SELF-SUPERVISED CLUSTERING - NO TURNING ‘BACK’

Rishabh Sheoran, Saisha Ajay Chhabria, Sourav Dey, Umaiyal Ramanathan

NUS School of Computing CS5260 Group Project Progress Update, Apr 2022

Background

Recent results [1] suggest that the Transformer architecture has superseded earlier Convolution Neural Networks (CNNs) in many areas of classification including self-supervised learning (SSL) and classification [2]. Sample below.

Method	Backbone	CIFAR10
SimCLR	ResNet-56	78.75%±0.24
SiT-NonLinProj	Transformer	83.50%±0.11

*SiT-NonLinProj was trained to minimize loss on 3x pretext tasks

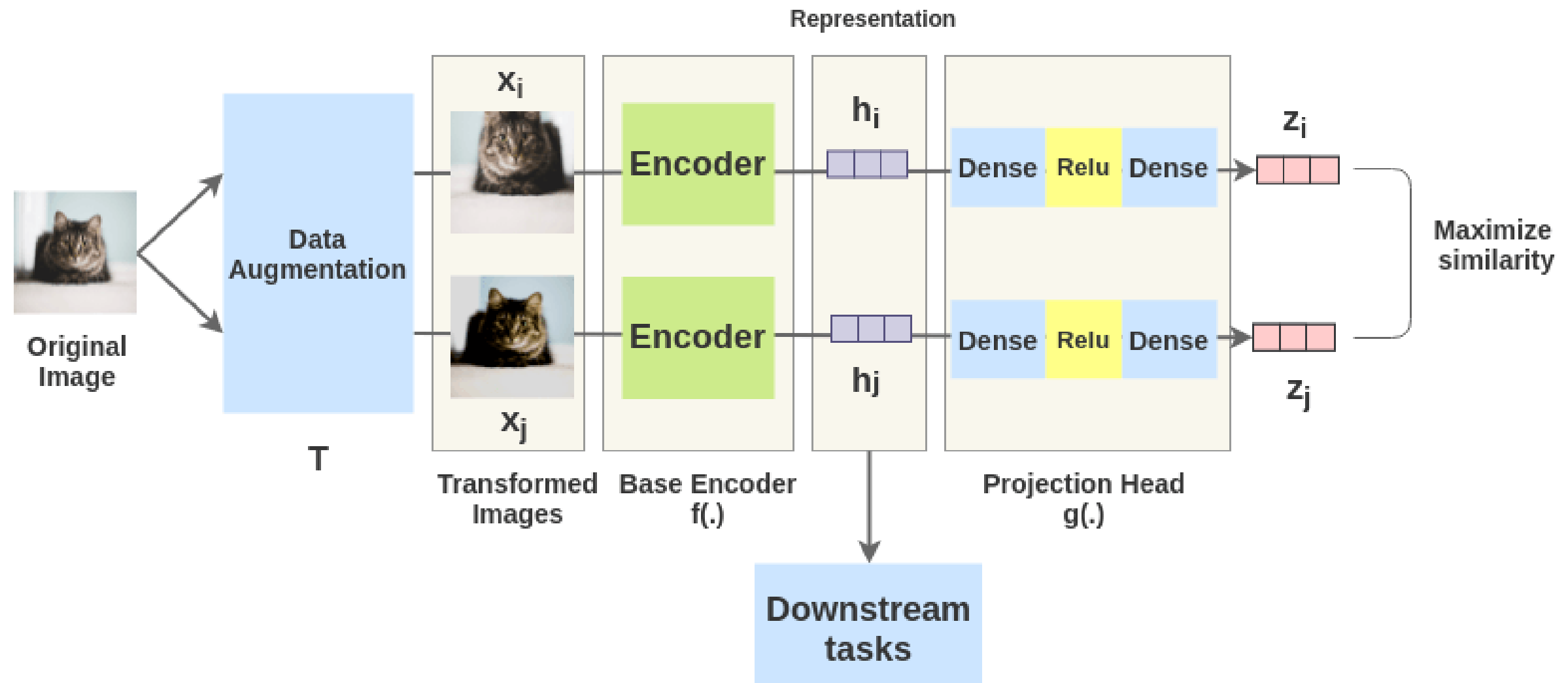
Meanwhile, to revive CNN, ConvNeXt was developed. ConvNeXt is based on a new hybrid architecture [3] that retains the translation invariance property from CNN whilst incorporating the multi-headed self-attention mechanism in Transformers. The ConvNeXt architecture has yet to be tested as a backbone in SSL.

