valeo.ai

MINES PARIS | PSL

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

RangeViT: Towards Vision Transformers for 3D Semantic Segmentation in Autonomous Driving

Angelika Ando[1,2], Spyros Gidaris[1], Andrei Bursuc[1], Gilles Puy[1], Alexandre Boulch[1], Renaud Marlet[1,3]

[1]Valeo.ai, Paris, France [2]Centre for Robotics, Mines Paris - Université PSL, Paris, France
[3]LIGM, Ecole des Ponts, Univ. Gustave Eiffel, CNRS, Marne-la-Vallée, France

TUE-PM-105


Angelika Ando


Spyros Gidaris

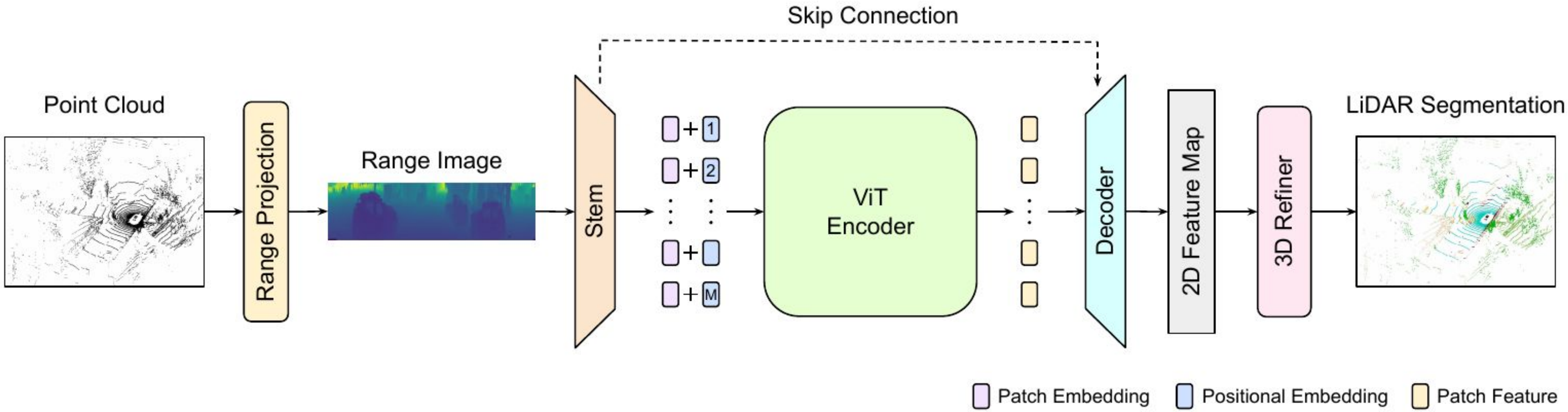

Andrei Bursuc


Gilles Puy


Alexandre Boulch

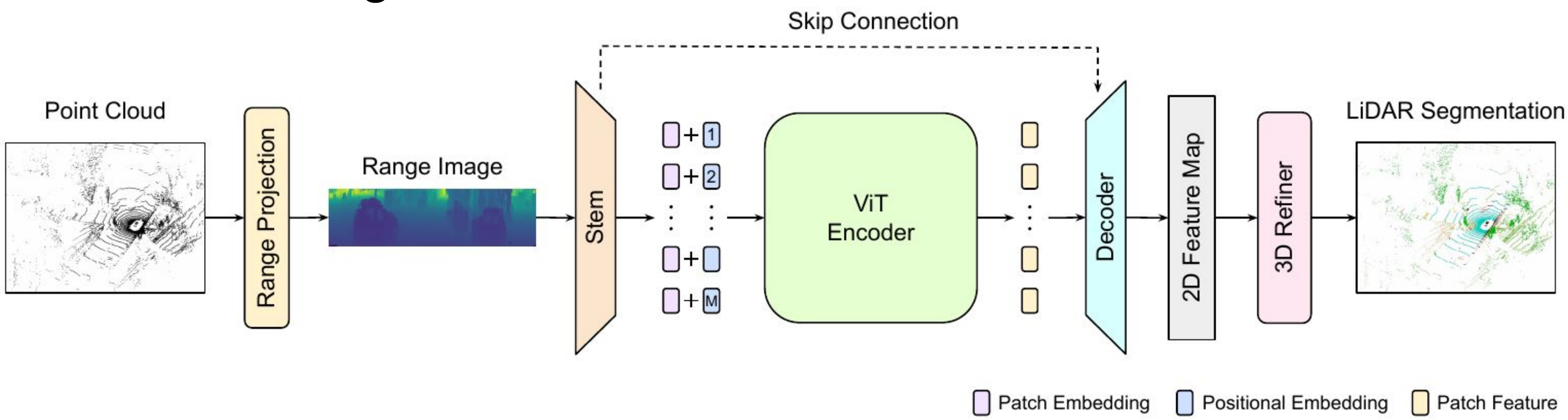

Renaud Marlet

Overview of RangeViT



Can 3D LiDAR semantic segmentation benefit from the latest improvements on Vision Transformers?

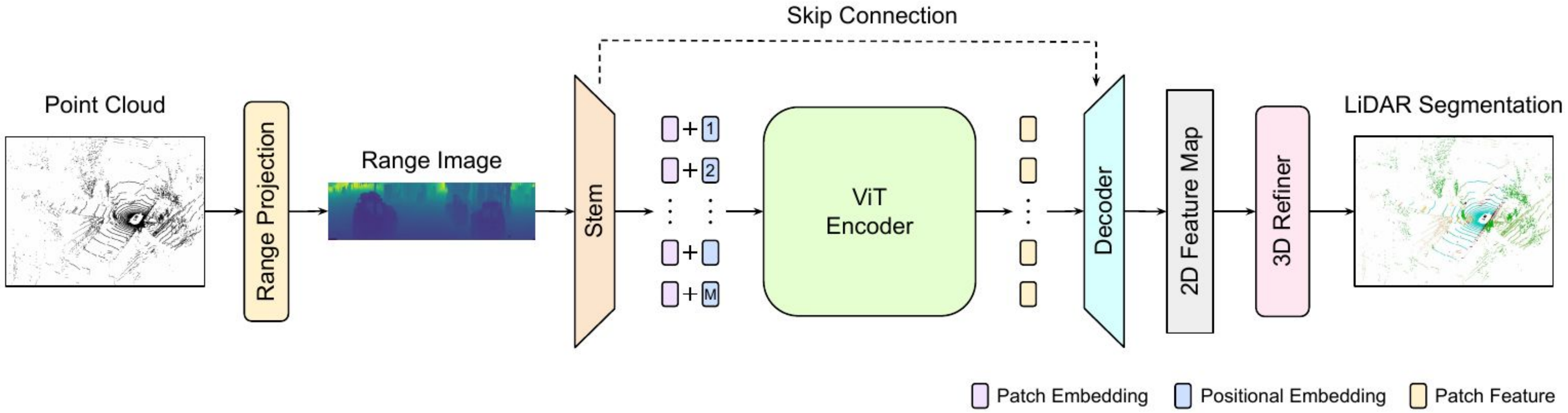
Overview of RangeViT



Can 3D LiDAR semantic segmentation benefit from the latest improvements on Vision Transformers?

