Implicit-Zoo: A Large-Scale Dataset of Neural Implicit Functions for 2D Images and 3D Scenes

Qi Ma, Danda Pani Paudel
Ender Konukoglu, Luc Van Gool

NEURAL INFORMATION PROCESSING SYSTEMS

Expand dataset in 2D and 3D INRs

The lack of large-scale INR datasets has hindered further advancements in INRs, as existing datasets are limited in scale and scope. To address this, we introduce ***Implicit-Zoo***, a dataset comprising over 1.5 million implicit functions across diverse 2D and 3D tasks.

- Creation of Implicit-Zoo: Developed using nearly 1,000 GPU days with iterative refinement for high-quality data (PSNR ≥ 30).

- Comprehensive Benchmarks: Tasks include 2D image classification, segmentation, and 3D pose regression with a novel baseline.

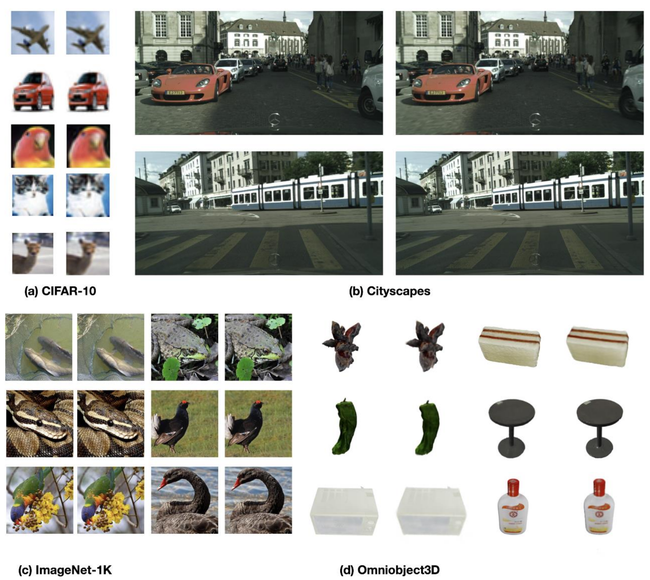
- Learnable Tokenization: Enhances benchmarks across tasks using adaptive patch centers, scales, and pixel-/point-level approaches.

This work introduces learnable tokenization as a novel research direction, showcasing its transformative impact on INR-based tasks.

Method	Task	Scenes	Model(Depth/Width)	GPU (days)	Overall Size (GB)	PSNR
CIFAR-10 [1]	2D	60000	3 / 64	5.96	1.44	31.01
ImageNet-1K [2]	2D	1431167	4 / 256	831.53	749.93	30.12
CityScapes [3]	2D	23473	5 / 256	50.15	18.40	34.10
Omniobject-3D [4]	3D	5914	4 / 128	69.81	5.96	31.54

