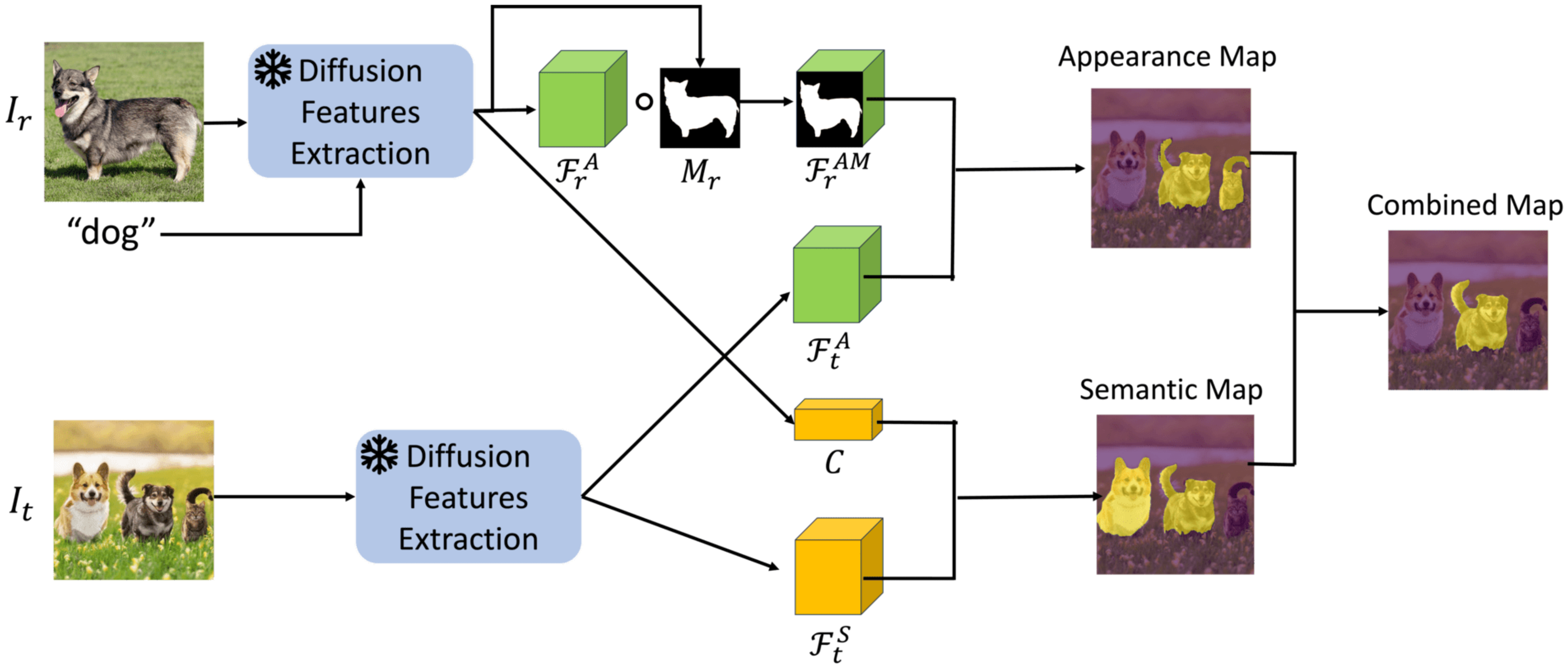
Are instance features encoded in diffusion models?

Semantic Features:
Cross-attention maps link the textual input prompt to image patches, creating a coarse semantic segmentation map that highlights potential object locations.



Personalized **D**iffusion Feature **M**atching (**PDM**)

PDM uniquely leverages diffusion features for instance-based tasks without extra training, combining both appearance and semantic characteristics of an instance.



Benchmark datasets

Existing benchmarks fall short as they tend to feature either a single, distinct object or multiple objects from different categories in each image.

$\mathcal{R}$Oxford

DAVIS

PerSeg

Benchmark datasets

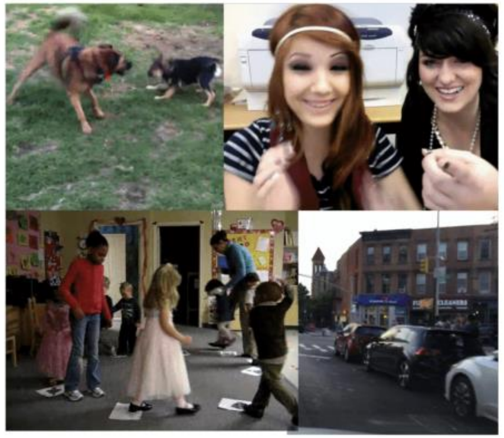
- **New Benchmarks**: PerMIR (Personalized Multi-Instance Retrieval) and PerMIS (Personalized Multi-Instance Segmentation).
- Built from BURST [1] Dataset with challenging multi-instance images.
- Highlights strengths of instance-based approaches.

