Multi-Modal Contrastive Masked Autoencoders: A Two-Stage Progressive Pre-training Approach for RGBD Datasets

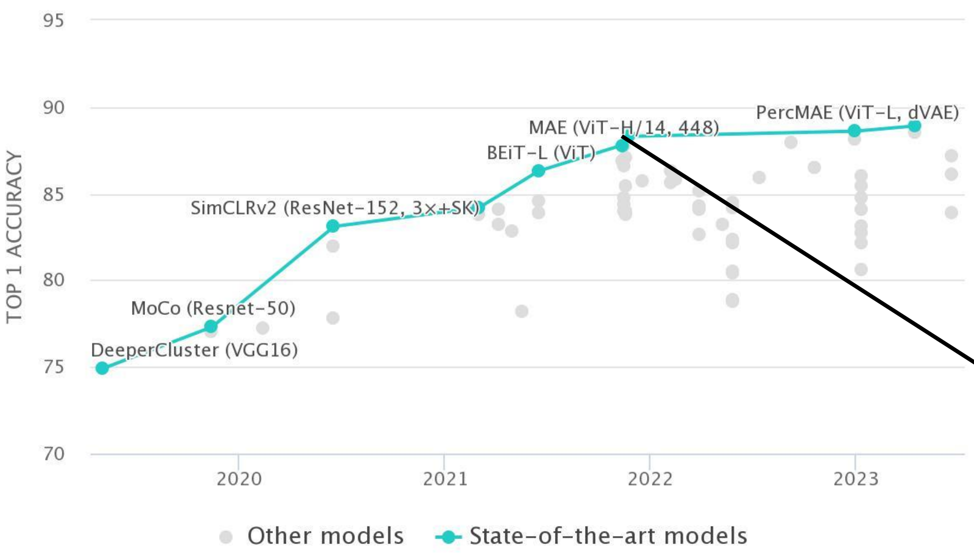
Muhammad Abdullah Jamal, Omid Mohareri

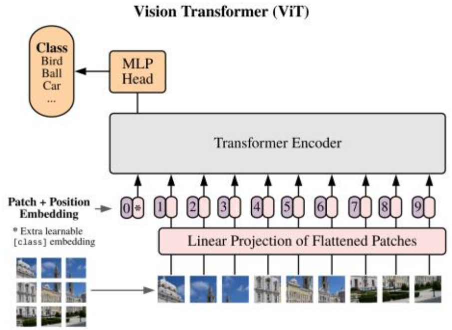
INTUITIVE

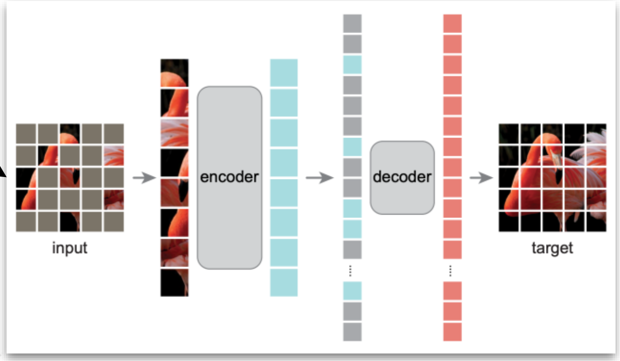
CVPR Nashville JUNE 11-15, 2025

Remarkable success of Vision Transformers (ViT) in computer vision tasks

ImageNet Classification

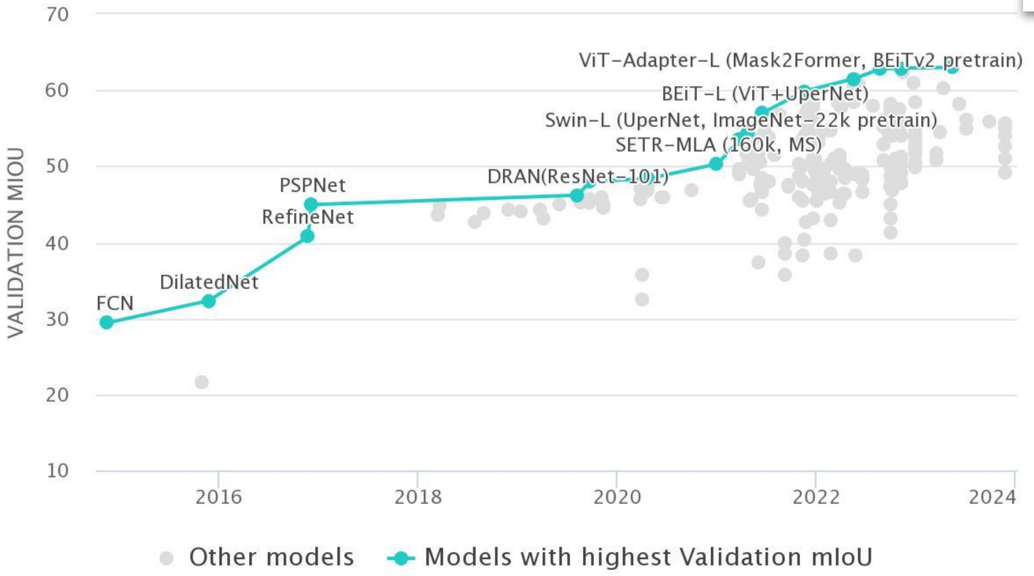




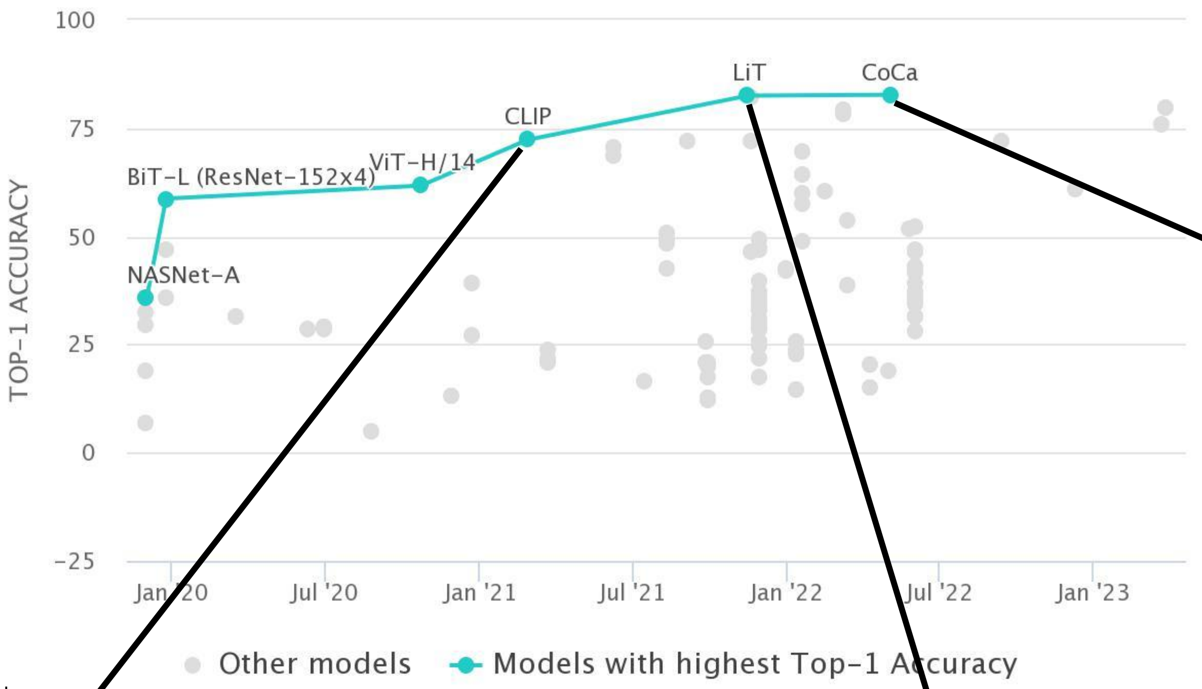


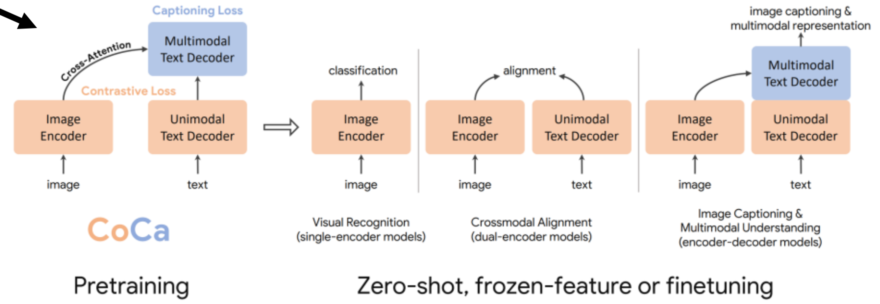