Yes, with RangeViT: **a simple ViT-based point cloud segmentation method.**

Overview of RangeViT

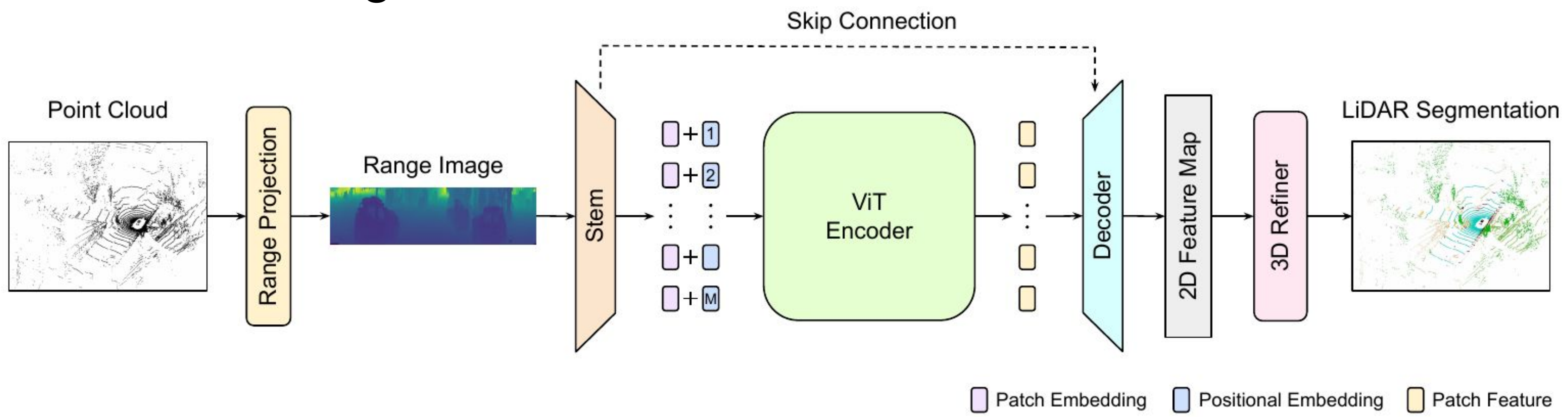


Can 3D LiDAR semantic segmentation benefit from the latest improvements on Vision Transformers?

Yes, with RangeViT: **a simple ViT-based point cloud segmentation method.**

- **Exploits the strong representation learning** capacity of ViTs for LiDAR segmentation.

Overview of RangeViT

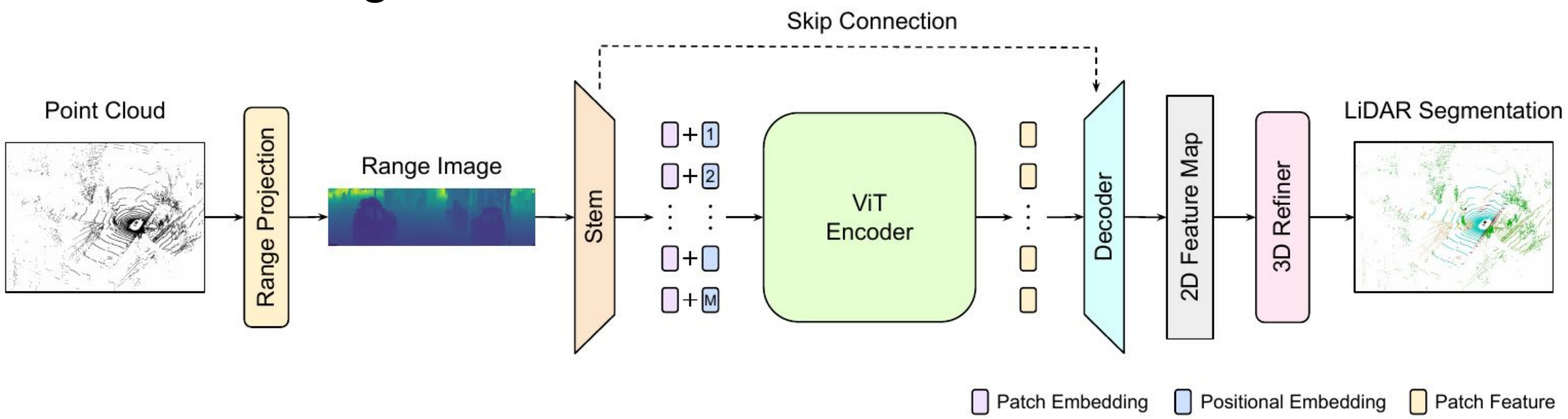


Can 3D LiDAR semantic segmentation benefit from the latest improvements on Vision Transformers?

Yes, with RangeViT: **a simple ViT-based point cloud segmentation method.**

- **Exploits the strong representation learning** capacity of ViTs for LiDAR segmentation.
- **Unify architectures** in LiDAR and image domain ⇒ Any advance in one domain benefits to both.