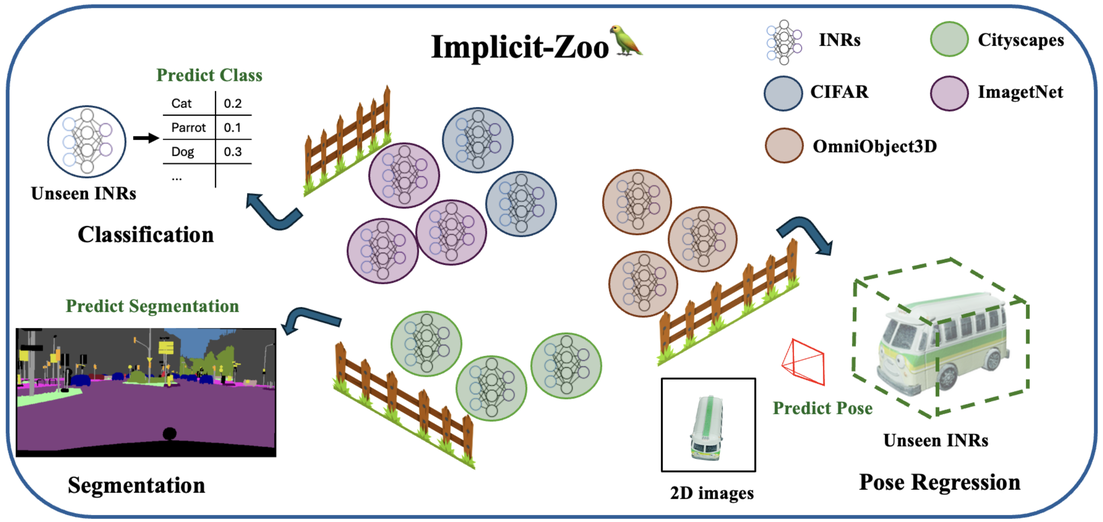
Example of Dataset

Optional section descriptor in 21pt font



INRs reconstruction exmaples

Application



Learnbale Token

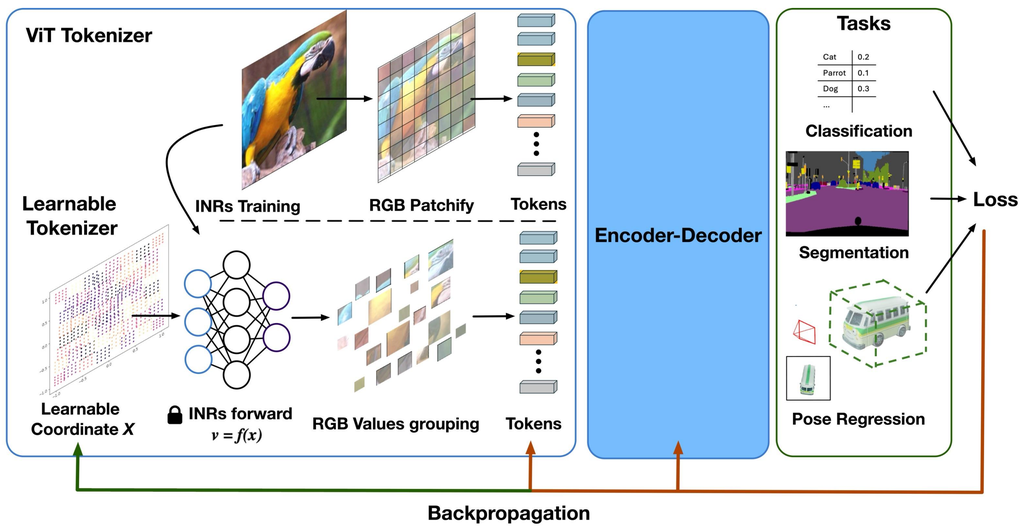
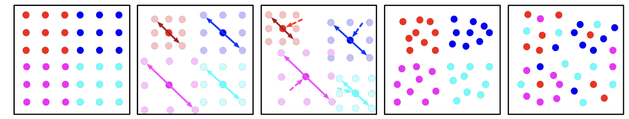
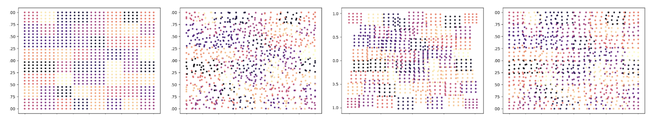


Illustration of learnable tokenizer. Instead of retrieve RGB value from images we query learnable coordinates to pre-trained freezed INRs and grouping RGB values to create tokens. Note that during backpropogation the Coordinate x will also be jointly optimized with ViT modules.

RGB Grouping





Experiment

Method	Acc ↑	Precision ↑	F1 ↑
ViT[18]	80.82±0.86%	80.76± 0.87 %	80.75± 0.86 %
ViT[18] + S	80.24± 0.47%	80.49± 0.63%	80.44± 0.57%
ViT[18] + LC	81.33± 0.23%	81.29± 0.22%	81.30± 0.23%
ViT[18] + LP + rand	59.43 ± 1.21 %	59.56 ±1.32 %	59.65 ±1.29%
ViT[18] + LP	79.51± 0.23%	79.37± 0.34%	79.37± 0.35%
ViT[18] + LP + Reg	**81.57± 0.29%**	**81.53± 0.30%**	**81.51± 0.30%**

Method	Acc ↑	Precision ↑	F1 ↑
VGG19[68]	82.06 ± 0.67 %	82.11± 0.71 %	82.07 ± 0.70 %
VG19+LP+Reg	**82.28 ± 0.45 %**	**82.30 ± 0.63 %**	**82.22 ± 0.59 %**
ResNet18[69]	83.34 ± 0.61 %	83.72 ± 0.67 %	83.48 ± 0.64 %
ResNet18+LP+Reg	**83.57 ± 0.59 %**	**83.94 ± 0.51%**	**83.71 ± 0.55 %**

Classification results on CIFAR-INR with various grouping methods. Using VGG11 and ResNet18, our learnable token approach enhances performance in both CNN architectures and supports learnable convolutions.