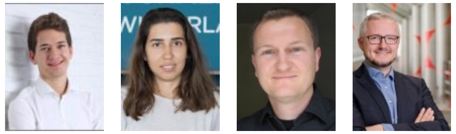
Universität St.Gallen

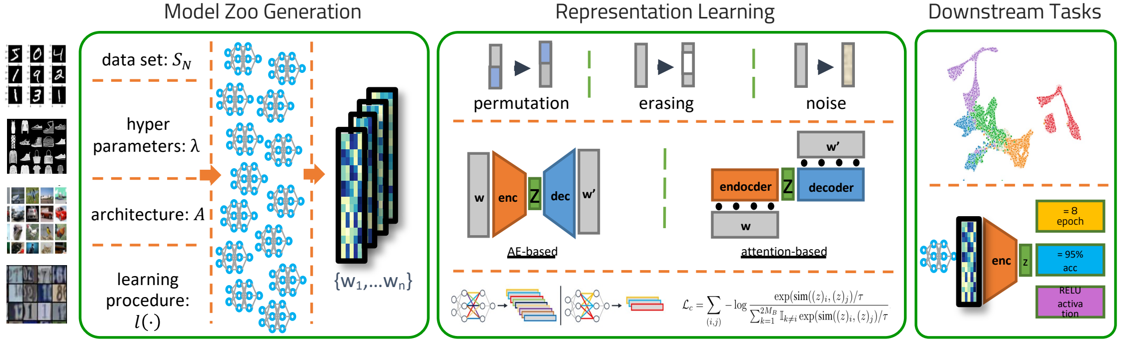
Efficient Representation Learning for Earth Observation & Remote Sensing

Linus Scheibenreif, Joëlle Hanna, Michael Mommert, Damian Borth

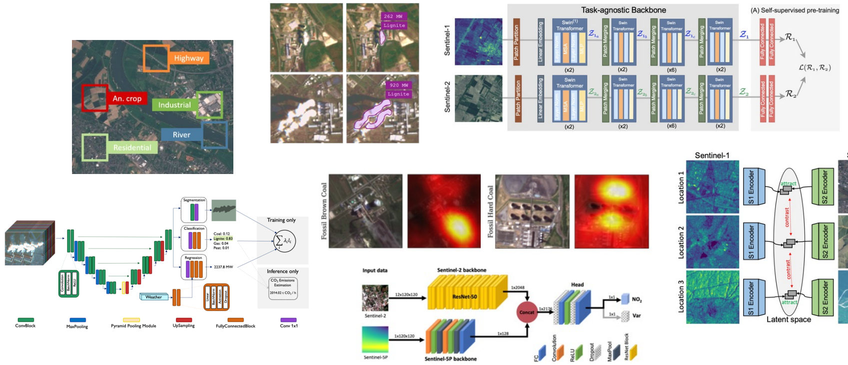


Current Research

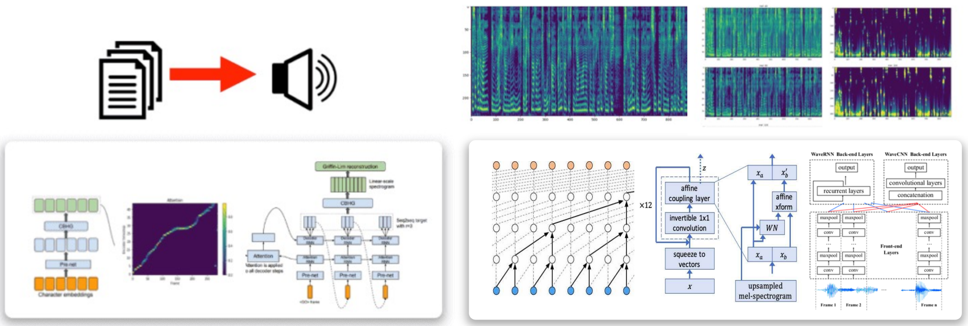
Representation Learning of Deep Neural Networks