MAE
He et al.
CVPR 2022

ADE-20k Semantic Segmentation



Large scale Multi-modal Foundation models improves the performance of various downstream tasks



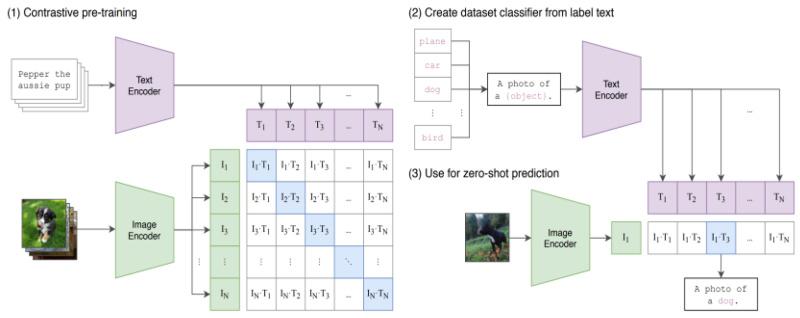


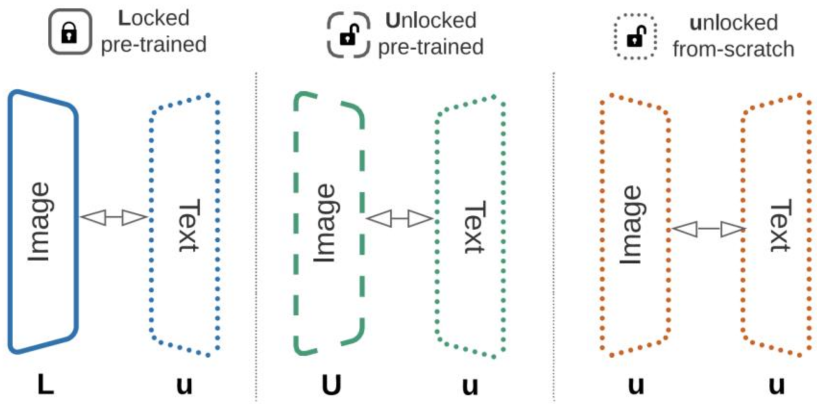
CoCa

Yui et al.
TMLR 2023

CLIP

Radford et al.
ICML 2021

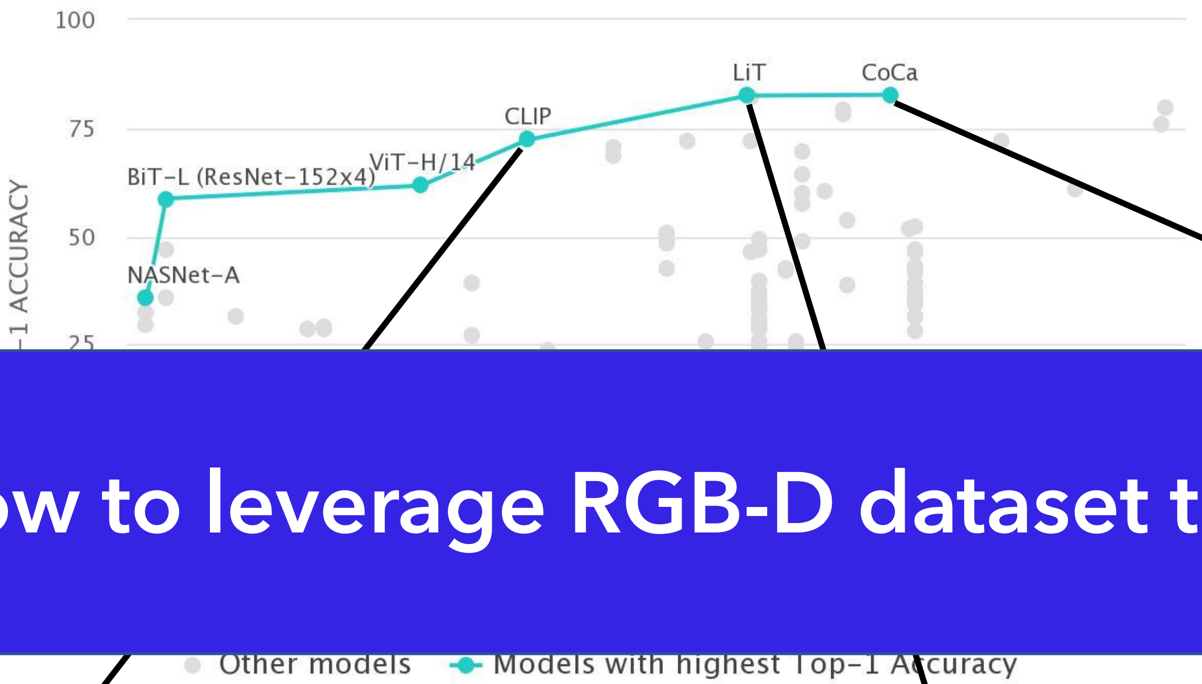


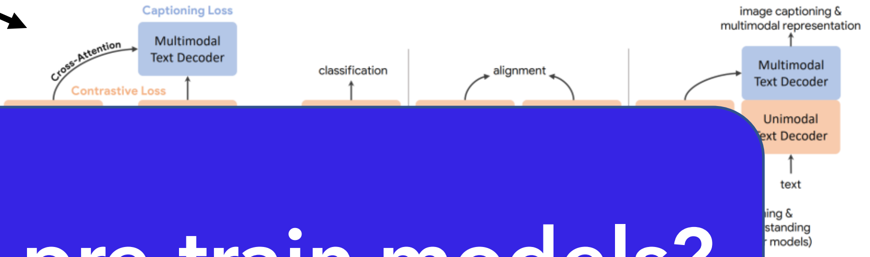


LiT

Zhai et al.
CVPR 2022

Large scale Multi-modal Foundation models improves the performance of various downstream tasks