Overview of RangeViT

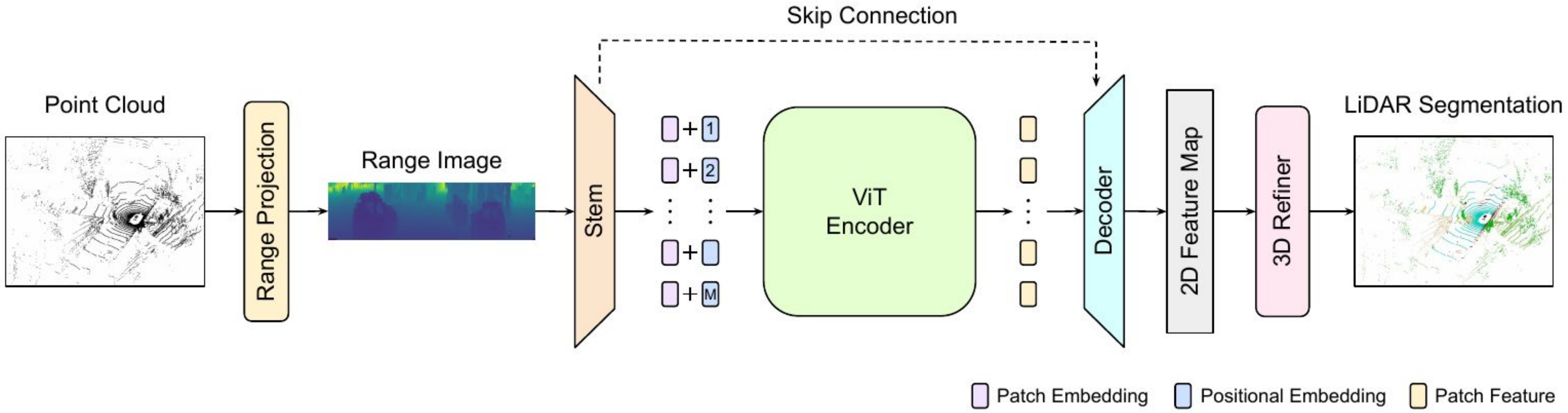


Can 3D LiDAR semantic segmentation benefit from the latest improvements on Vision Transformers?

Yes, with RangeViT: **a simple ViT-based point cloud segmentation method.**

- **Exploits the strong representation learning** capacity of ViTs for LiDAR segmentation.
- **Unify architectures** in LiDAR and image domain ⇒ Any advance in one domain benefits to both.
- **Leverages ViTs pre-trained** on large **RGB image datasets** for LiDAR segmentation.

Overview of RangeViT



Can 3D LiDAR semantic segmentation benefit from the latest improvements on Vision Transformers?

Yes, with RangeViT: **a simple ViT-based point cloud segmentation method.**

- **Exploits the strong representation learning** capacity of ViTs for LiDAR segmentation.
- **Unify architectures** in LiDAR and image domain ⇒ Any advance in one domain benefits to both.
- **Leverages ViTs pre-trained** on large **RGB image datasets** for LiDAR segmentation.
- Strong LiDAR segmentation results ⇒ **Surpasses prior projection-based methods.**

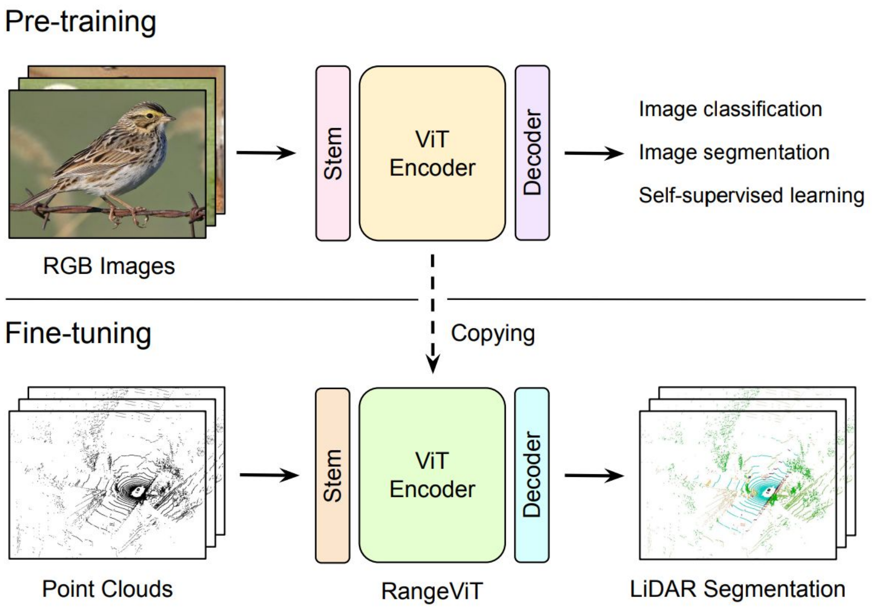
Method	nuScenes mIoU (%)	SemKITTI mIoU (%)
Voxel-based		
Cylinder3D	**76.1**	**67.8**
2D Projection-based		
RangeNet++	65.5	52.2
PolarNet	71.0	54.3
SalsaNext	72.2	59.5
KPRNet	-	63.1
Lite-HDSeg	-	63.8
RangeViT-CS (ours)	**75.2**	**64.0**

Motivation

Can 3D LiDAR semantic segmentation benefit from the latest improvements on Vision Transformers?

Yes, with RangeViT!

- Simple ViT-based point cloud segmentation method.
- **Same ViT backbone (and pre-trained weights) as in the image domain.**