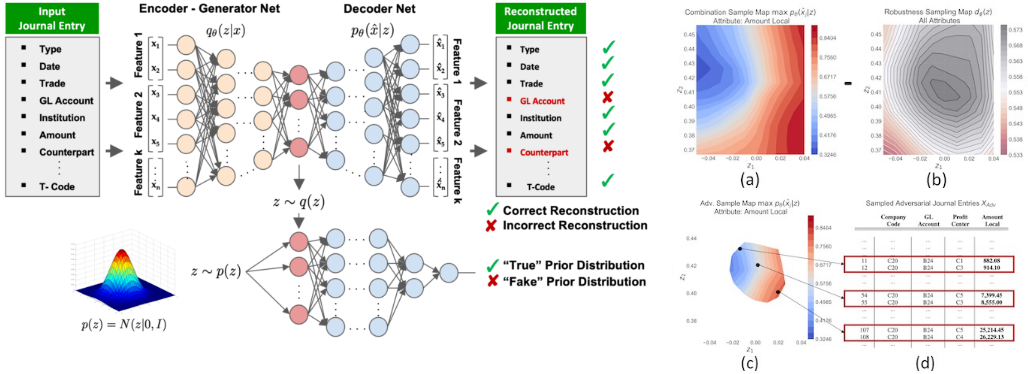
Remote Sensing & Earth Observation



Text-to-Speech Synthesis



Anomaly Detection in Transaction Data



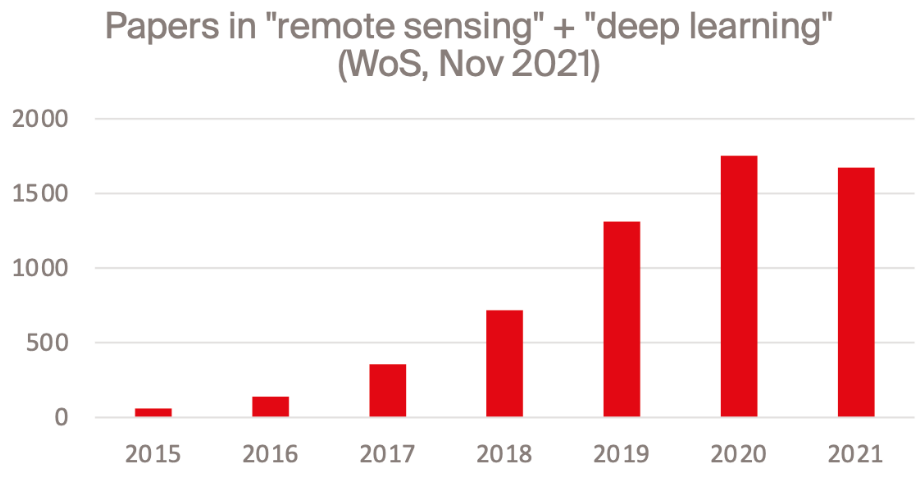
ESA / NASA & St.Gallen Symposium 2022

Invited panel discussion with ESA General Secretary **Josef Aschbacher**, NASA Astronaut **Edward Chang-Diaz** and Space Law Expers **Elyssia Gossler**