Objective

1. To determine the optimal base encoder backbone architecture (ResNet50 vs ViT vs ConvNeXt) for self-supervised clustering when training on a single pretext task to minimize contrastive loss. [SimCLR]
2. To investigate optimization and scheduling schemes best suited for minimizing contrastive loss.
3. To apply an appropriate SSL architecture from above on remote sensing images and evaluate the performance.

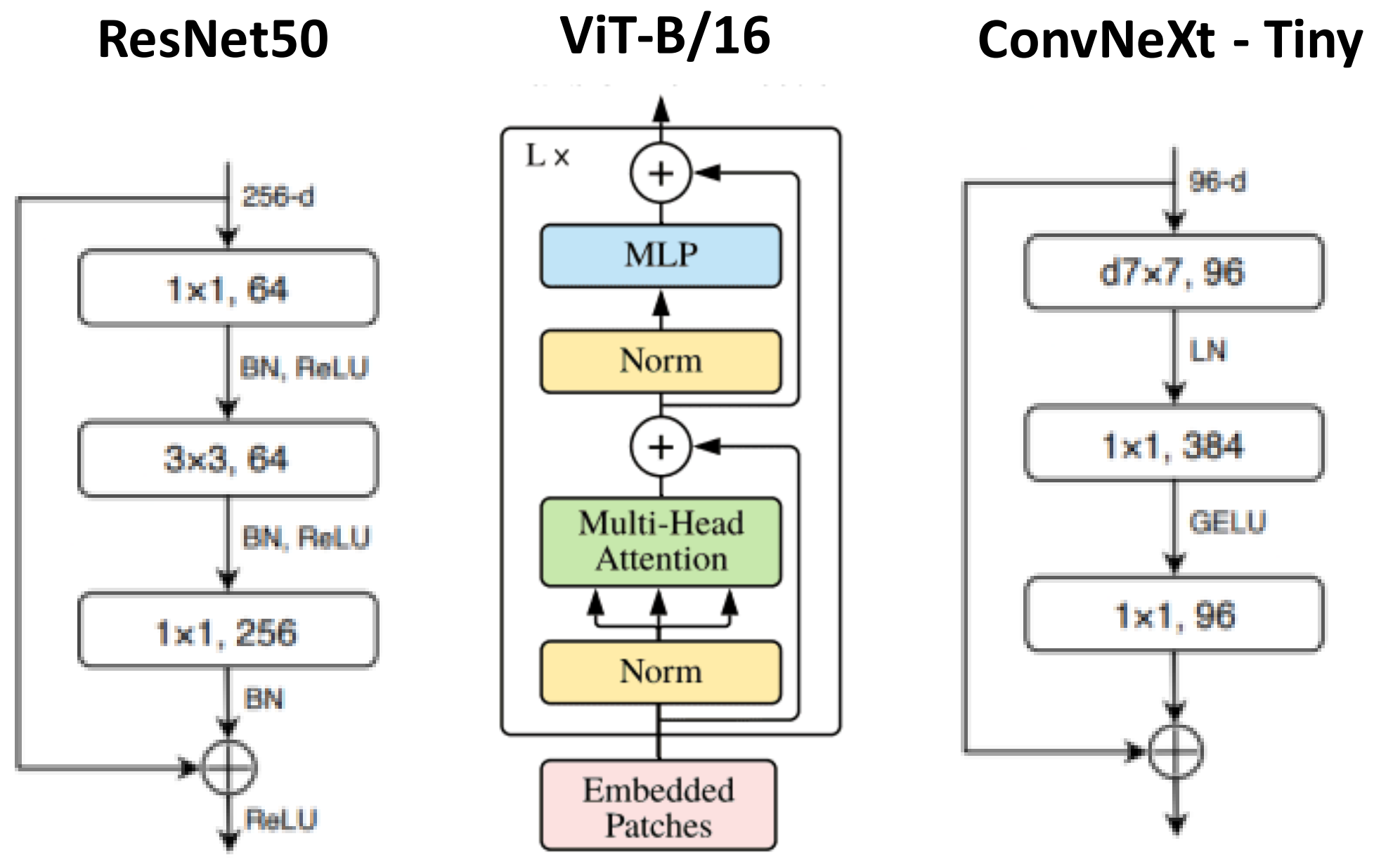
Methodology

SimCLR Framework



The Encoder in the SSL architecture is implemented and tested using three different backbones:

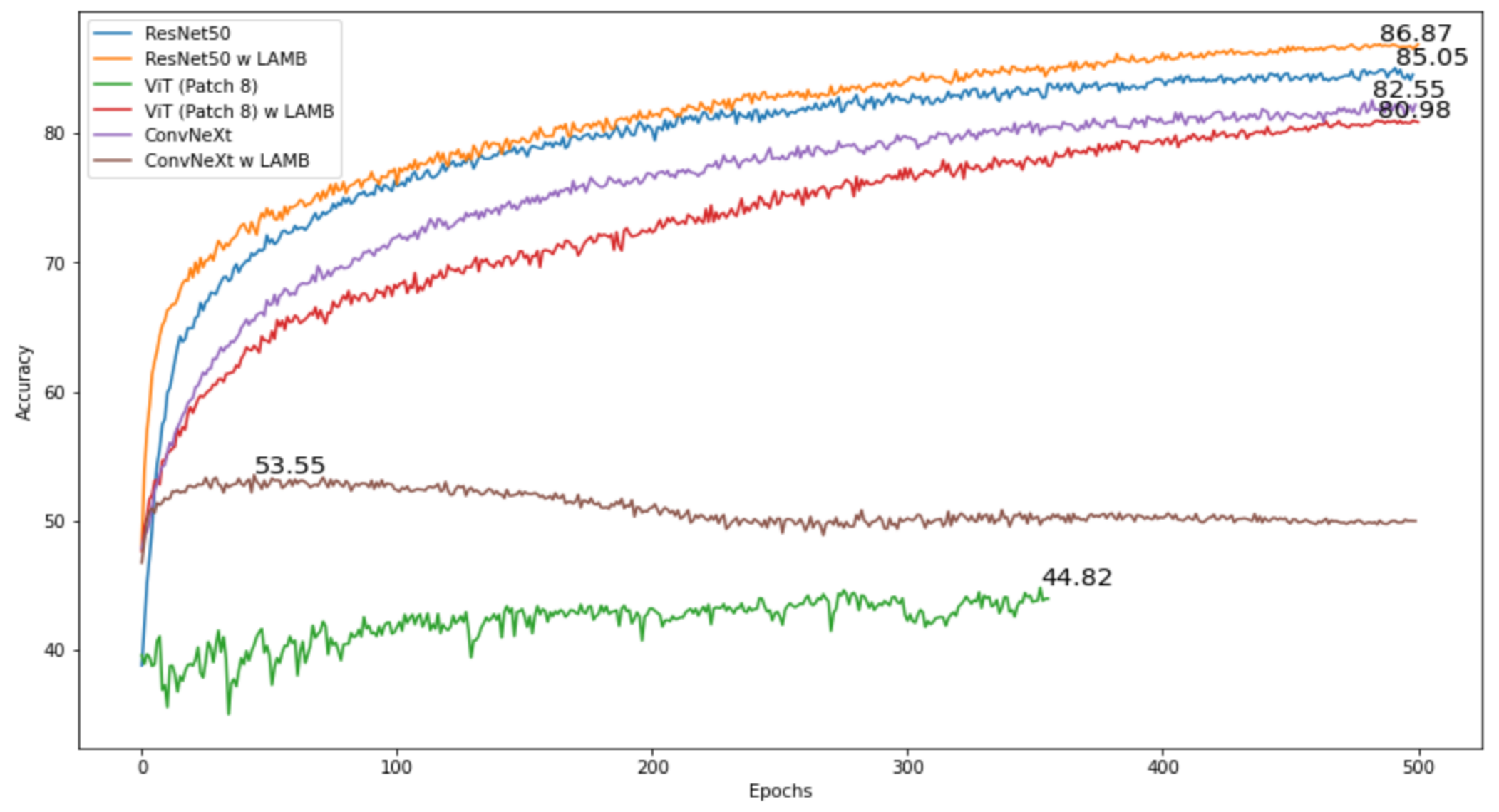
1. Resnet50
2. ViT-B
3. ConvNeXt