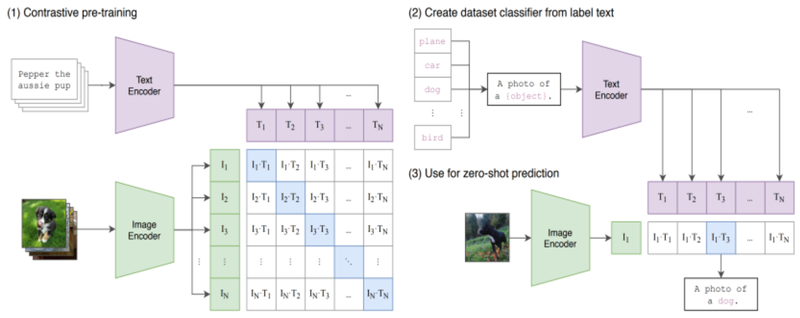


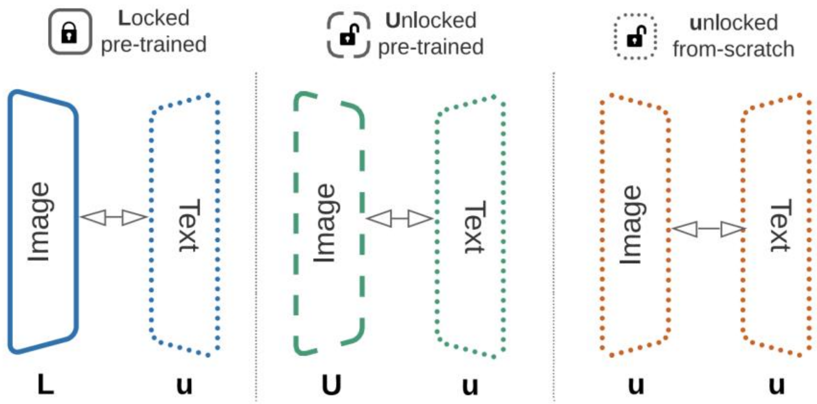


How to leverage RGB-D dataset to pre-train models?

CLIP

Radford et al.
ICML 2021



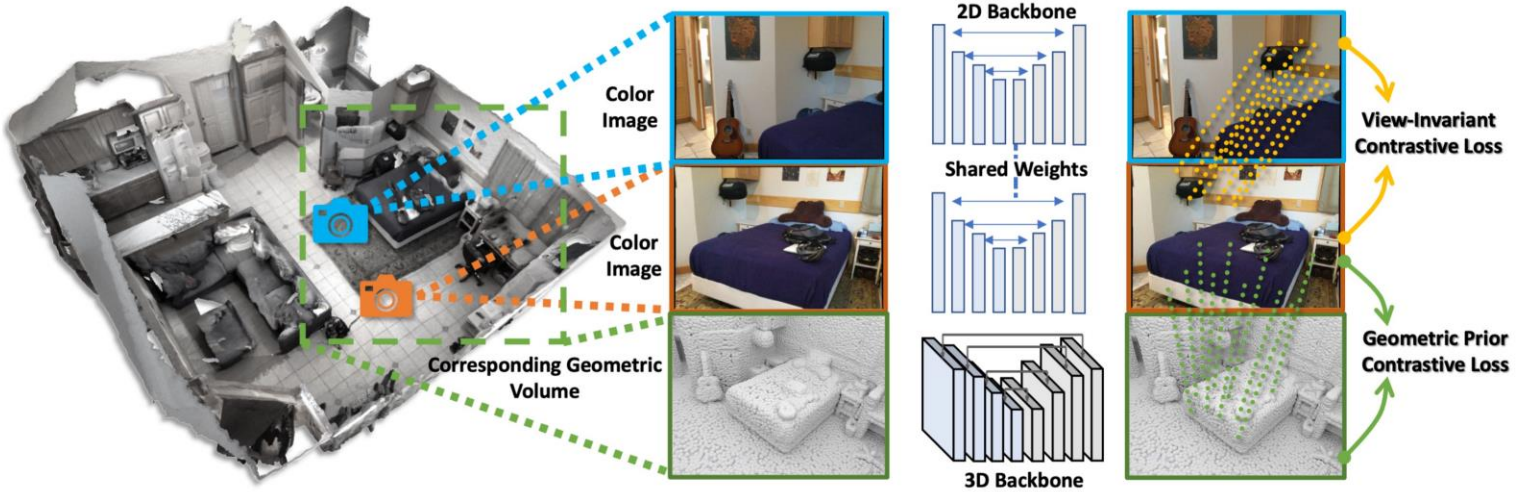


LiT

Zhai et al.
CVPR 2022

Contrastive Learning for RGB-D dataset

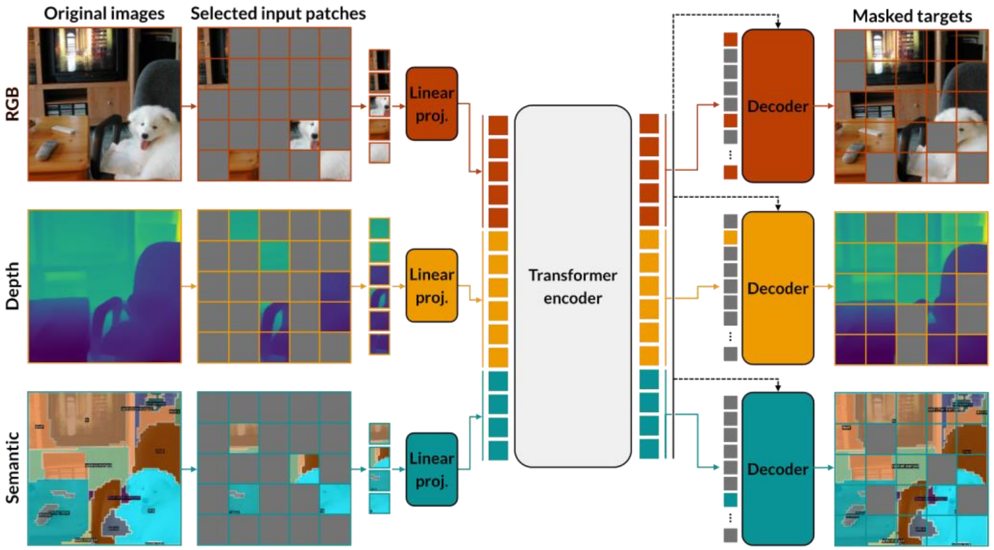
Pri3D
Hou et al.
ICCV 2021



- Embed 3D priors using 2D-3D correspondence
- Multi-view input
- Requires camera pose registration
- Doesn't capture discriminative features beyond mere cross-model correspondence

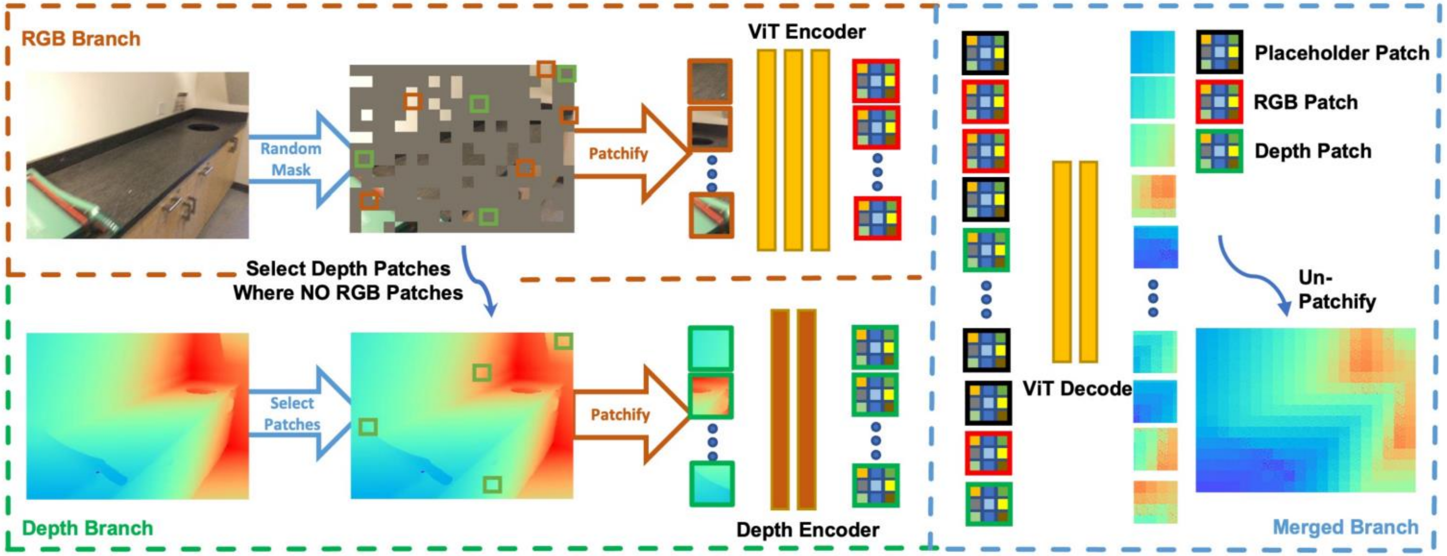
Masked Autoencoding pre-training for RGB-D