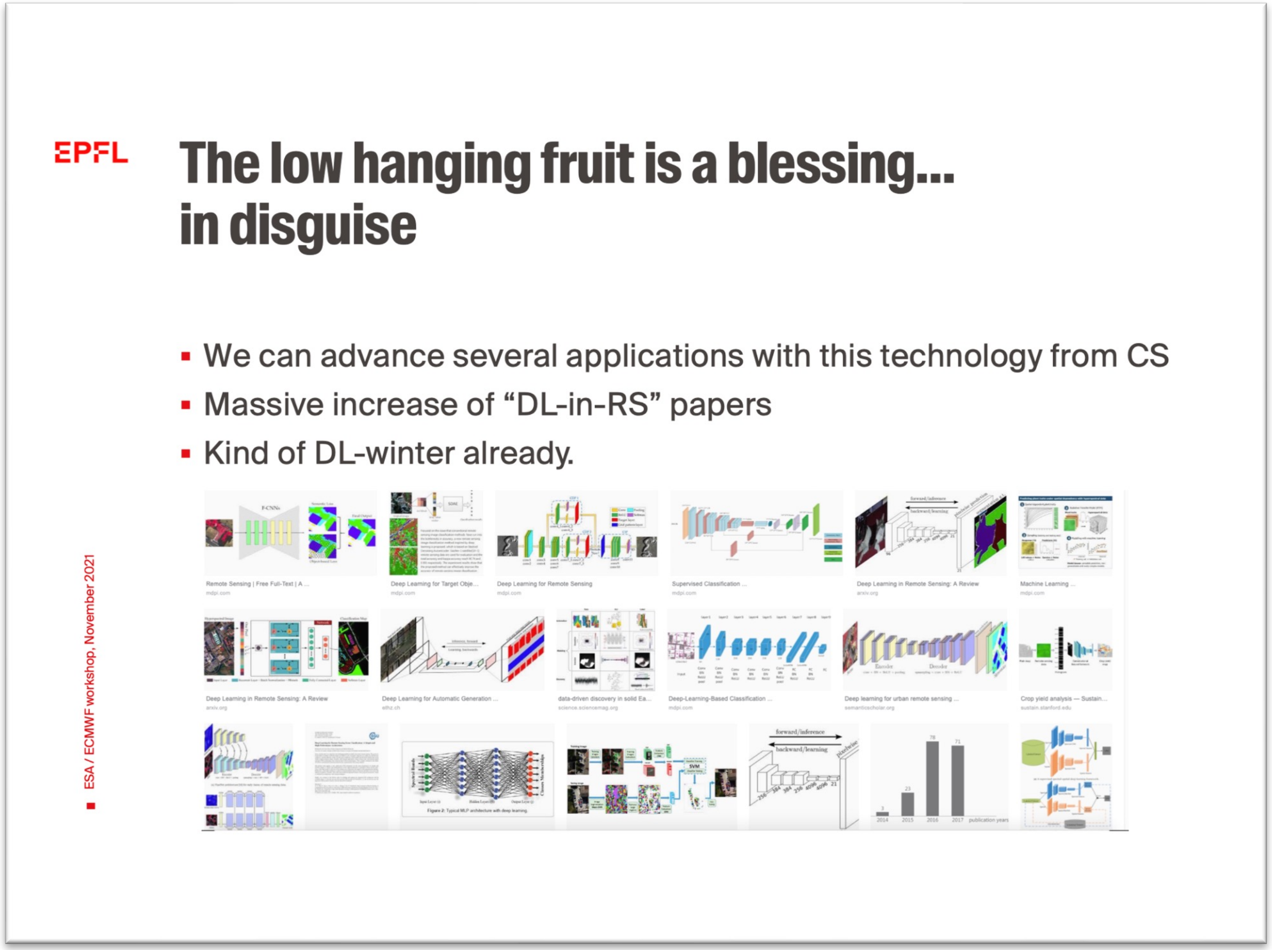
Efficient Representation Learning

Deep Neural Networks & Remote Sensing

Last Year’s ESA/ECMWF Machine Learning Workshop

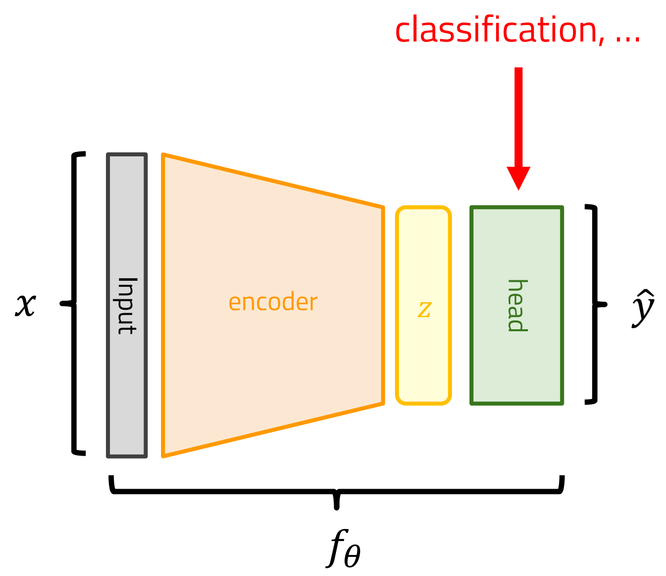


Prof. Devis Tuia, EPFL