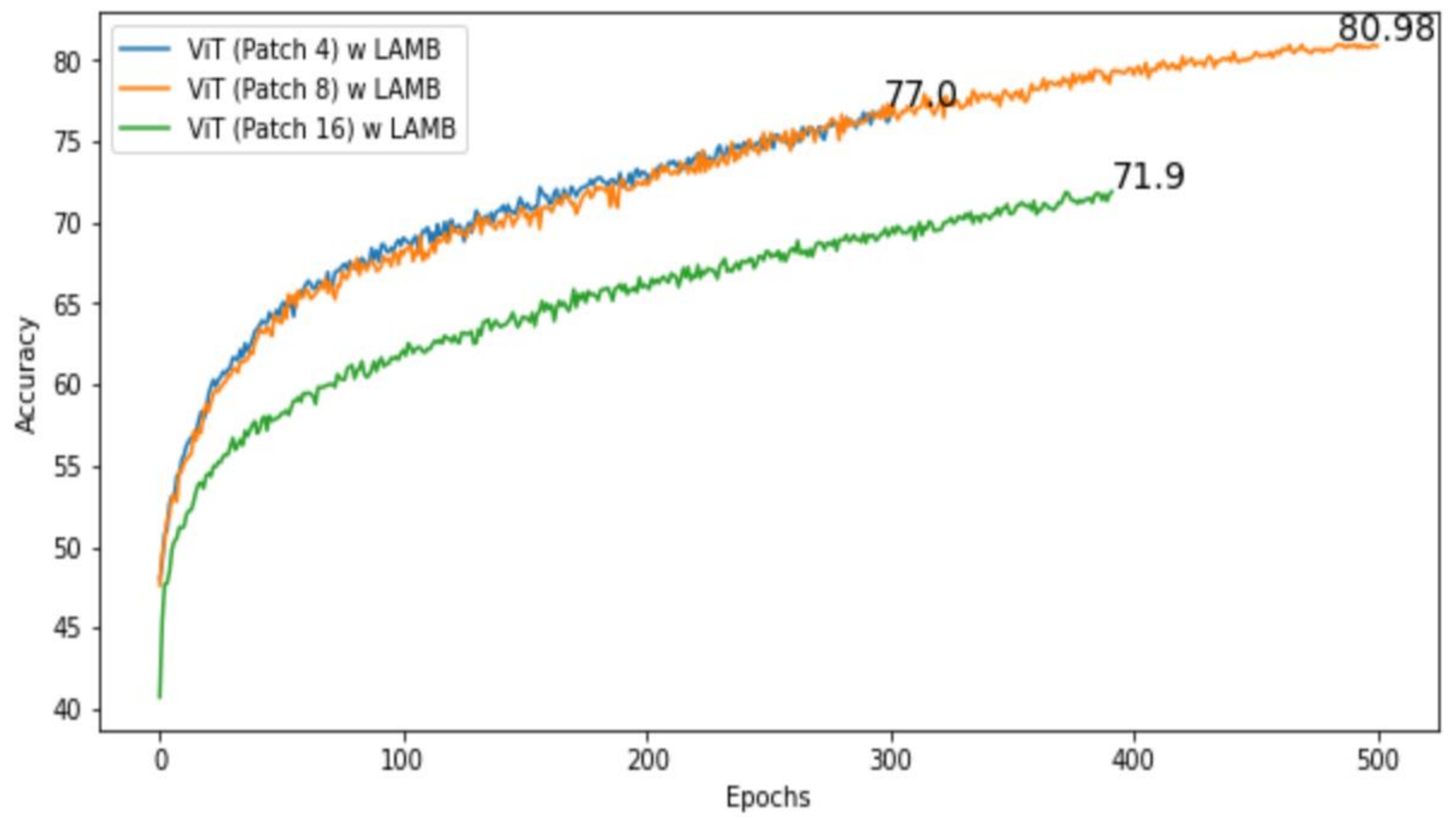
	ResNet50	ViT-B/16	ConvNeXt-Tiny
#Parameters	24.62M	62.75M	28.69M