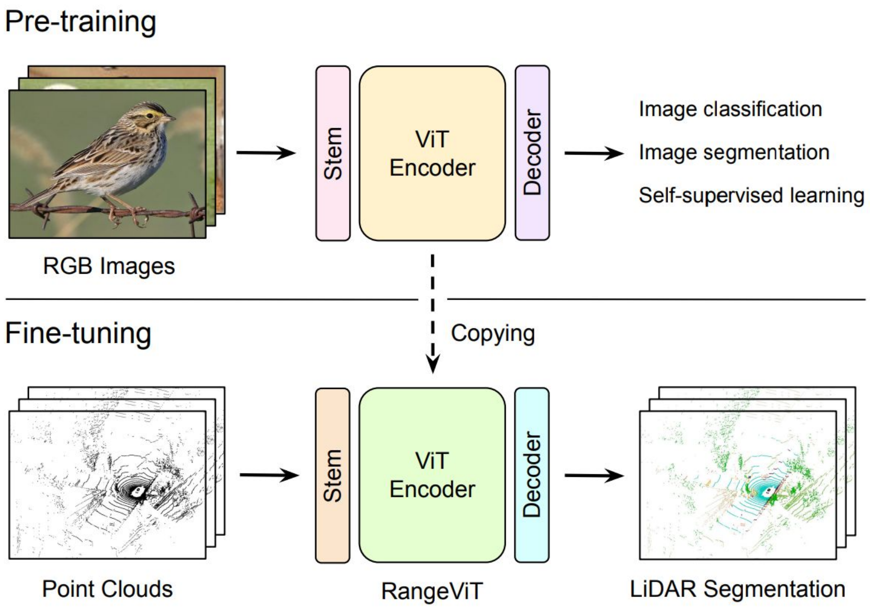


Motivation

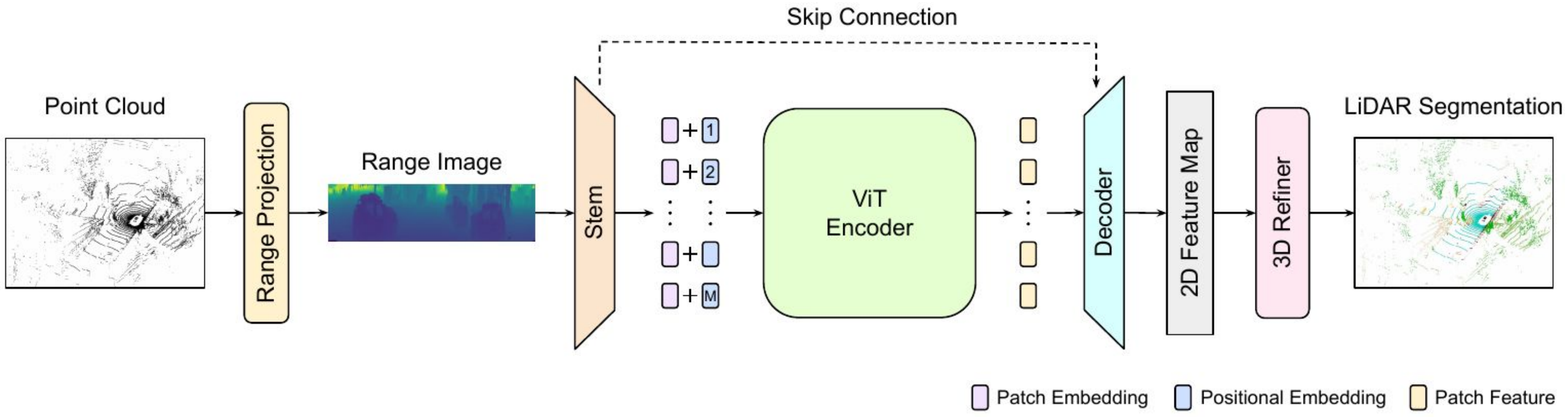
Can 3D LiDAR semantic segmentation benefit from the latest improvements on Vision Transformers?

Yes, with RangeViT!

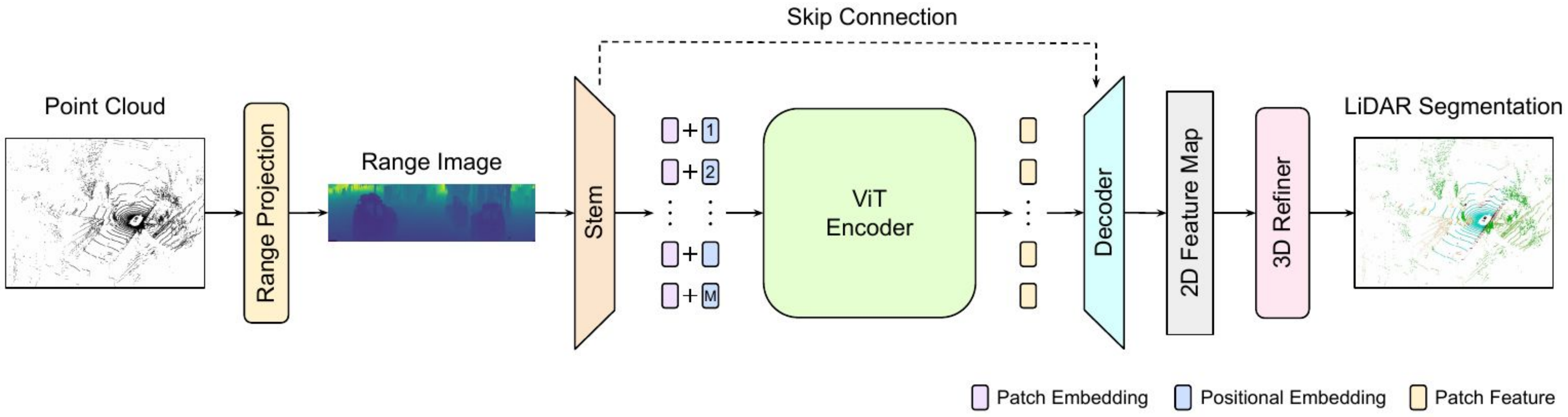
- Simple ViT-based point cloud segmentation method.
- **Same ViT backbone (and pre-trained weights) as in the image domain.**
- ViT **tokenization adapted for LiDAR data.**
- **Fast point cloud processing** by 2D range projection.



Overview of RangeViT

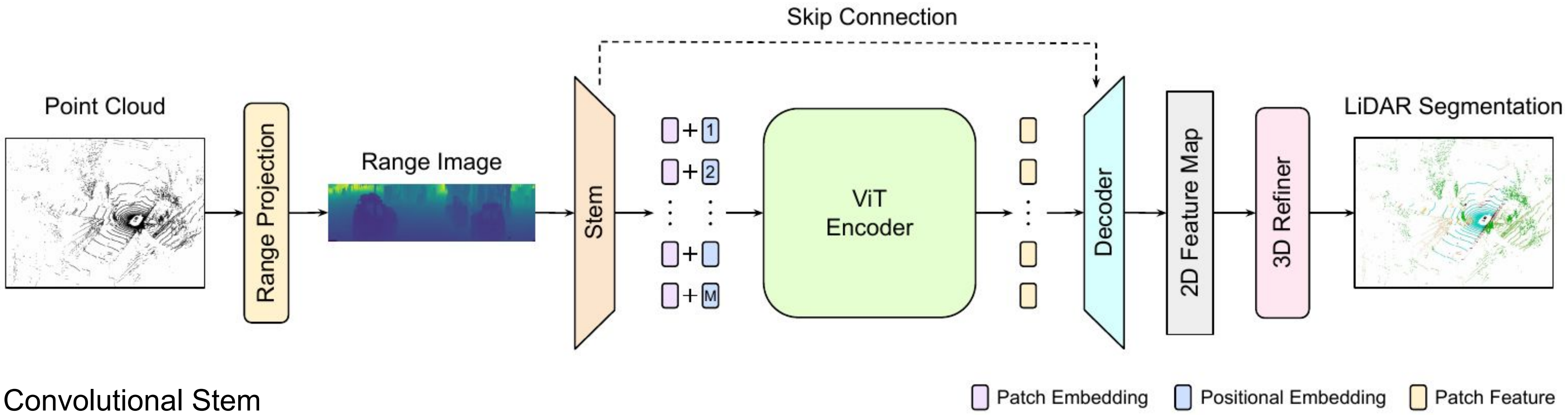


Overview of RangeViT

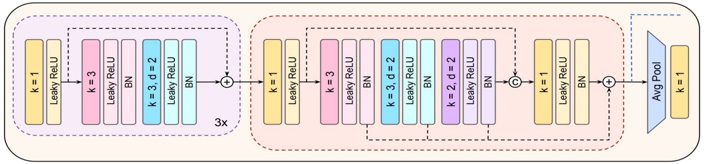


What makes an effective ViT architecture for 3D LiDAR segmentation?