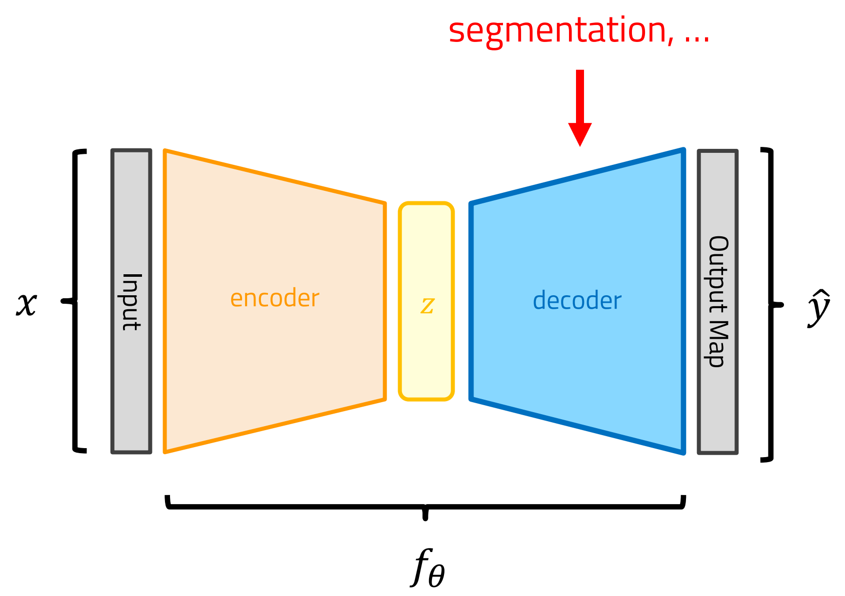


Deep Neural Networks = Representation Learning

Discriminative Tasks



Generative Tasks



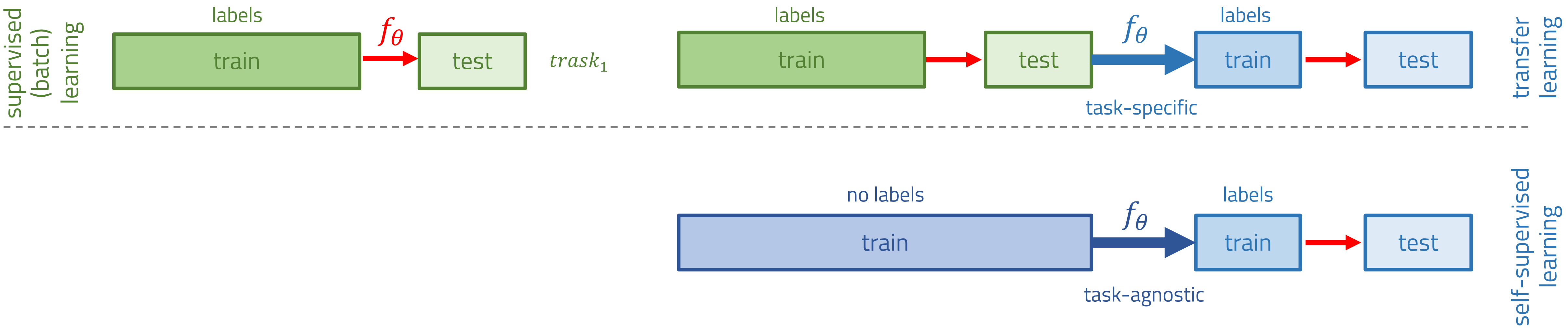
How can we be become more efficient in learning these representations?