- The pretext task is performed on CIFAR-10 training data to reduce the contrastive loss.
- The downstream task is performed on CIFAR-10 test data to perform unsupervised classification.

Preliminary Results of SimCLR Backbone Comparison on CIFAR 10



SimCLR w/ ResNet50 outperforms other backbones, with or without LAMB optimization. Likely, SiT outperforms SimCLR w/ ResNet50 backbone in [2] because SiT contains more pretext tasks than SimCLR and not due to backbone choice.



Smaller patch size in the Vision Transformer Encoder improved the model performance. This is likely due to the smaller patch sizes enabling the model to encode more granular information from the image.

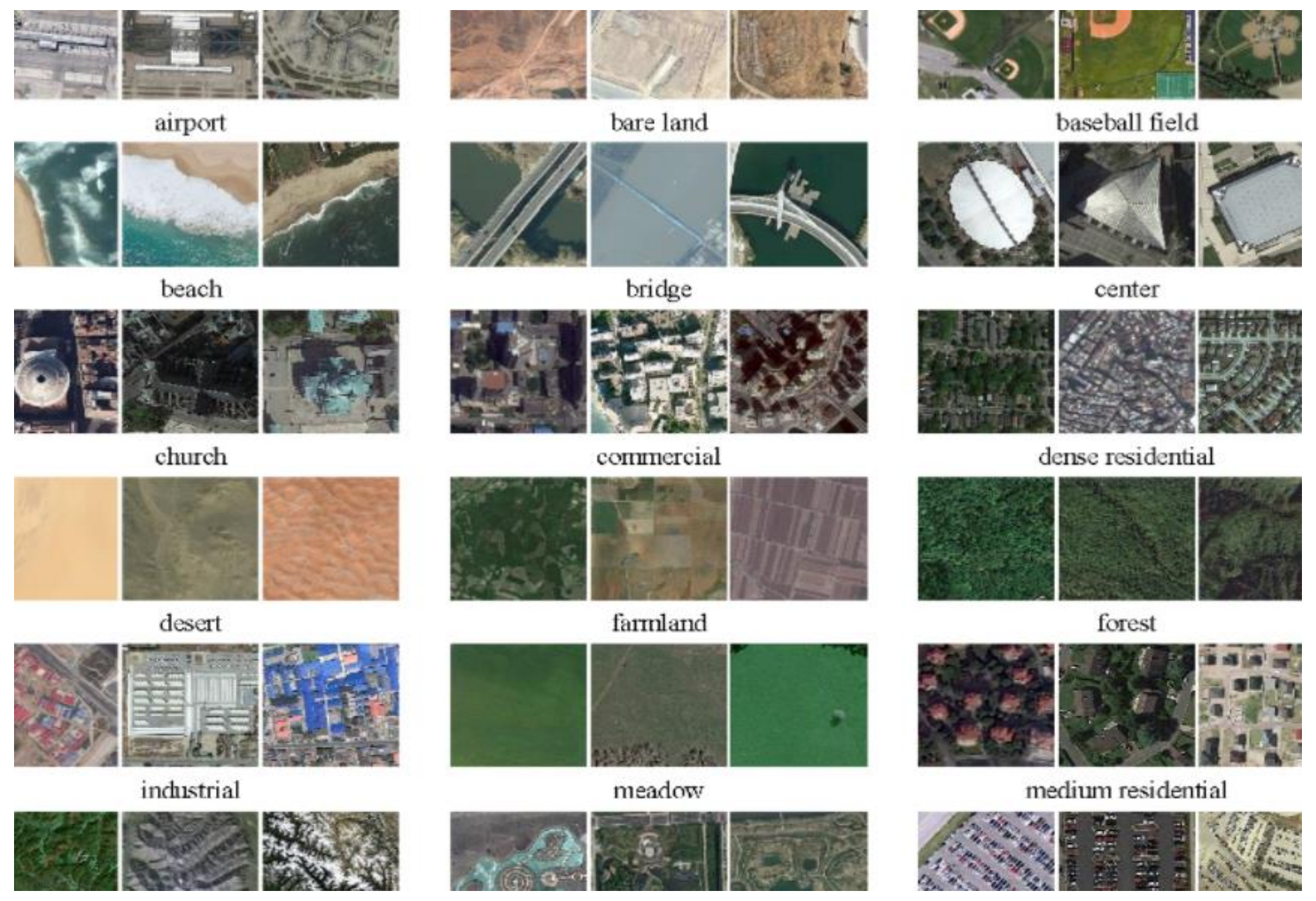