MultiMAE
Roman et al.
ECCV 2022



- Requires segmentation labels during pre-training
- Relies on multi-modal data during finetuning.
- Cross-attention model for cross-modal prediction

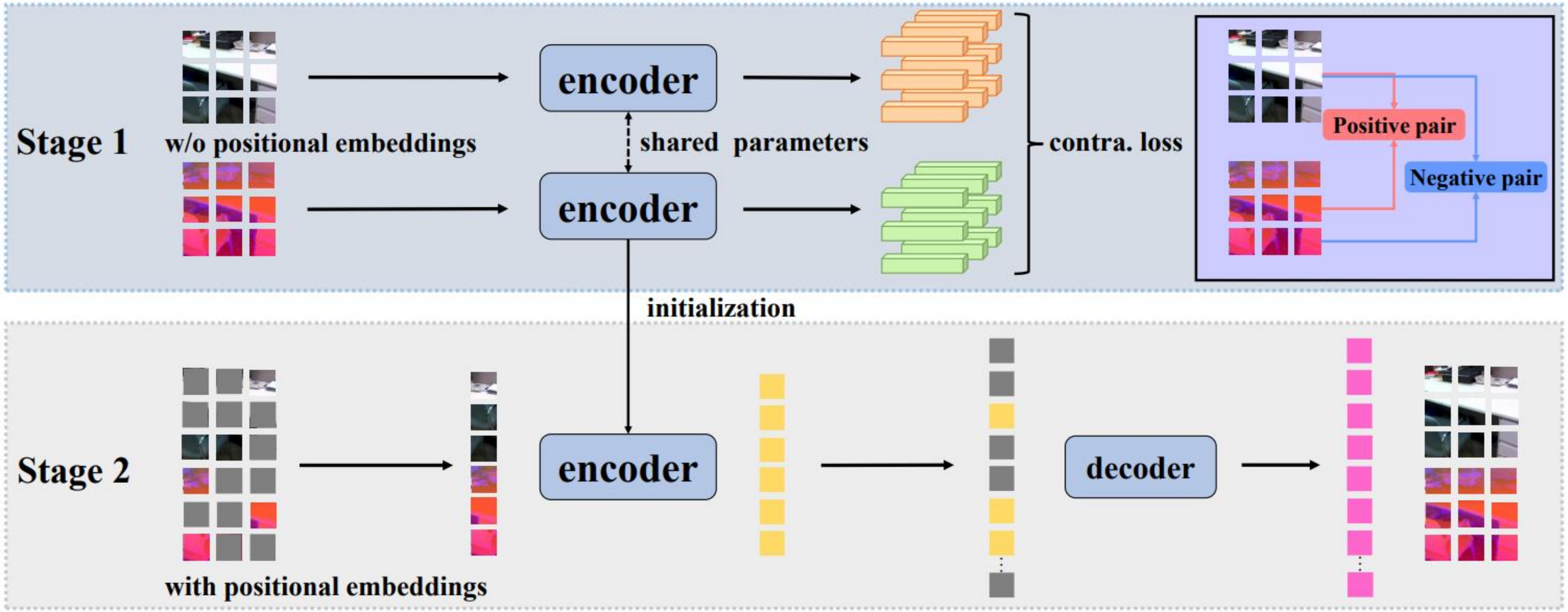
Mask3D
Hou et al.
CVPR 2023



- Limited to 2D image understanding tasks
- Relies on MAE pre-training for any cross-modal representation learning.

Contrastive and Masked Autoencoding pre-training for RGB-D

CoMAE
Jiange et al.
AAAI 2023



- Small-scale RGB-D datasets
- Requires both RGB and depth data during finetuning and evaluation
- Fails to distill features learned in stage-1

Contrastive and Masked Autoencoding pre-training for RGB-D

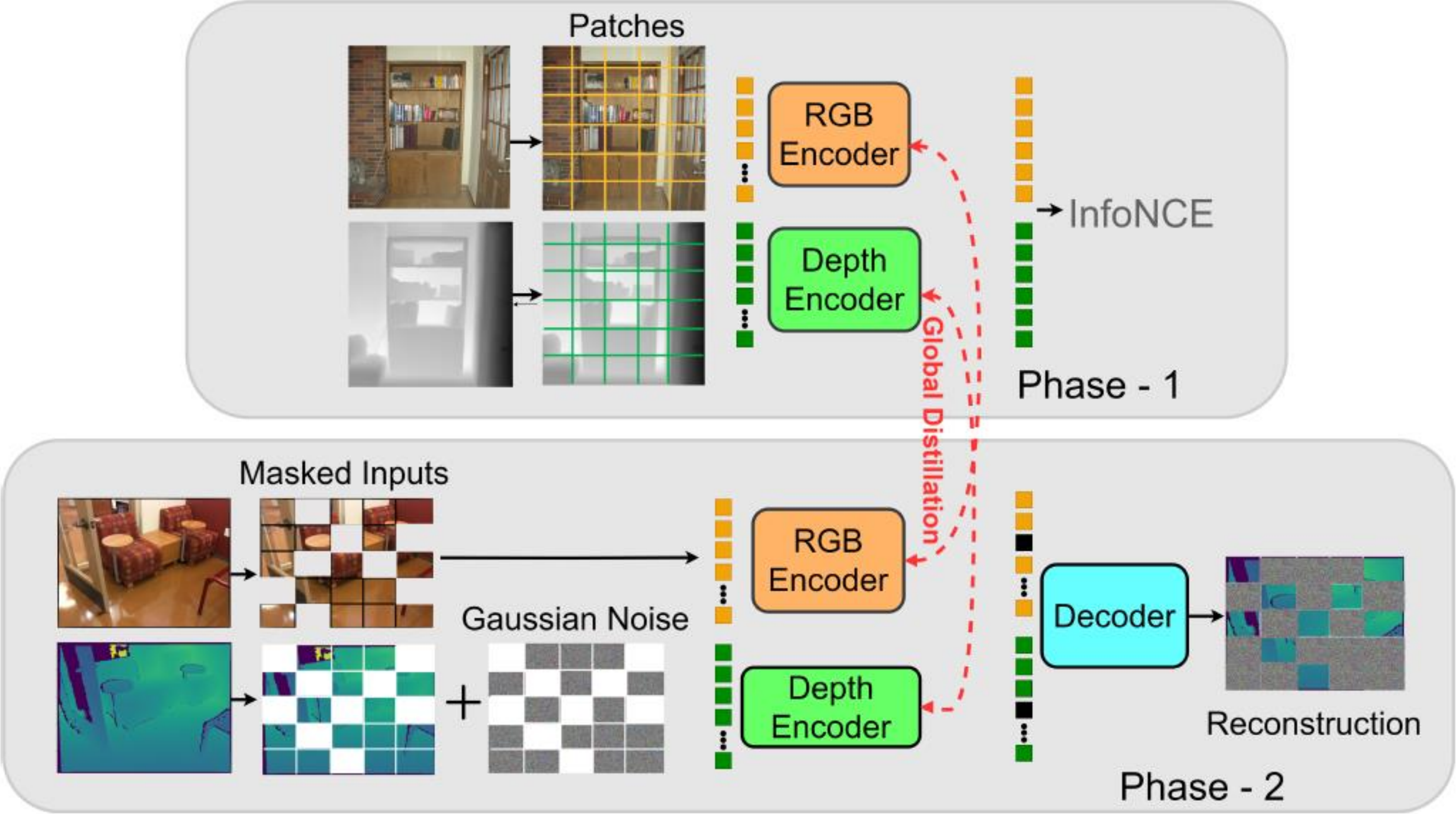
C
Jian
AAA

- All these approaches frequently struggle to capture high-frequency components of the data. Can denoising inspired from diffusion models encourage the model to extract high-frequency features?