Data and Model Efficiency

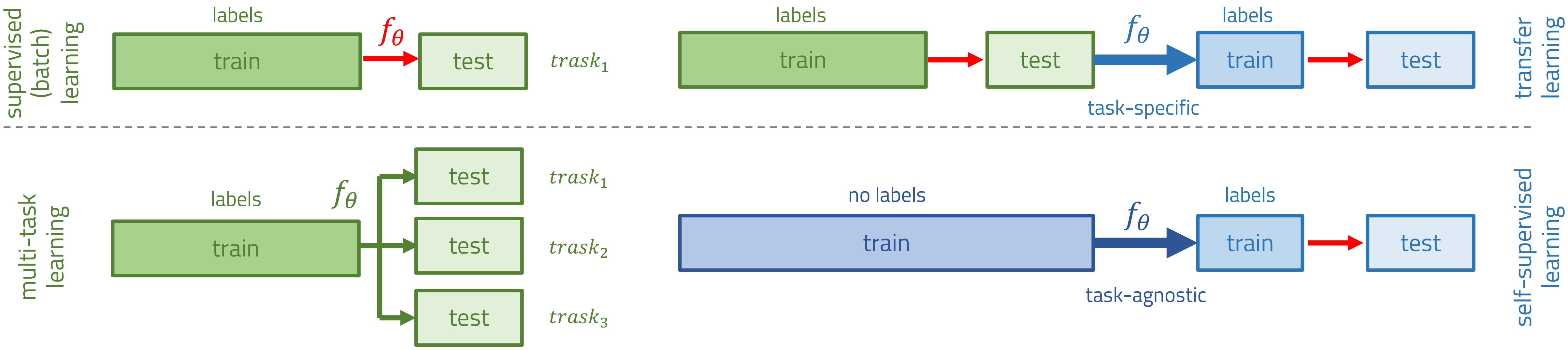
Data and Model Efficiency

Data and Model Efficiency

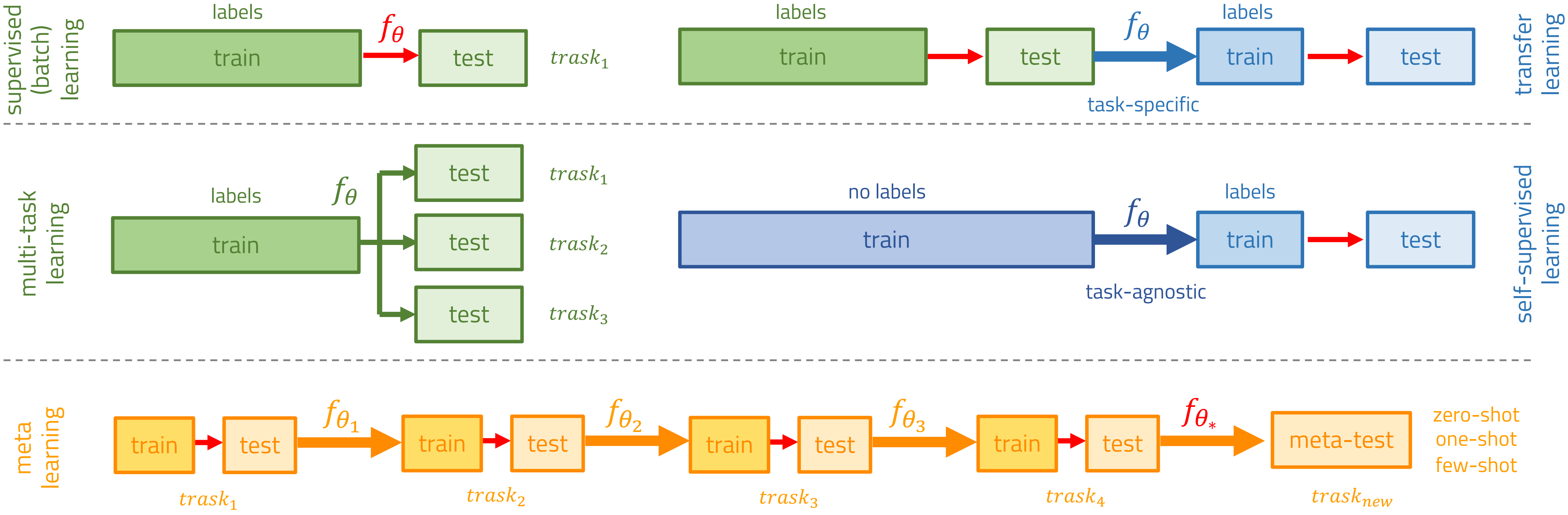
Data and Model Efficiency



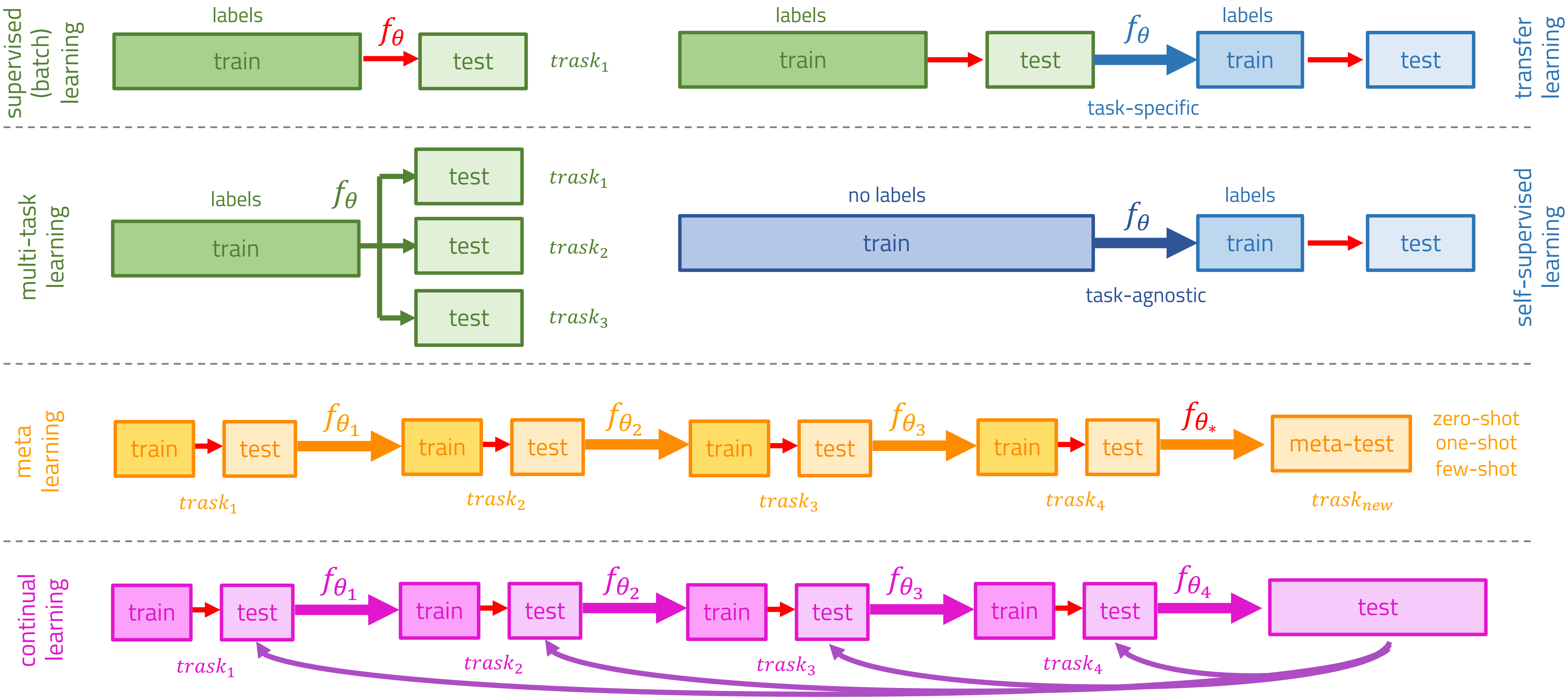
Data and Model Efficiency



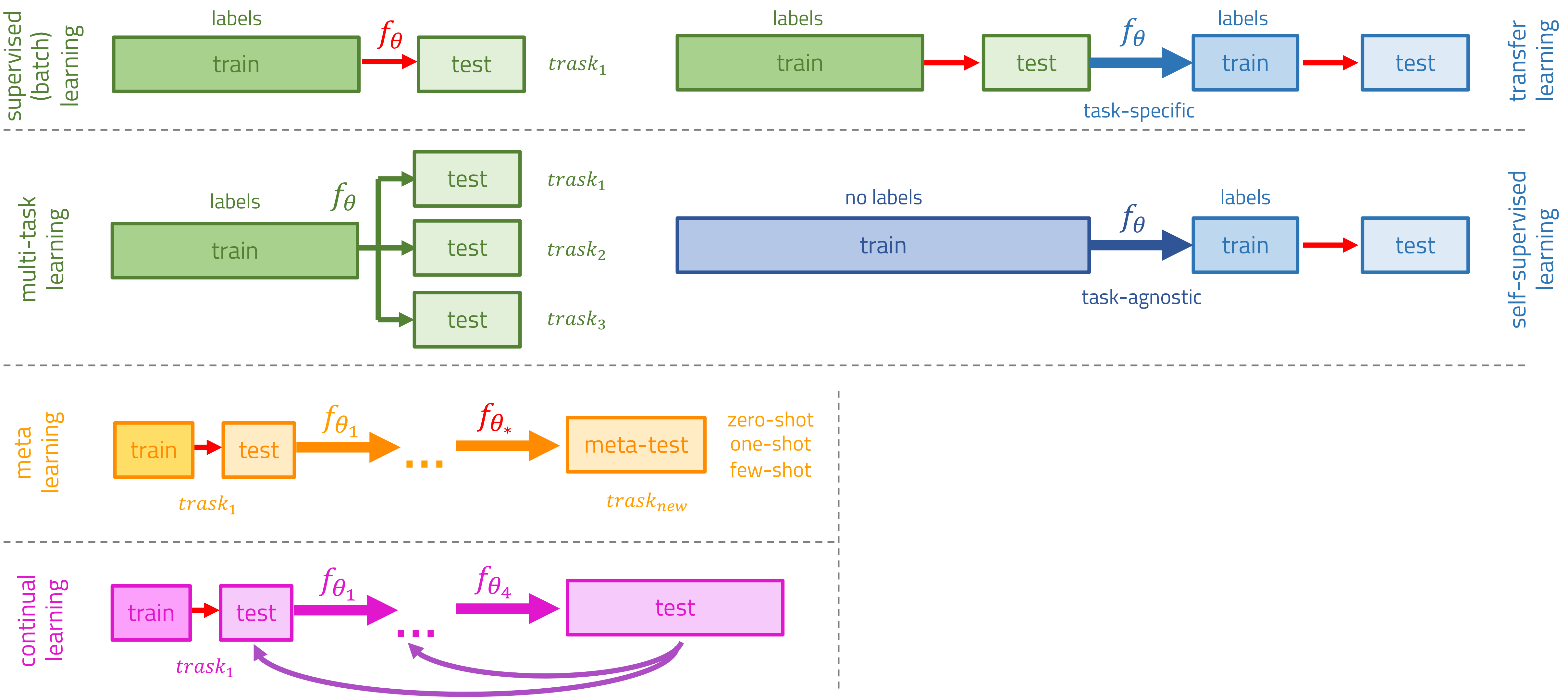
Data and Model Efficiency