Application on Remote Sensing Imagery

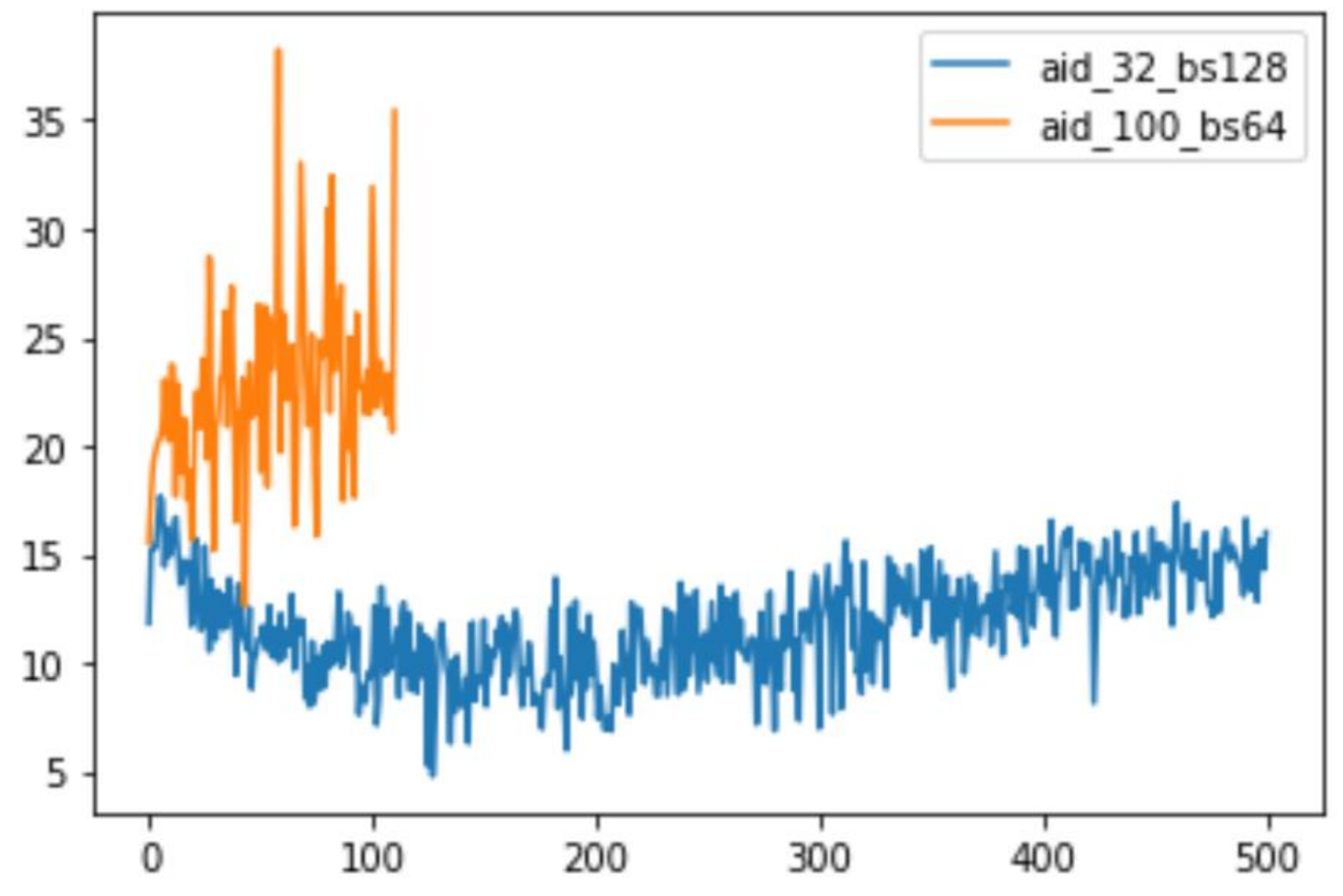
An exponential increase (>10000 PB) is forecasted in aerial imagery [4] in the next decade owing to the increasing number of high-resolution constellation satellites. These sensors range across several spectral bands (Panchromatic, Multispectral, VHF, C band, X band Synthetic Aperture Radar-SAR) and their imaging geometries are vastly different. Their volume, variety and velocity make labeling and thus supervised classification an impossible task. Label free SSL and classification/semantic scene segmentation with continuous retraining is the way to go to automate remote sensing imagery exploitation.