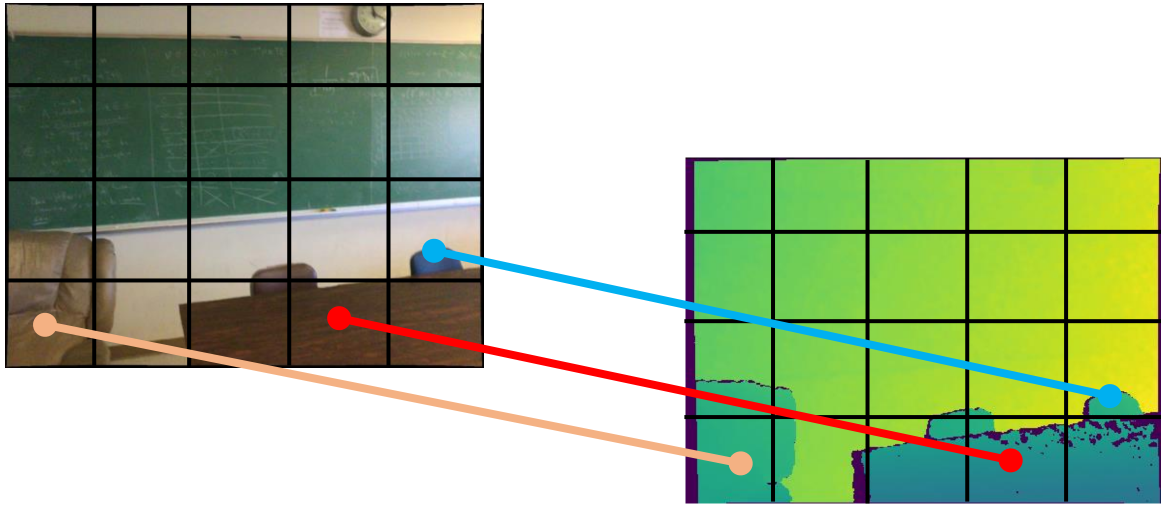
- A pre-training approach that combines powerful self-supervised learning methods for pre-training RGB-D ?

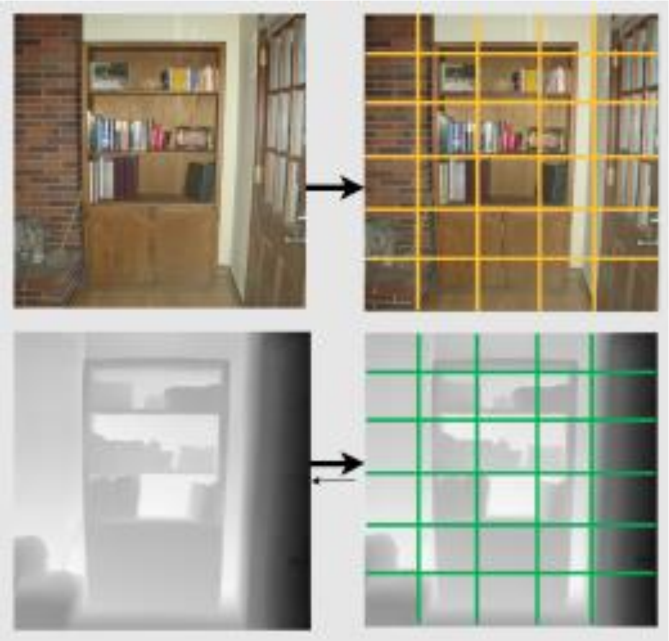
- Requires both RGB and depth data during finetuning and evaluation
- Fails to distill features learned in stage-1

