Generative Neural Fields by Mixtures of Neural Implicit Functions

Tackgeun You
POSTECH

Mijeong Kim
Seoul National University

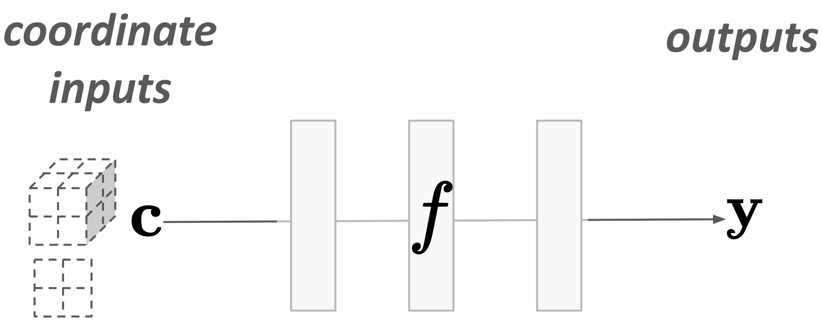
Jungtaek Kim
University of Pittsburgh

Bohyung Han
Seoul National University

NeurIPS 2023, New Orleans, USA

Implicit Neural Representation

- Implicit neural representations (or neural fields) represents a data using implicit function *f*.



$$\mathbf{y} = f(\mathbf{c}) = f^{(L+1)} \circ f^{(L)} \cdots \circ f^{(0)}(\mathbf{c})$$

$$\mathbf{h}^{(i+1)} = f^{(i)}(\mathbf{h}^{(i)}) = \mathrm{Act}(\mathbf{W}^{(i)}\mathbf{h}^{(i)} + \mathbf{b}^{(i)})$$

$$\mathbf{c} = \mathbf{h}^{(0)} \qquad \mathbf{y} = \mathbf{h}^{(L+2)}$$

$$\mathcal{C}(\mathbf{c}) = \sum_{(\mathbf{c}^{(i)}, y^{(i)}) \in \mathcal{D}} ||f(\mathbf{c}^{(i)}) - y^{(i)}||^2$$

Image

$$(c_x, c_y) \mapsto (y_R, y_G, y_B)$$

Voxel

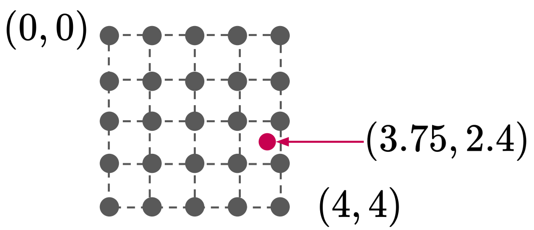
$$(c_x, c_y, c_z) \mapsto (y_\sigma)$$

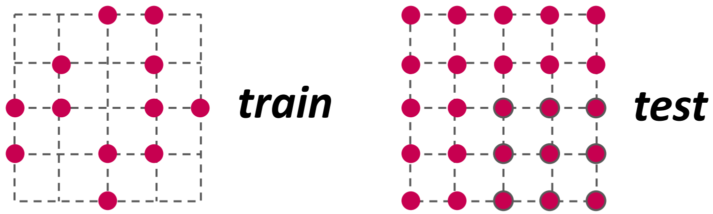
Radiance Fields

$$(c_x, c_y, c_z, c_\theta, c_\phi) \mapsto (y_R, y_G, y_B, y_\sigma)$$

Merits of INR

- Input coordinate queries on continuous grids
  - Input coordinate queries are independent each other.





- Compression ability
  - *"Our method [NeRF] requires only 5 MB for the network weights, which is even less memory than the input images alone for a single scene from any of our datasets."*
  - Neural Representation for Video [NeRV] compresses a video comparable to off-the-shelf video codecs, such as H.264.

[NeRF] Ben Mildenhall, et al., **Representing Scenes as Neural Radiance Fields for View Synthesis**, ECCV 2020
[NeRV] Hao Chen, et al., **NeRV: Neural Representations for Videos**, NeurIPS 2021

Limitation of INR

- Only memorizing a single object or scene

A single NeRF scene

- Generation 3D scene is tricky due to a large parameter space

→ Motivation for generative mechanism for INR or neural fields

[Gaussian Splatting] 3D Gaussian Splatting for Real-Time Radiance Field Rendering, SIGGRAPH 2023

Motivation

- Model Averaging for Implicit Neural Representation
    - Simple conditioning mechanism for INR
    - Increasing capacity of generative neural fields without additional inference costs

Mixtures of INR weights within parameter space
for compact generative neural fields

Ensemble of Model Weights

- Ensembling weights within parameter space
  - Several evidences from distinct tasks and architectures, such as domain generalization [SWA], generative model [EMA-GAN] and generic transfer learning [ModelSoup]
  - No requirement for additional inference cost