Preliminary Results on AID Dataset

AID[5] Aerial Data Set

- 5000/5000 train & test images
- Size: 600x600 belonging to 30 classes





The chip-size certainly affects the performance of remote sensing images. The larger the chip, the more contextual features available for better classification. But current GPU constraints cannot fit more than 32x32 chip-size for batch size 128.

Progress Checklist

Backbone	Optimizer	Variable	Domain	Status
ResNet	ADAM	-	CIFAR10	Completed
	LAMB	-		Completed
ViT	ADAM	Patch: 16		In Progress
	LAMB	Patch: 4		In Progress
		Patch: 8		Completed
		Patch: 16		In Progress
ConvNeXt	ADAM	Tiny		Completed
	LAMB	Tiny		Completed
Best Backbone & Optim		-	AID	In Progress
		-	Pol SAR	To commence

*All runs to be repeated 3x

Novelty

✓ **Novelty of Algorithm**

- No literature on a fair comparison of CNN vs Transformer backbone when trained with a single contrastive loss minimization pretext task (SimCLR).
- Whether ConvNeXt can supersede Transformer architecture for SSL classification is yet to be published.

✓ **Novelty of Application**

- Recent first works of applying SSL on Multispectral images for unsupervised classification and semantic scene segmentation have been published [6, 7].
- But no publication yet on polarimetric SAR images.

Challenges

- Obtaining GPU resources for compute intensive tasks.
- Semantic scene segmentation.

References

1. Dosovitskiy, A. et al(2021). “An Image is worth 16X16 Words: Transformers for Image Recognition at Scale, ICLR 2021.” from the arXiv database
2. Atito, S. Awais, M. and Josef Kittler, J. (Nov 2021). “Self-supervised vision Transformer.” from the arXiv database
3. Liu, Z. Mao, H. Wu, Feichtenhofer, C. Darrell, T. Xie, S. (Mar 2022). “A ConvNet for the 2020s.” from the arXiv database
4. “DLR-Earth Observation Center – 60 Petabytes for the German Satellite Data Archive D-SDA.” Startseite – DLR Portal, https://www.dlr.de/eoc/en/desktopdefault.aspx/tabid-12632/22039_read-51751
5. Risojevi´c, V. Stojni´c, V. (Dec 2021) “The Role of Pre-Training in High-Resolution Remote Sensing Scene Classification.” from the arXiv database
6. Li, H. Li, Y. Zhang, G. Liu, R. Huang, H. Zhu, Q. Tao, C. (Jan 2022) “Global and Local Contrastive Self-Supervised Learning for Semantic Segmentation of HR Remote Sensing Images.” from the arXiv database.
7. Caron, Mathilde, et al. “Emerging Properties in Self-Supervised Vision Transformers.” 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021,

RESEARCH POSTER PRESENTATION DESIGN © 2019
www.PosterPresentations.com