Use non-linear convolution stem

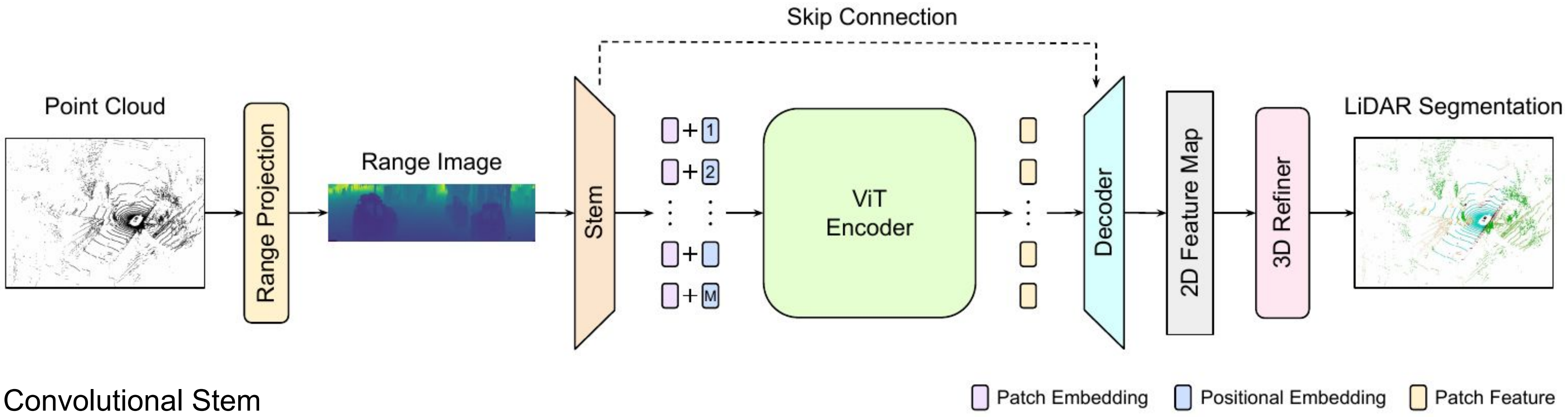


Convolutional Stem

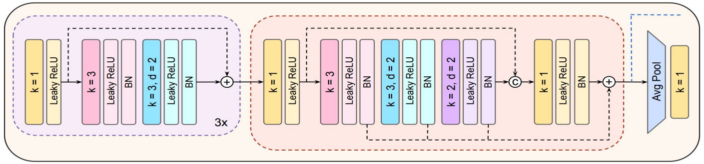


Stem	Decoder	Refiner	mIoU	#Params
Linear	Linear		65.52	22.0M
Conv	Linear		69.82	22.8M
Conv	UpConv		73.83	24.6M
Conv	UpConv	✓	**74.60**	25.2M

Use non-linear convolution stem

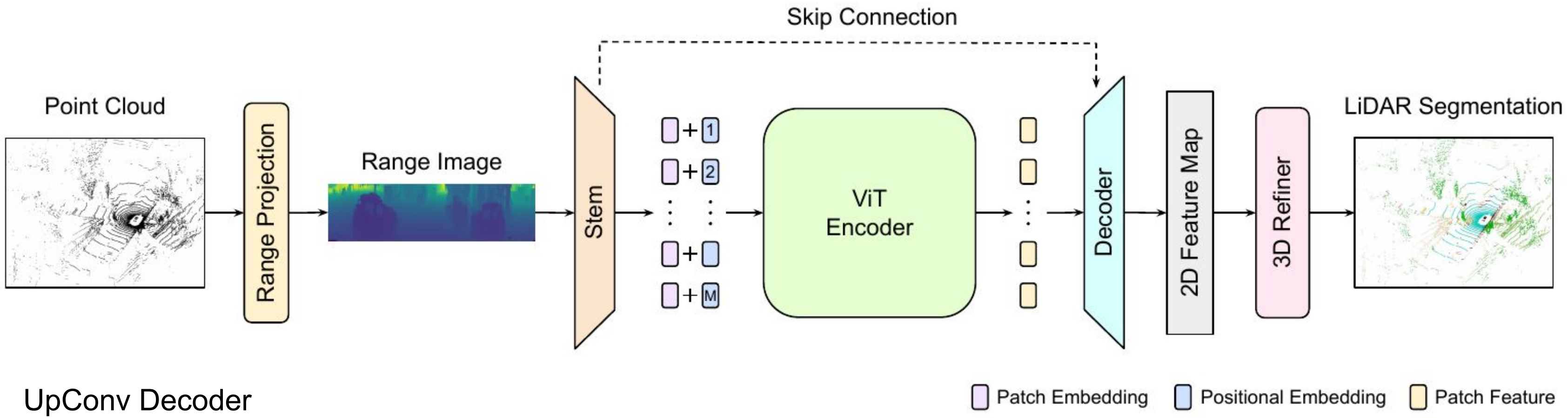


Convolutional Stem

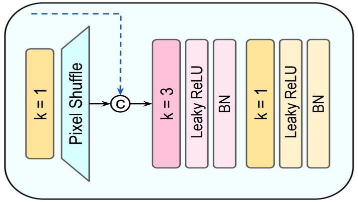


Stem	Decoder	Refiner	mIoU	#Params
Linear	Linear		65.52	22.0M
Conv	Linear		69.82	22.8M
Conv	UpConv		73.83	24.6M
Conv	UpConv	✓	**74.60**	25.2M