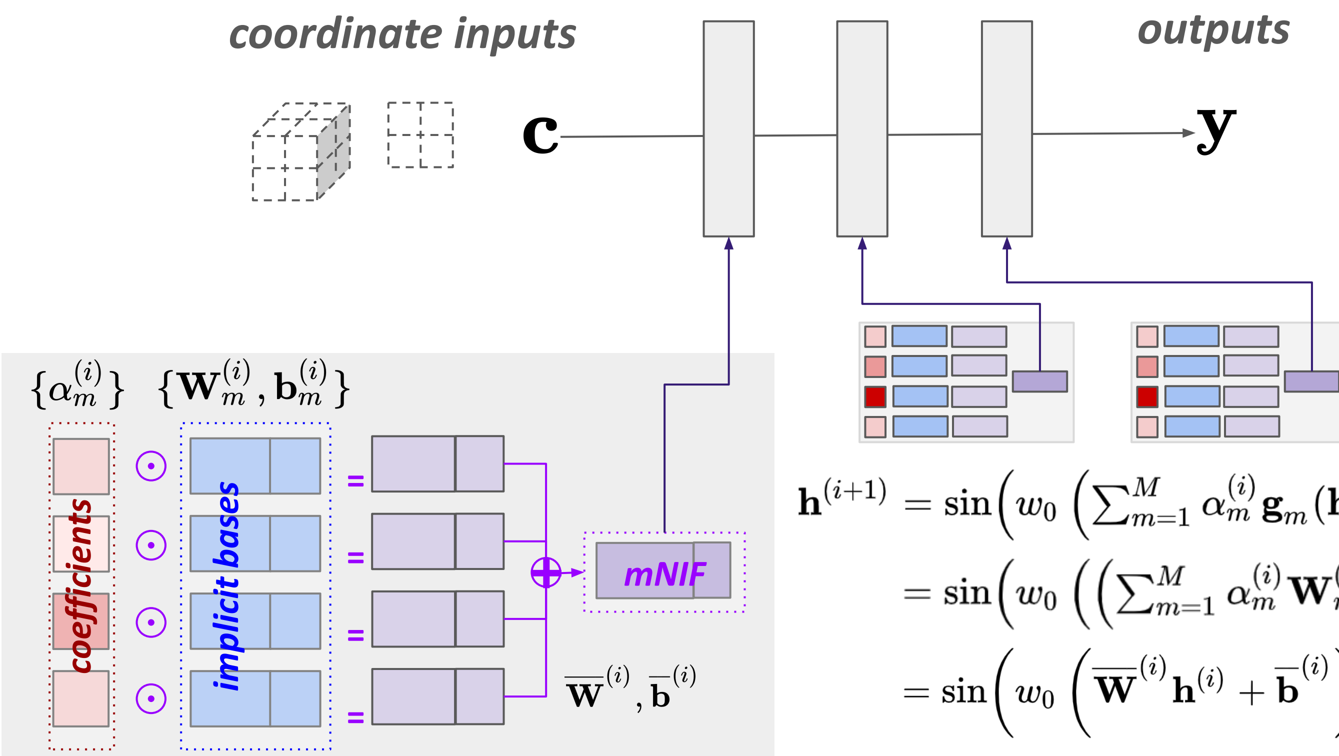
$$\mathbf{y} = f(\mathbf{x}; \boldsymbol{\theta}) \quad \Longrightarrow \quad \mathbf{y} = f(\mathbf{x}; \overline{\boldsymbol{\theta}})$$

$$\overline{\boldsymbol{\theta}} = \sum_{m=1}^{M} \alpha_m \cdot \boldsymbol{\theta}_m$$

[SWA] Pavel Izmailov, et al., **Averaging Weights Leads to Wider Optima and Better Generalization**, UAI 2018
[EMA-GAN] Yasin Yazıcı, et al., **The Unusual Effectiveness of Averaging in GAN Training**, ICLR 2019
[ModelSoup] Mitchell Wortsman, et al., **Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time**, ICML 2022