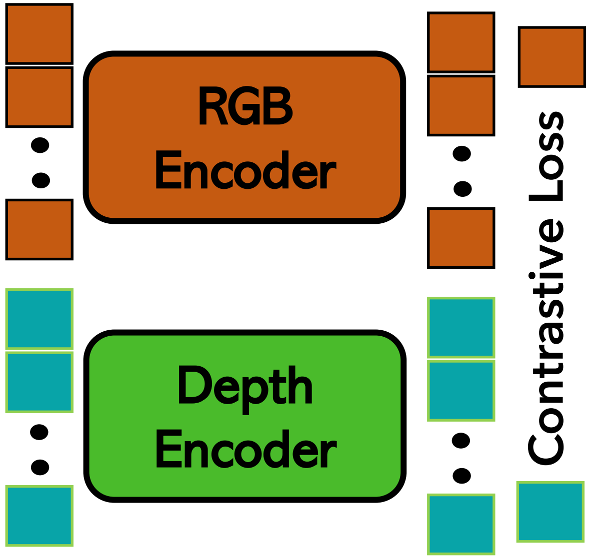
Two-stage progressive pre-training



Stage -1: Cross-Modal Representation Learning



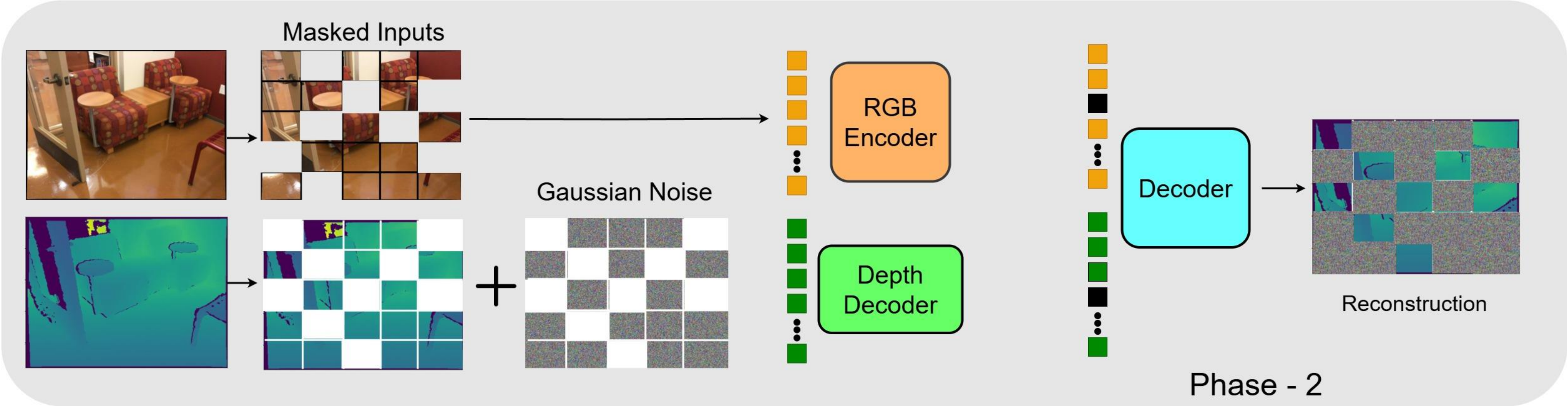




RGB-Depth patch level correspondence

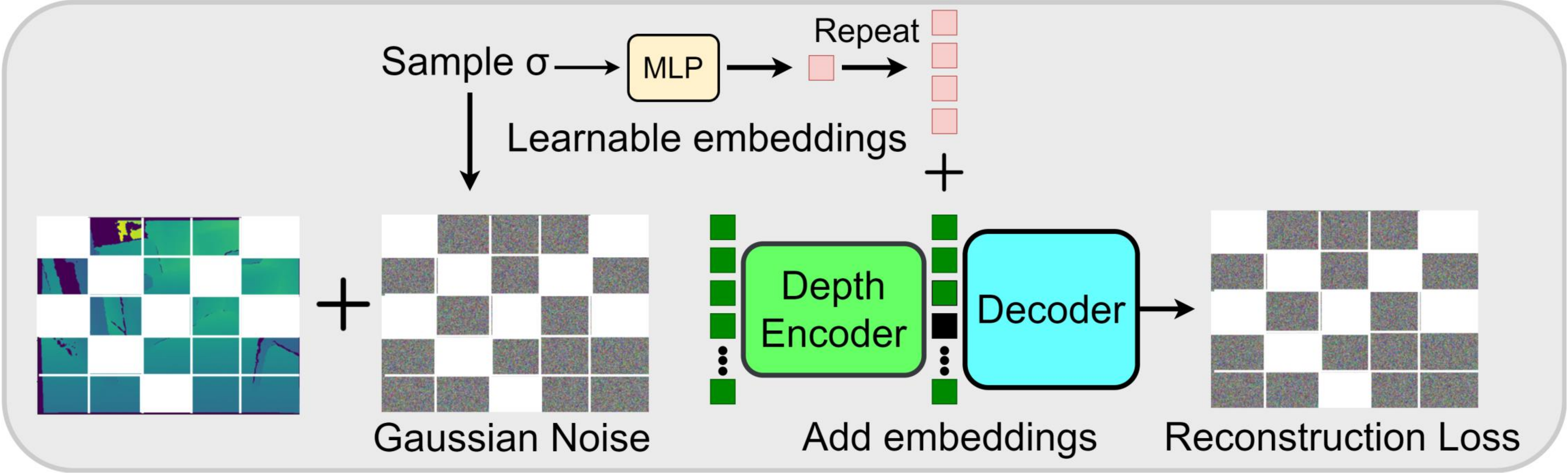
$$\mathcal{L}_{\mathrm{c}} = -\frac{1}{N}\sum_{i=1}^{N}\log\left[\frac{\exp(s_{i,i}/\tau)}{\sum_{k\neq i}\exp(s_{i,k}/\tau)+\exp(s_{i,i}/\tau)}\right]$$

Stage-2: Multi-Modal Masked Autoencoding



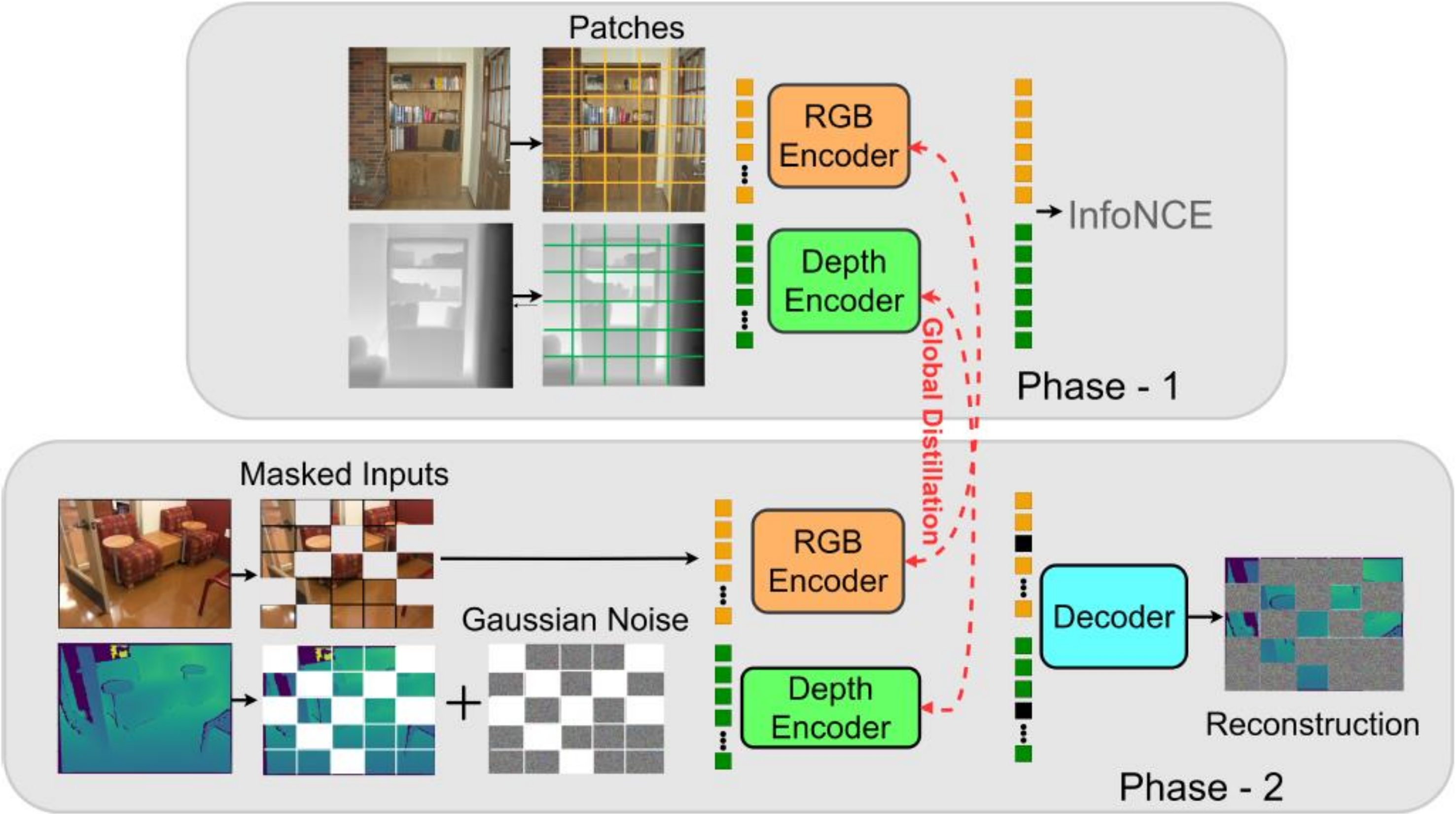
$$\mathcal{L}_{\text{depth}} = \frac{1}{n} \sum_{i=1}^{n} \|\mathbf{M}_i^{depth} \circ (\mathbf{x}_i^{depth} - \hat{\mathbf{x}}_i^{depth})\|_2^2$$

Denoising



$$\mathcal{L}_{\text{denoise}} = \frac{1}{n}\sum_{i=1}^{n} \|(1 - \mathbf{M}_i^{depth}) \circ (\sigma_i^{depth}\mathbf{e}_i^{depth} - \hat{\mathbf{x}}_i^{depth})\|_2^2$$