Use non-linear (UpConv) decoder

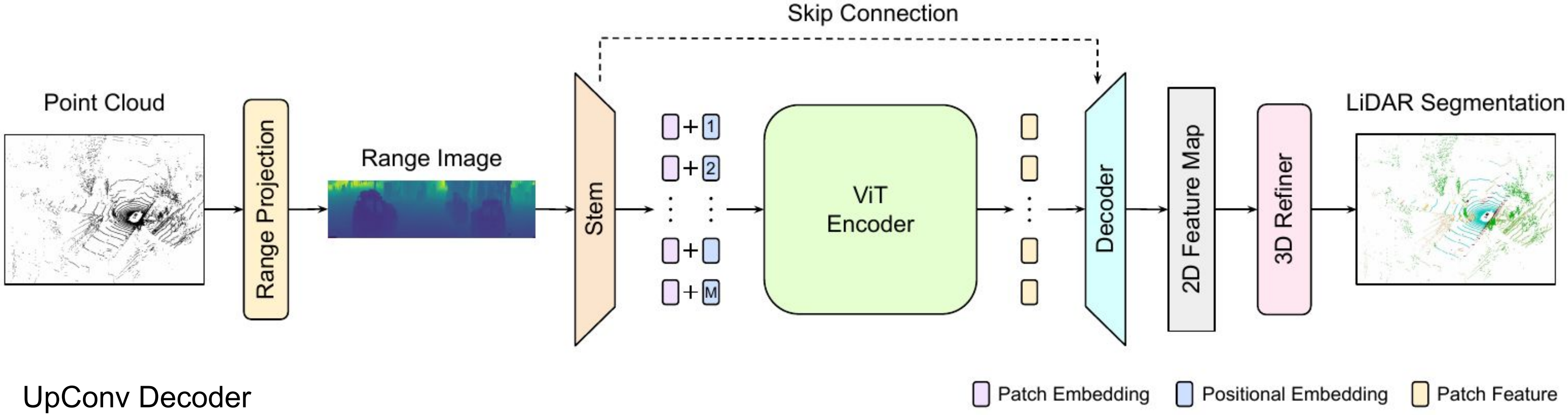


UpConv Decoder

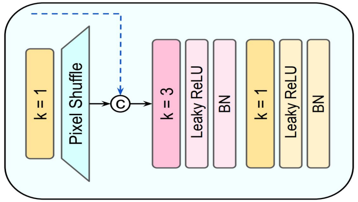


Stem	Decoder	Refiner	mIoU	#Params
Linear	Linear		65.52	22.0M
Conv	Linear		69.82	22.8M
Conv	UpConv		73.83	24.6M
Conv	UpConv	✓	**74.60**	25.2M

Use non-linear (UpConv) decoder + 3D Refiner

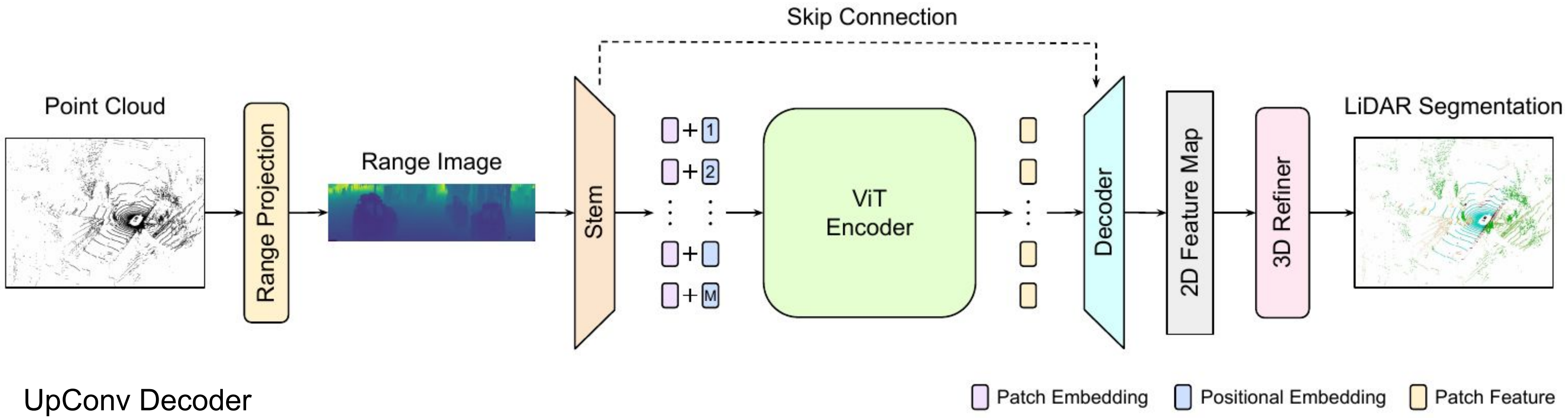


UpConv Decoder

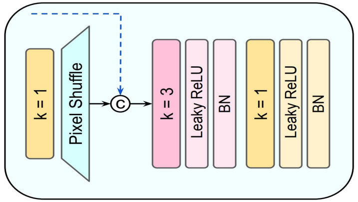


Stem	Decoder	Refiner	mIoU	#Params
Linear	Linear		65.52	22.0M
Conv	Linear		69.82	22.8M
Conv	UpConv		73.83	24.6M
Conv	UpConv	✓	**74.60**	25.2M