Mixture of Neural Implicit Functions

- Mixtures of neural implicit functions

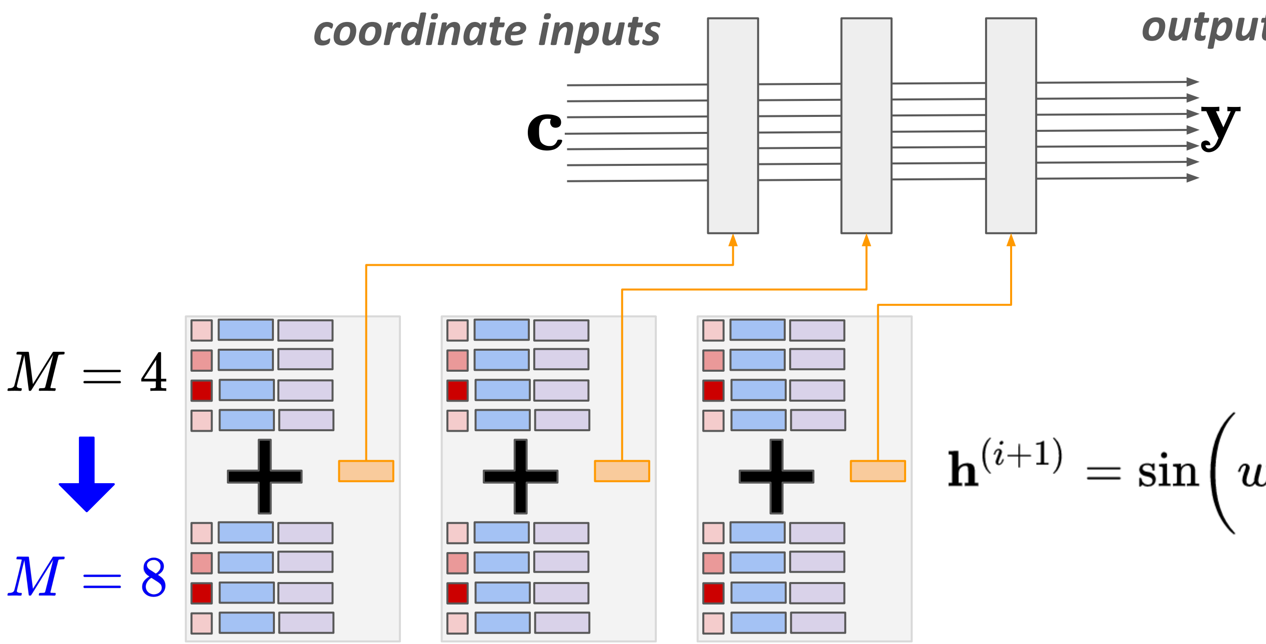


$$
\begin{aligned}
\mathbf{h}^{(i+1)} &= \sin\left(w_0 \left(\sum_{m=1}^{M} \alpha_m^{(i)} \mathbf{g}_m(\mathbf{h}^{(i)})\right)\right) \\
&= \sin\left(w_0 \left(\left(\sum_{m=1}^{M} \alpha_m^{(i)} \mathbf{W}_m^{(i)}\right) \mathbf{h}^{(i)} + \sum_{m=1}^{M} \alpha_m^{(i)} \mathbf{b}_m^{(i)}\right)\right) \\
&= \sin\left(w_0 \left(\overline{\mathbf{W}}^{(i)} \mathbf{h}^{(i)} + \overline{\mathbf{b}}^{(i)}\right)\right)
\end{aligned}
$$

$M = 4$

Merits of mNIF

- Easy to increase the capacity of generative INR
- Efficient computation for numerous coordinate queries
- Compact size of neural field instance



independent from the number of mixtures (M)

$$\mathbf{h}^{(i+1)} = \sin\left(w_0 \left(\overline{\mathbf{W}}^{(i)} \mathbf{h}^{(i)} + \overline{\mathbf{b}}^{(i)} \right) \right)$$

Benchmarks

- Unconditioned generation of Image, Voxel, Radiance Fields

Image

$(c_x, c_y) \mapsto (y_R, y_G, y_B)$

- CelebA-HQ 64^2
- Train (27,000) Test (3,000)
- Protocol from [Functa]
- Fréchet Inception Distance (FID)

Voxel

$(c_x, c_y, c_z) \mapsto (y_\sigma)$

- ShapeNet 64^3
- Train (35,019) Test (8,762)
- Protocol from [GEM]
- Coverage & Maximum Mean Discrepancy (MMD) with Chamfer distance

Radiance Fields

$(c_x, c_y, c_z) \mapsto (y_R, y_G, y_B, y_\sigma)$

- SRN Cars with 128^2 pixels
- Train (2,458) Test (704)
- Protocol from [Functa]
- Simplified NeRF rendering without ray direction (θ, ϕ)
- FID of images from pre-defined 251 views

[Functa] Emilien Dupont, et al., **From data to functa: Your data point is a function and you can treat it like one**, ICML 2022
[GEM] Yilun Du, et al., **Learning Signal-Agnostic Manifolds of Neural Fields**, NeurIPS 2021