$\mathcal{R}$Oxford

DAVIS

PerSeg

PerMI{R/S} (**Ours**)

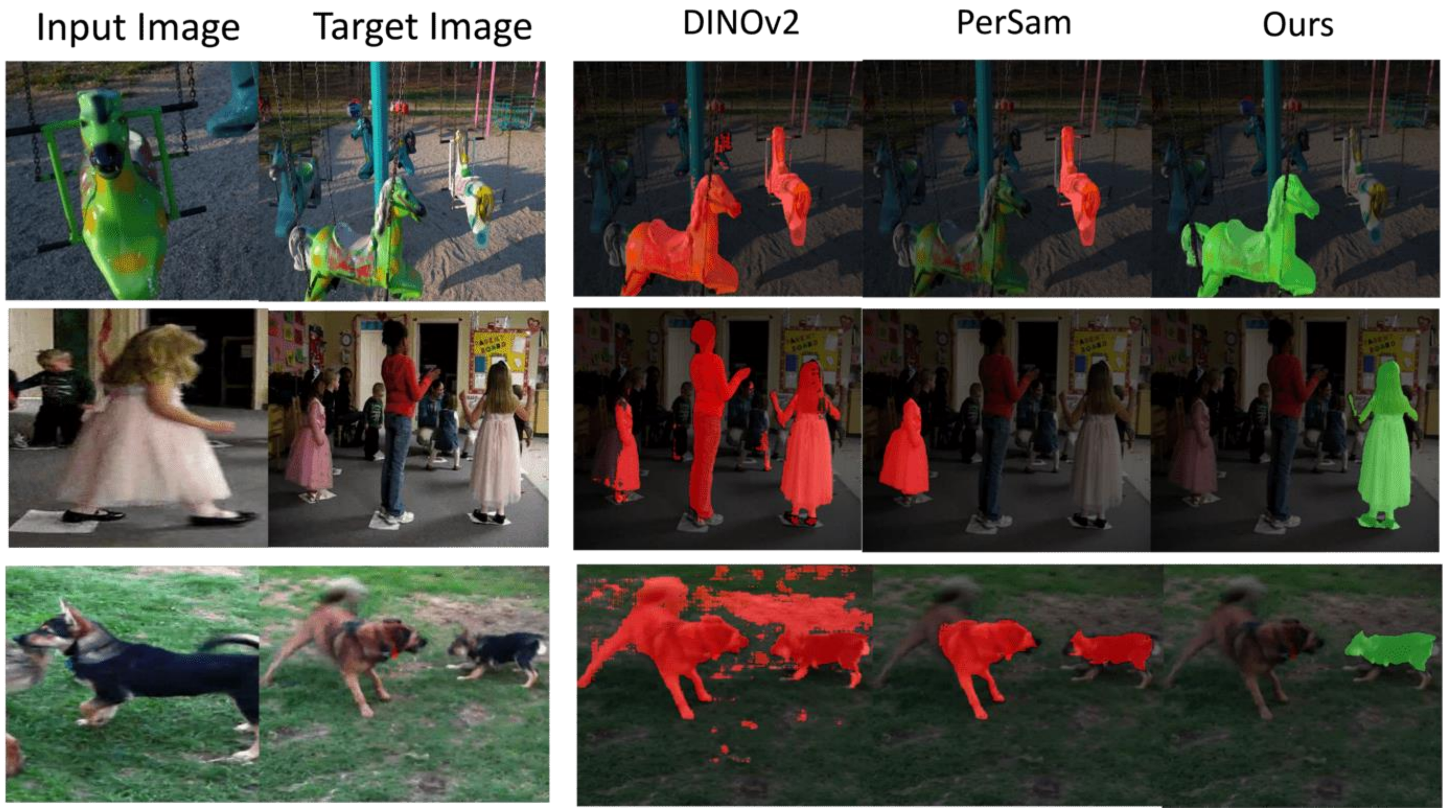


[1] Athar et al, BURST: A Benchmark for Unifying Object Recognition, Segmentation and Tracking in Video. (2022)

Results: Personalized Segmentation

PDM achieves top results on segmentation accuracy. Has high consistency in segmenting the exact instance among similar objects.