Data and Model Efficiency

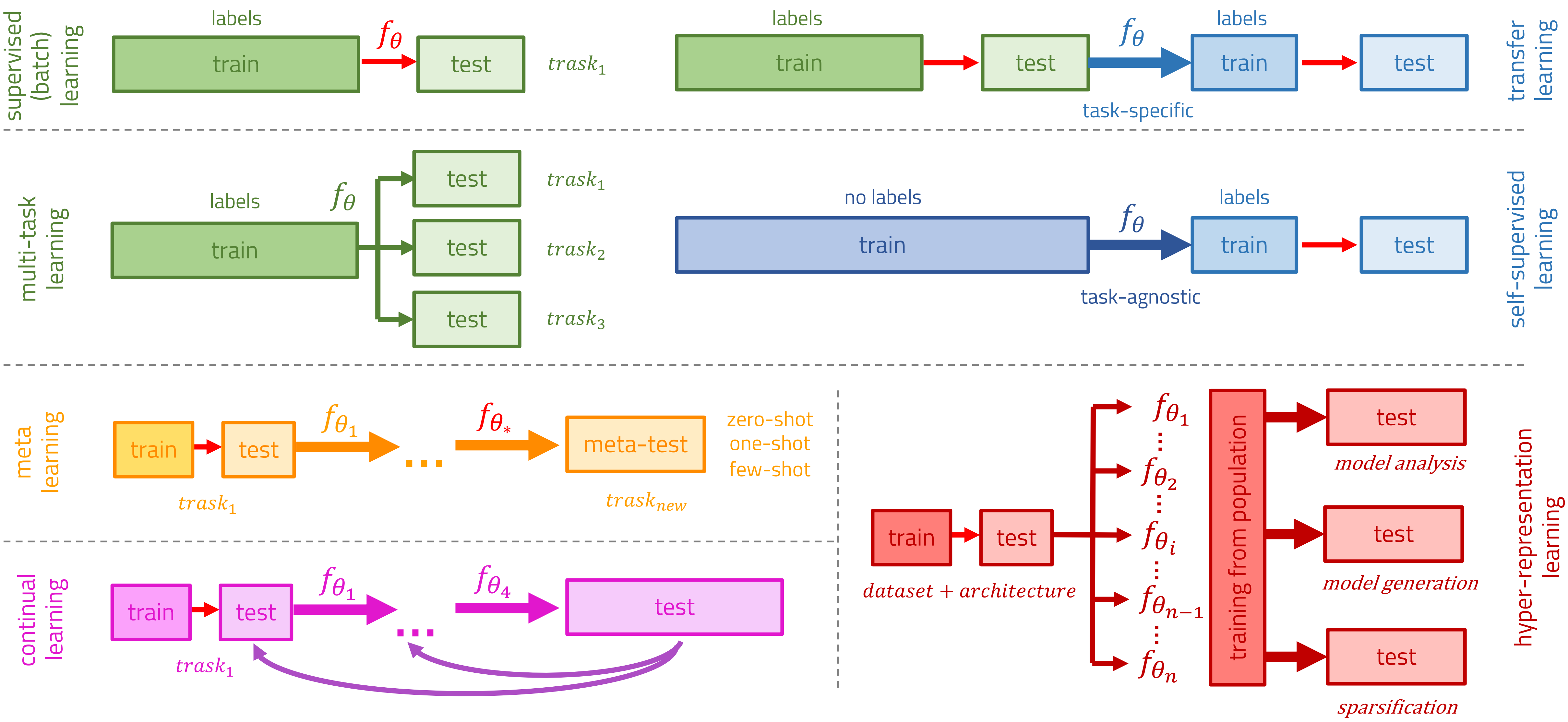


Data and Model Efficiency



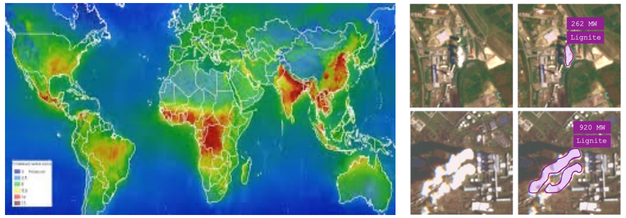
Universität St.Gallen

Data and Model Efficiency



Overview

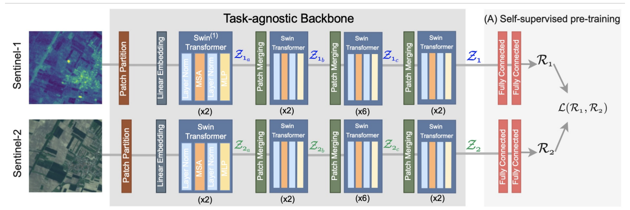
Shared-Backbones/Heads



Approach:
- Muilti-modal Fusion
- Multi-task Learning
- Auxiliary Tasks

Application
- NO2 estimation
- Power Production
- CO2 estimation

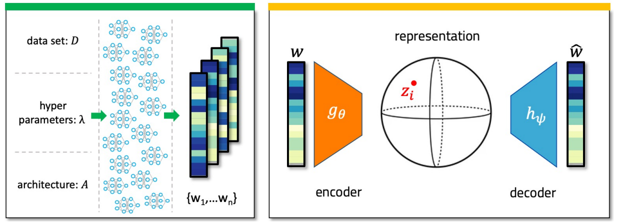
Self-supervised Learning



Approach:
- Contrastive Learning
- Augmentation free
- CNNs & Transformer

Application
- Land-use Classification
- Single-class / Multi-class
- Segmentation