Performance on Image

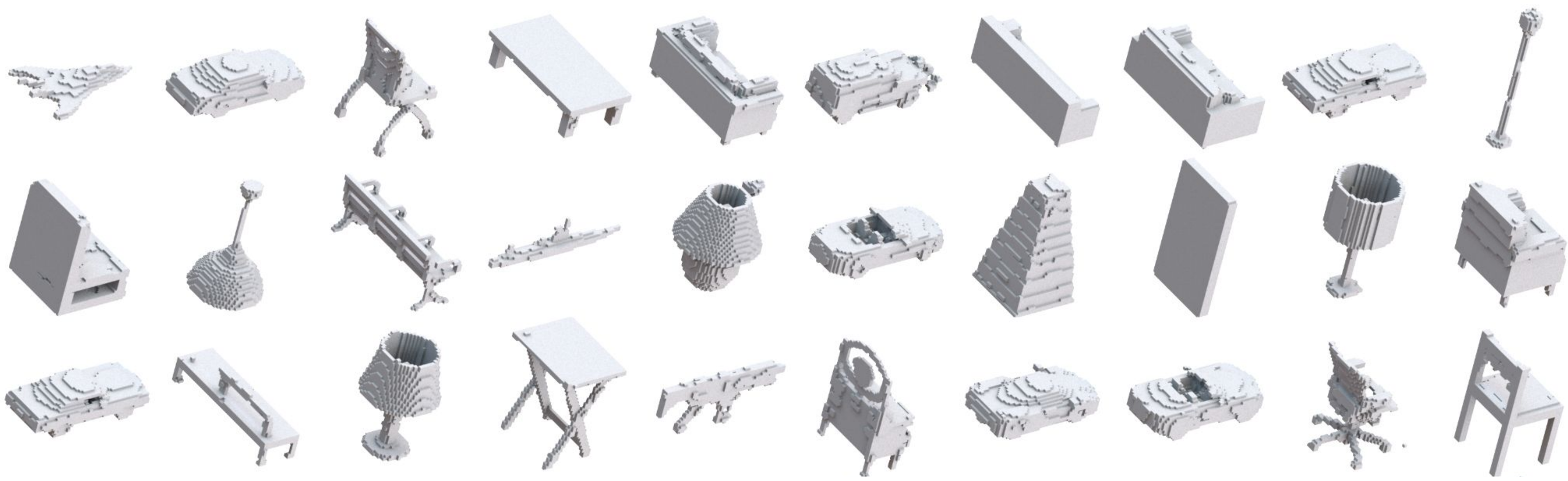
- Image generation performance on CelebA-HQ 64^2.

Model	# Params		Reconstruction (train)		Generation (train)				Efficiency	
	Learnable	Inference	PSNR ↑	rFID ↓	FID ↓	Precision ↑	Recall ↑	F1 ↑	fps	GFLOPS
Functa	3.3 M	2,629.6 K	26.6	28.4	40.4	0.577	0.397	0.470	332.9	8.602
GEM	99.0 M	921.3 K	-	-	30.4	0.642	0.502	0.563	559.6	3.299
GASP	34.2 M	83.1 K	-	-	13.5	0.836	0.312	0.454	1949.3	0.305
DPF	62.4 M	-	-	-	13.2	0.866	0.347	0.495	-	-
mNIF (S)	4.6 M	17.2 K	31.5	10.9	21.0	0.787	0.324	0.459	2958.6	0.069
mNIF (L)	33.4 M	83.3 K	34.5	5.8	13.2	0.902	0.544	0.679	891.3	0.340

Performance on Voxel

- Voxel generation performance on ShapeNet 64^3.

Model	# Params		Reconstruction (test)		Generation (test)		Inference Efficiency	
	Learnable	Inference	MSE ↓	PSNR ↑	Coverage ↑	MMD ↓	fps	GFLOPS
GASP	34.2 M	83.1 K	0.0296	16.52	0.341	0.0021	180.9	8.742
GEM	99.0 M	921.3 K	0.0153	21.32	0.409	0.0014	16.7	207.0
DPF	62.4 M	-	-	-	0.419	0.0016	-	-
mNIF (S)	4.6 M	17.2 K	0.0179	20.26	0.430	0.0014	191.5	4.363
mNIF (L)	33.4 M	83.3 K	0.0166	21.38	0.437	0.0013	69.6	21.613



Performance on Radiance Fields

- Radiance field generation performance on SRN Cars.

Model	# Params		Reconstruction (train)	Generation (test)	Inference Efficiency		
	Learnable	Inference	PSNR ↑	FID ↓	fps	TFLOPS	Memory
Functa	3.9 M	3,418.6 K	24.2	80.3	2.0	1.789	28.01 GB
mNIF (S)	4.6 M	17.2 K	25.9	79.5	97.7	0.009	1.26 GB

view 64 view 128 view 192

Interpolation of Context Vectors

Thank you for watching the video

Poster Session 3

Great Hall & Hall B1+B2 **Poster #537**

Wed 13th Dec 10:45 a.m. CST — 12:45 p.m. CST