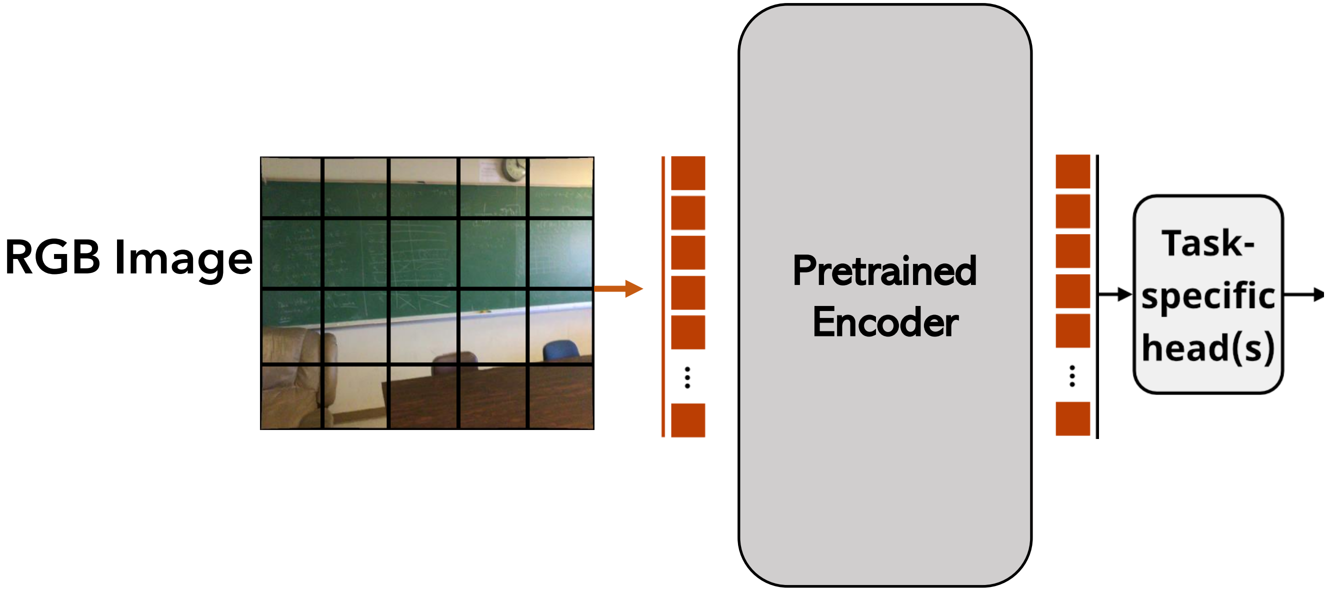
Feature distillation



$$\mathcal{L}_{\text{distill}}(\mathbf{f}_2, \mathbf{f}_1) = \begin{cases} \frac{1}{2}(\mathbf{f}_2 - \mathbf{f}_1)^2/\beta, & |\mathbf{f}_2 - \mathbf{f}_1| \leq \beta \\ (|\mathbf{f}_2 - \mathbf{f}_1| - \frac{1}{2}\beta), & \text{otherwise} \end{cases},$$

Transfer from Our approach

Flexibly transfer using RGB modality for multiple downstream tasks



- Semantic Segmentation
- Depth Estimation
- Instance Segmentation

Semantic Segmentation on ScanNet

Methods	Reconstruction task	Backbone	Pre-train	Fine-tune Modality	mIoU
Scratch	-	ViT-B	None	RGB	32.6
Pri3D [44]		ViT-B	ImageNet+ScanNet	RGB	59.3
Pri3D [44]	-	ResNet-50	ImageNet+ScanNet	RGB	60.2
DINO [12]	-	ViT-B	ImageNet+ScanNet	RGB	58.1
MAE [38]	RGB	ViT-B	ImageNet	RGB	64.8
MAE [38]	RGB	ViT-B	ImageNet+ScanNet	RGB	64.5
MultiMAE* [6]	RGB + Depth	ViT-B	ImageNet+ScanNet	RGB	65.1
Mask3D** [45]	Depth	ViT-B	ImageNet+ScanNet	RGB	66.2
Mask3D [45]	RGB + Depth	ViT-B	ImageNet+ScanNet	RGB	65.5
Ours	Depth	ViT-B	ImageNet+ScanNet	RGB	**67.5**
MultiMAE [6]	RGB + Depth + Segmentation	ViT-B	ImageNet	RGB	66.4

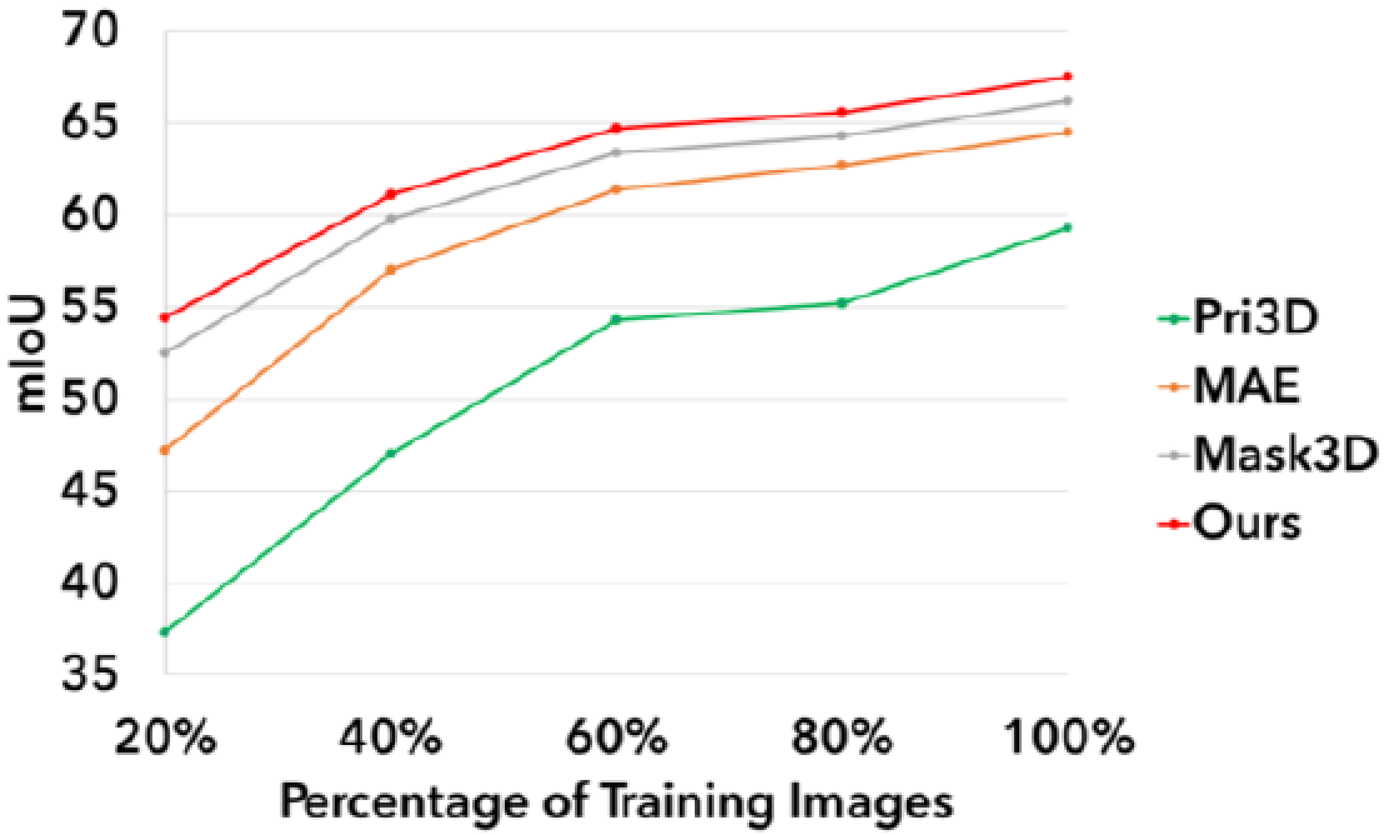
Instance Segmentation on ScanNet

Methods	Pre-train	AP
Scratch	None	12.2
ImageNet baseline	Supervised ImageNet	17.6
Pri3D [44]	ImageNet + ScanNet	18.3
MoCov2 [86]	ImageNet + ScanNet	18.3
MAE [38]	ImageNet + ScanNet	20.7
MultiMAE [6]	ImageNet + ScanNet	22.4
Mask3D [45]	ImageNet + ScanNet	22.8
Ours	ImageNet + ScanNet	**23.7**