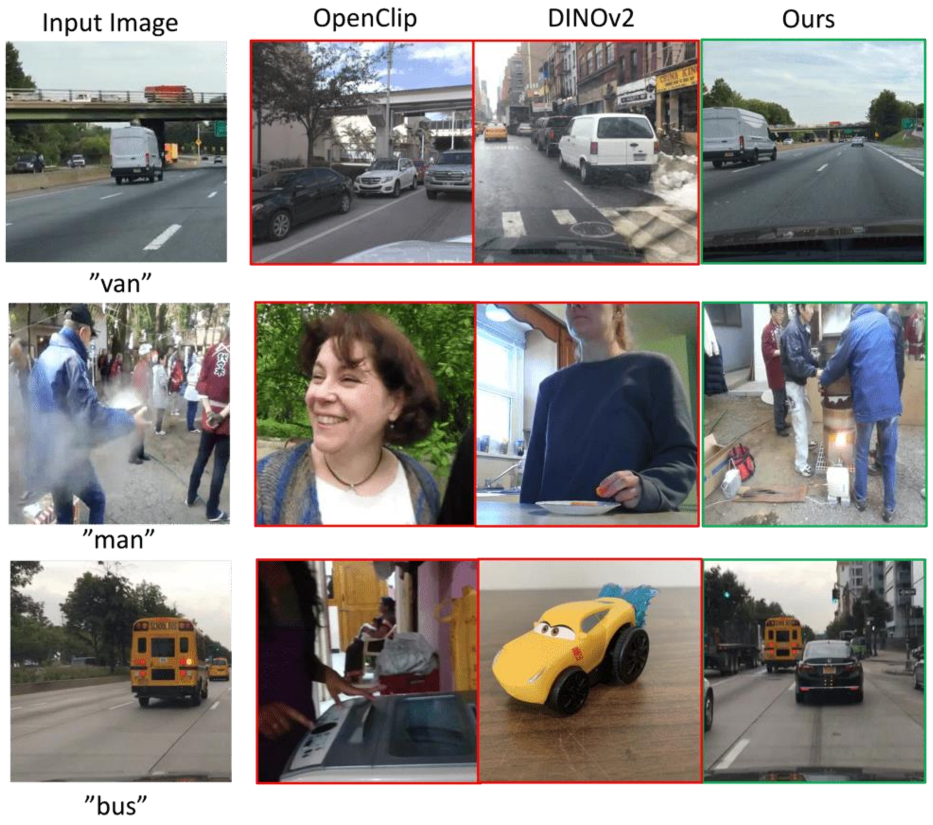
	Perosnalized Image Segmentation			
	PerSeg		**PerMIS (Image)**	
Model	mIoU	bIoU	mIoU	bIoU
SEEM [42]	87.1	55.7	14.3	35.8
SegGPT [36]	94.3	76.5	18.7	39.5
MAST [15]	-	-	-	-
SFC [12]	-	-	-	-
DINOv2 [18]	68.7	27.6	20.2	41.9
DIFT [31]	63.2	26.9	21.9	43.1
DiffSeg [33]	38.6	37.9	7.9	6.4
PerSAM(SAM) [39]	95.3	77.9	16.5	38.3
PDM (**ours**)	**95.4**	**79.8**	**42.3**	**86.8**
PerSAM(PDM) (**ours**)	**97.4**	**81.9**	**49.7**	**89.3**



Results: Personalized Retrieval

PDM outperforms self-supervised and weakly supervised baselines.
Significant gains in complex, multi-instance scenarios.

Methods	ROxford Medium	ROxford Hard	RParis Medium	RParis Hard	PerMIR
Self & Weakly Supervised					
MAE [11]	11.7	2.2	19.9	4.7	-
iBOT [40]	39.0	12.7	70.7	47.0	-
DINOv2 [18]	75.1	54.0	92.7	83.5	29.7
CLIP [24]	28.5	7.0	66.7	41.0	20.9
OpenClip [13]	50.7	19.7	79.2	60.2	26.7
PDM (**ours**)	**77.2**	**58.3**	**93.4**	**84.7**	**73.0**
OpenClip + PDM (**ours**)	**70.1**	**57.7**	**90.1**	**82.0**	**69.9**
DINOv2 + PDM (**ours**)	**80.4**	**62.1**	**93.6**	**85.1**	**70.8**
Supervised					
GSS [17]	80.6	64.7	93.4	85.3	-
HP [2]	85.7	70.3	92.6	83.3	-
SuperGlobal [29]	90.9	80.2	93.9	86.7	33.5
GSS + PDM (**ours**)	89.3	76.1	92.9	84.8	**62.0**
SuperGlobal + PDM (**ours**)	**91.2**	**80.3**	**94.0**	**86.8**	**69.1**



Thank you!