Use non-linear (UpConv) decoder + 3D Refiner

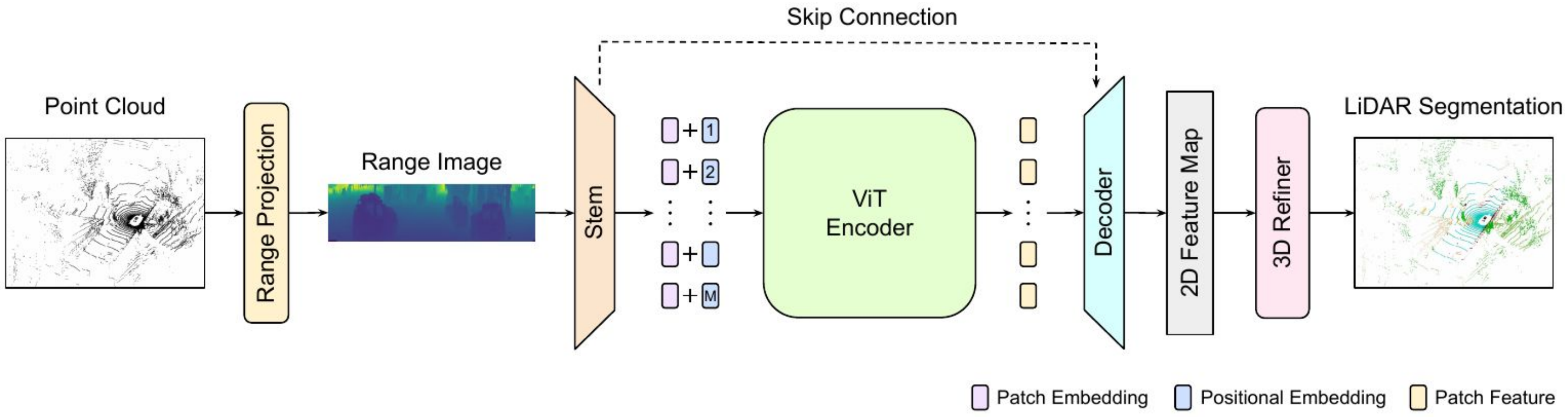


UpConv Decoder



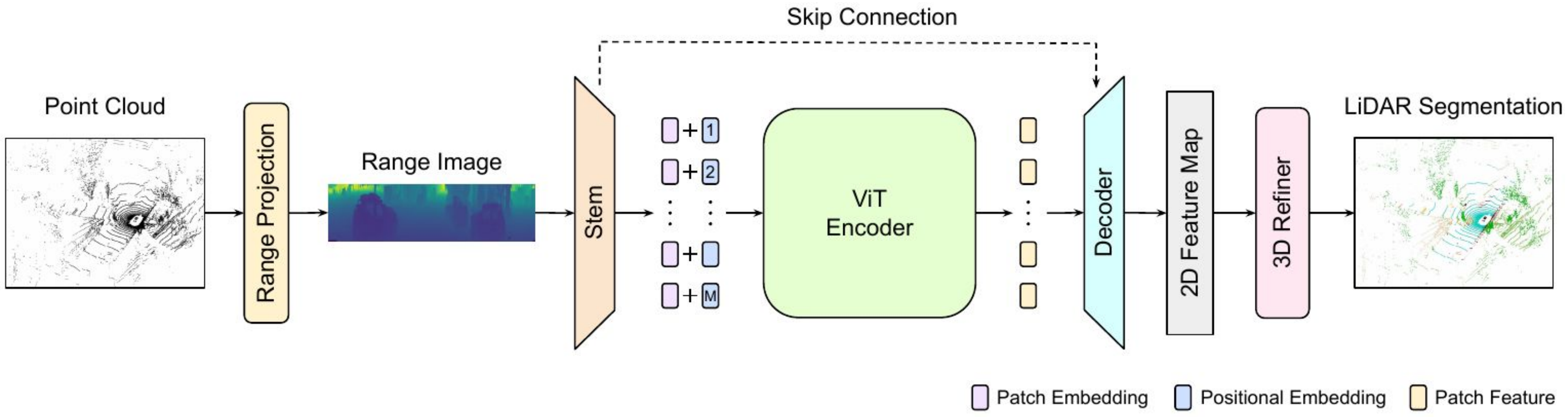
Stem	Decoder	Refiner	mIoU	#Params
Linear	Linear		65.52	22.0M
Conv	Linear		69.82	22.8M
Conv	UpConv		73.83	24.6M
Conv	UpConv	✓	**74.60**	25.2M

What patch-size for range image "tokenization"?



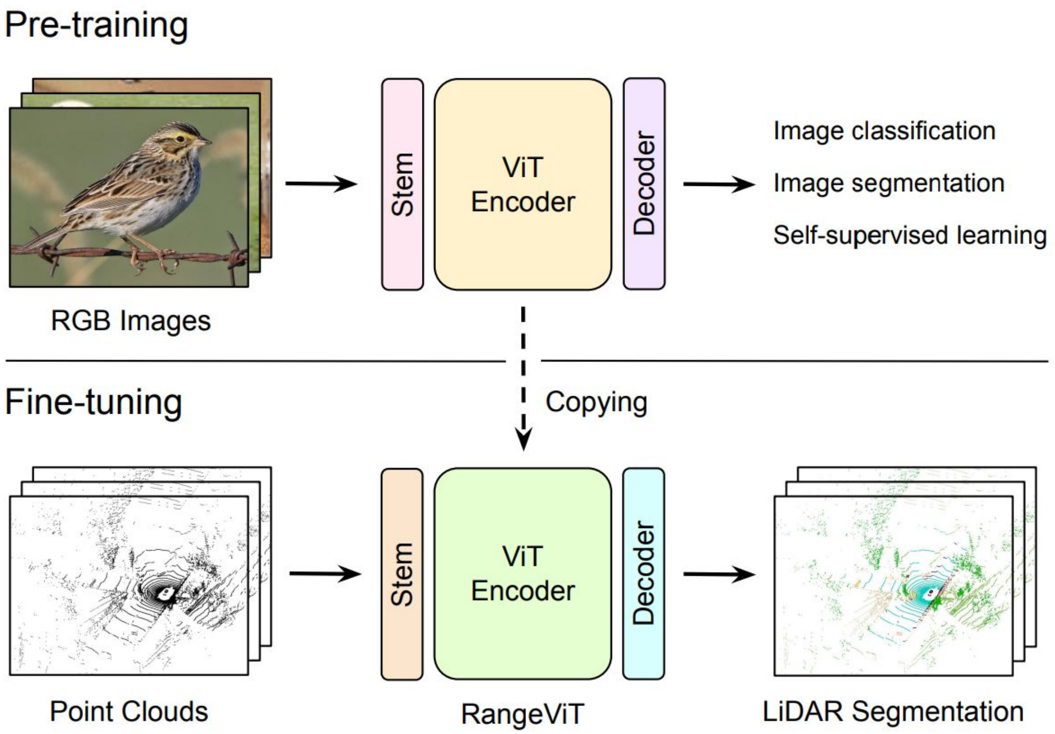
Patch size	16×16	8×8	4×16	4×8	4×4	2×16	2×8
mIoU	68.45	72.04	72.72	73.30	73.70	73.88	**75.21**
#Tokens	49	193	193	385	769	385	769
Train time	$\times 1$	$\times 1.02$	$\times 1.02$	$\times 1.13$	$\times 1.43$	$\times 1.13$	$\times 1.43$

What patch-size for range image “tokenization”?



Patch size	16×16	8×8	4×16	4×8	4×4	2×16	2×8
mIoU	68.45	72.04	72.72	73.30	73.70	73.88	**75.21**
#Tokens	49	193	193	385	769	385	769
Train time	$\times 1$	$\times 1.02$	$\times 1.02$	$\times 1.13$	$\times 1.43$	$\times 1.13$	$\times 1.43$

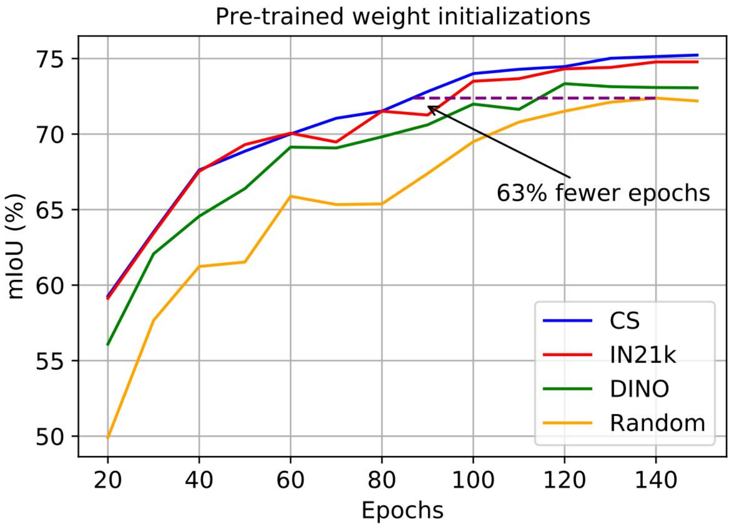
Exploiting image-pre-trained ViTs for LiDAR segmentation



Is pre-training on RGB images beneficial?