Depth Estimation on NYUv2

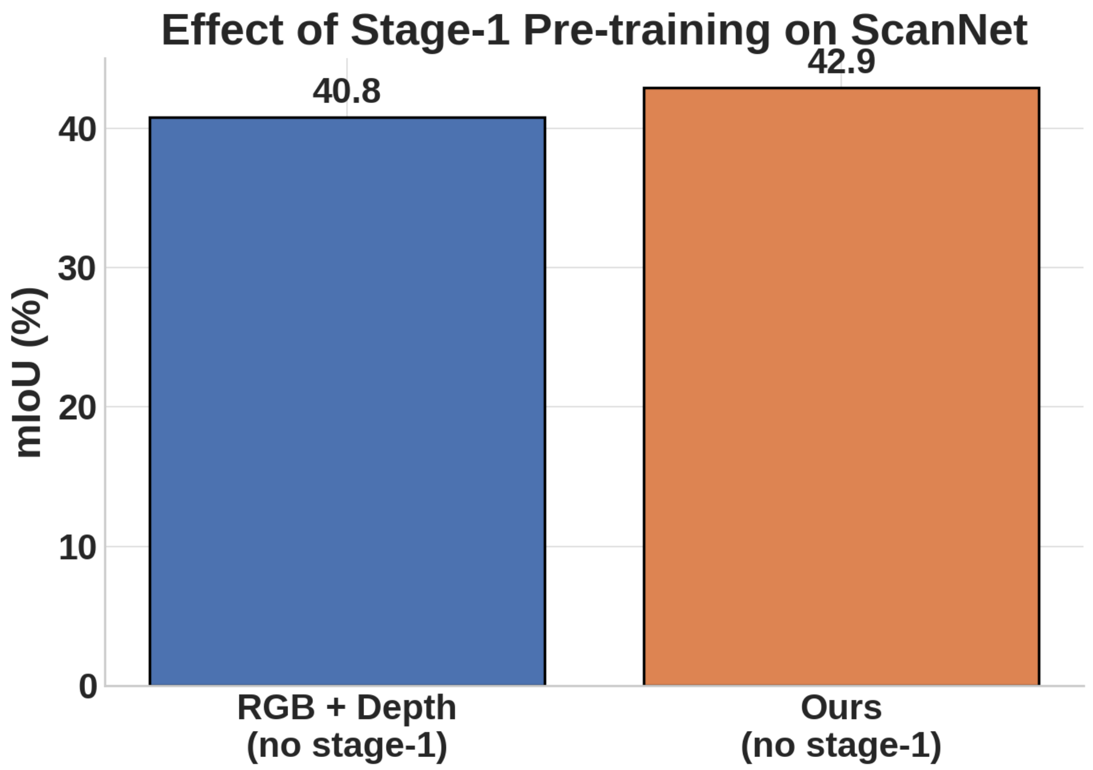
Methods	Reconstruction task	Backbone	Pre-train	Fine-tune Modality	δ_1
MAE [38]	RGB	ViT-B	ImageNet	RGB	85.1
Mask3D [45]	Depth	ViT-B	ImageNet+ScanNet	RGB	85.4
CroCo [80]	RGB + Depth	ViT-B	Habitat	RGB	85.6
MultiMAE* [6]	RGB + Depth + Segmentation	ViT-B	ImageNet	RGB	83.0
MultiMAE [6]	RGB + Depth	ViT-B	ImageNet+ScanNet	RGB	85.3
Ours	Depth	ViT-B	ImageNet+ScanNet	RGB	**87.1**
MultiMAE [6]	RGB + Depth + Segmentation	ViT-B	ImageNet	RGB	**86.4**

Data-Efficient Learner



Effect of different components

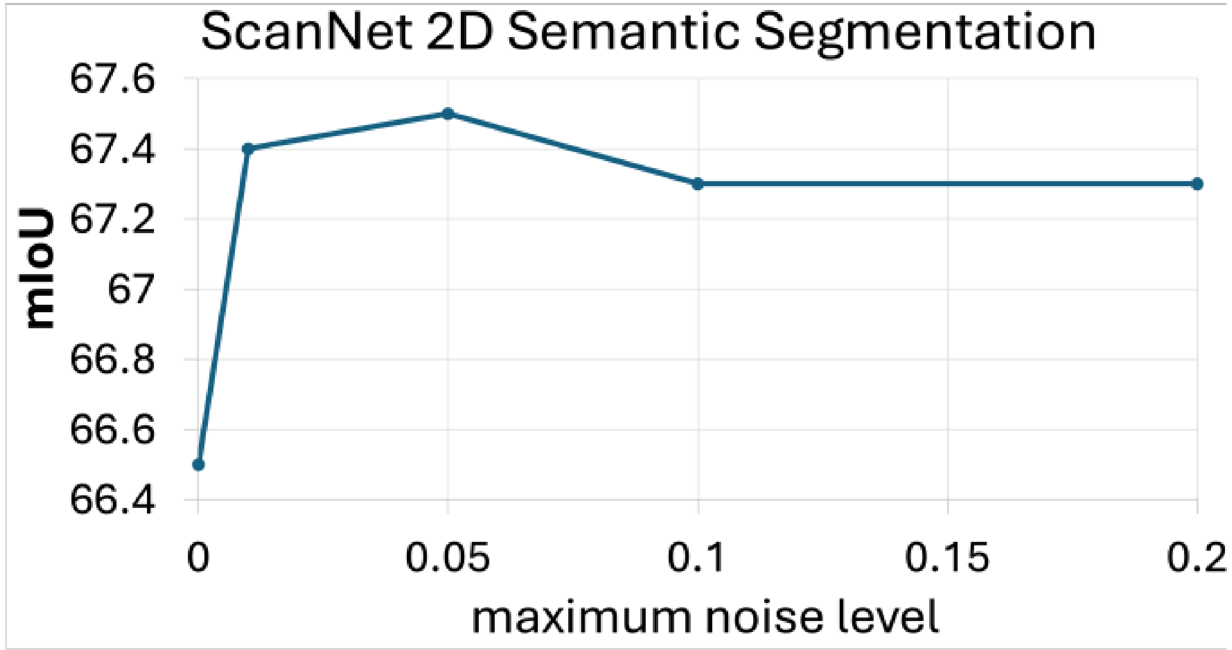
Contrastive	Reconstruction	Denoising	Distill	mIoU
✓	✗	✗	✗	63.4
✓	✓	✗	✗	66.3
✓	✓	✗	✓	66.5
✓	✓	✓	✗	67.0
✓	✓	✓	✓	**67.5**



Effect of noise

Method	Backbone	mIoU
w/out noise	ViT-B	66.5
noise only	ViT-B	66.9
Full	ViT-B	**67.5**

Table 6. Ablation study on the components of the denoising method. We report the performance on **ScanNet 2D semantic segmentation**.



Multi-Modal Contrastive Masked Autoencoders: A Two-Stage Progressive Pre-training Approach for RGBD Datasets

Muhammad Abdullah Jamal, Omid Mohareri

INTUITIVE

CVPR Nashville JUNE 11-15, 2025

- Simple & effective multi-modal pre-training to integrate strong self-supervised methods
- Trained without any semantic pseudo labels
- Inspired by diffusion models, integrates denoising to extract high-frequency components.
- Feature distillation to distill knowledge learnt in stage-1
- Notable performance gains on multiple datasets for multiple downstream tasks