DINO: self-supervised pre-trained on ImageNet1k. **IN21k:** supervised on ImageNet21k.

Cityscapes: supervised on ImageNet21k + supervised on Cityscapes.



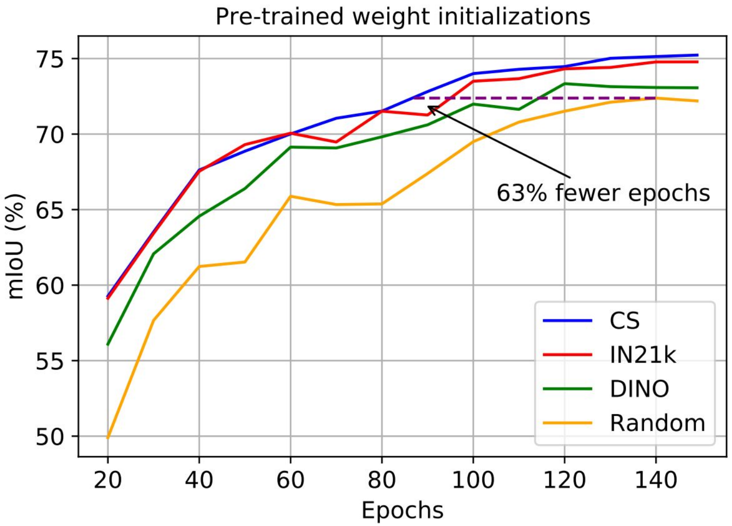
Pre-training	Rand	DINO	IN21k	Cityscapes
mIoU	72.37	73.33	74.77	**75.21**

Image-pretrained ViTs improve LiDAR segmentation performance and training efficiency.

Is pre-training on RGB images beneficial?

DINO: self-supervised pre-trained on ImageNet1k. **IN21k:** supervised on ImageNet21k.

Cityscapes: supervised on ImageNet21k + supervised on Cityscapes.



Pre-training	Rand	DINO	IN21k	Cityscapes
mIoU	72.37	73.33	74.77	**75.21**

Image-pretrained ViTs improve LiDAR segmentation performance and training efficiency.

Which ViT layers are better to fine-tune?

DINO: self-supervised pre-trained on ImageNet1k. **IN21k:** supervised on ImageNet21k.

Cityscapes: supervised on ImageNet21k + supervised on Cityscapes.

	Fine-tuning			IN21k	Cityscapes
Model	LN	ATTN	FFN	mIoU	
(a)	✓	✓	✓	74.79	75.21
(b)				67.88	68.03
(c)	✓			69.08	69.31
(d)	✓	✓		73.56	72.77
(e)	✓		✓	75.11	75.47

Partial fine-tuning of ViT backbone.

Which ViT layers is better to fine-tune?

DINO: self-supervised pre-trained on ImageNet1k. **IN21k:** supervised on ImageNet21k.

Cityscapes: supervised on ImageNet21k + supervised on Cityscapes.

Model	Fine-tuning LN	Fine-tuning ATTN	Fine-tuning FFN	IN21k mIoU	Cityscapes mIoU
(a)	✓	✓	✓	74.79	75.21
(b)				67.88	68.03
(c)	✓			69.08	69.31
(d)	✓	✓		73.56	72.77
(e)	✓		✓	75.11	75.47

Partial fine-tuning of ViT backbone.
The best results are achieved when the attention layers remain frozen.

Transferring to LiDAR segmentation: ViTs vs ResNet

Encoder	ViT-S†	ViT-S	RN50†	RN50
mIoU (%)	67.88	**74.77**	60.48	72.30

†: pretrained and fixed encoder
ViT-S and ResNet50 encoders pre-trained on IN21k.

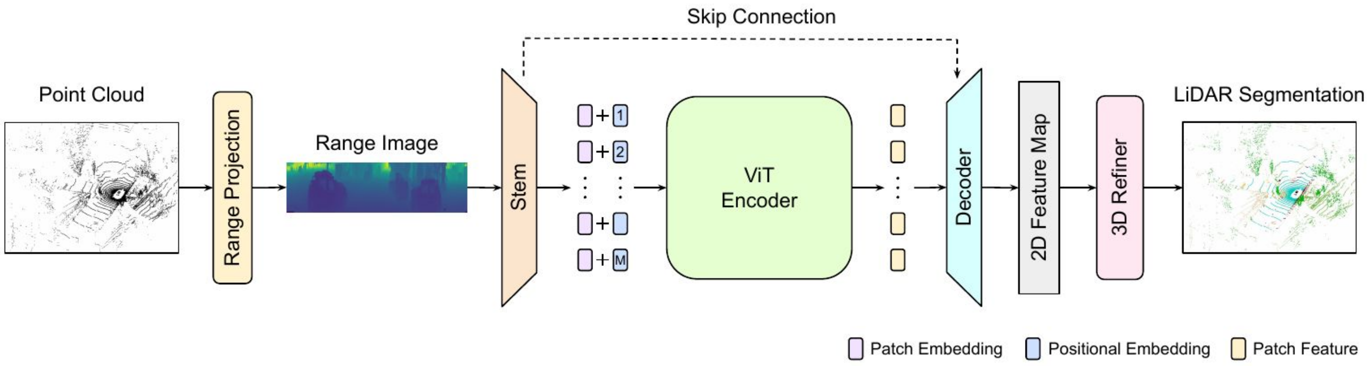
Image-pre-trained ViTs are more effectively transferred than ResNets.

Comparison to the state-of-the-art

Method	nuScenes mIoU (%)	SemKITTI mIoU (%)	#Params	Inference Time
Voxel-based				
Cylinder3D	**76.1**	**67.8**	55.9M	49 ms
2D Projection-based				
RangeNet++	65.5	52.2	-	-
PolarNet	71.0	54.3	-	-
SalsaNext	72.2	59.5	6.7M	28 ms
KPRNet	-	63.1	213.2M	-
Lite-HDSeg	-	63.8	-	-
RangeViT-CS (ours)	**75.2**	**64.0**	27.1M	25 ms

RangeViT outperforms prior projection-based segmentation methods, reducing the gap with the strong voxel-based Cylinder3D method.

Conclusions



- **RangeViT surpasses prior projection-based** methods.
- **Unifies architectures** in the LiDAR and image domains. ⇒ Any advance in one domain benefits both.
- **ViTs pre-trained on RGB images can be effectively transferred** for LiDAR point cloud segmentation.

We thank you very much for your attention!