Hyper-Representations



Approach:
- Contrastive Learning
- Model Zoos
- CNNs

Application
- Model analysis
- Sample unseen models
- Sparsificaiton

Universität St.Gallen

Shared Backbones / Heads

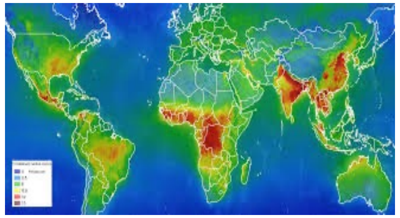
L Scheibenreif, M Mommert, D Borth
Toward Global Estimation of Ground-Level NO 2 Pollution With Deep Learning and Remote Sensing,
IEEE Transactions on Geoscience and Remote Sensing (TGSRS), March 2022

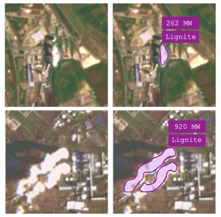
J Hanna, M Mommert, L Scheibenreif, D Borth
Multitask Learning for Estimating Power Plant Greenhouse Gas Emissions from Satellite Imagery,
NeurIPS Workshop on Tackling Climate Change with Machine Learning, 2021

Universität St.Gallen

Overview

Shared-Backbones/Heads

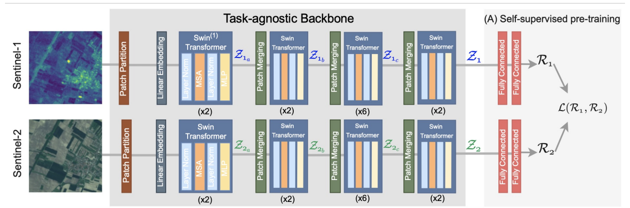




Approach:
- Muilti-modal Fusion
- Multi-task Learning
- Auxiliary Tasks

Application
- NO2 estimation
- Power Production
- CO2 estimation

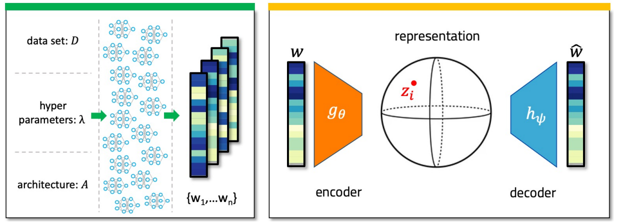
Self-supervised Learning



Approach:
- Contrastive Learning
- Augmentation free
- CNNs & Transformer

Application
- Land-use Classification
- Single-class / Multi-class
- Segmentation

Hyper-Representations



Approach:
- Contrastive Learning
- Model Zoos
- CNNs

Application
- Model analysis
- Sample unseen models
- Sparsificaiton

Ground Level NO_2 Pollution Estimation

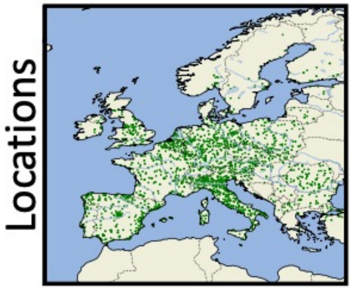
L Scheibenreif, M Mommert, D Borth
Toward Global Estimation of Ground-Level NO 2 Pollution With Deep Learning and Remote Sensing,
IEEE Transactions on Geoscience and Remote Sensing (TGSRS), March 2022

Universität St.Gallen

Ground Level NO_2 Pollution Estimation

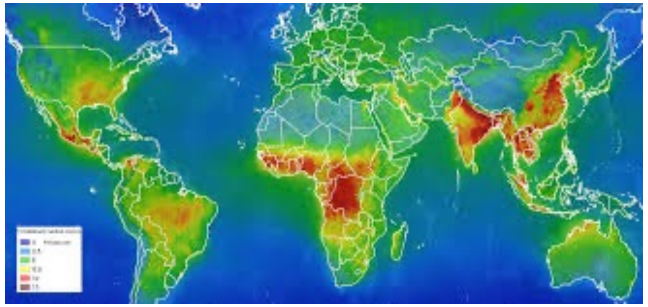
- European Environment Agency **Air Quality Stations**
  - Surface NO_2 measurements
  - 3000 locations in Europe

Ground truth NO_2



Google Streetview, 2 Keizerinlaan Steenokkerzee, Belgium

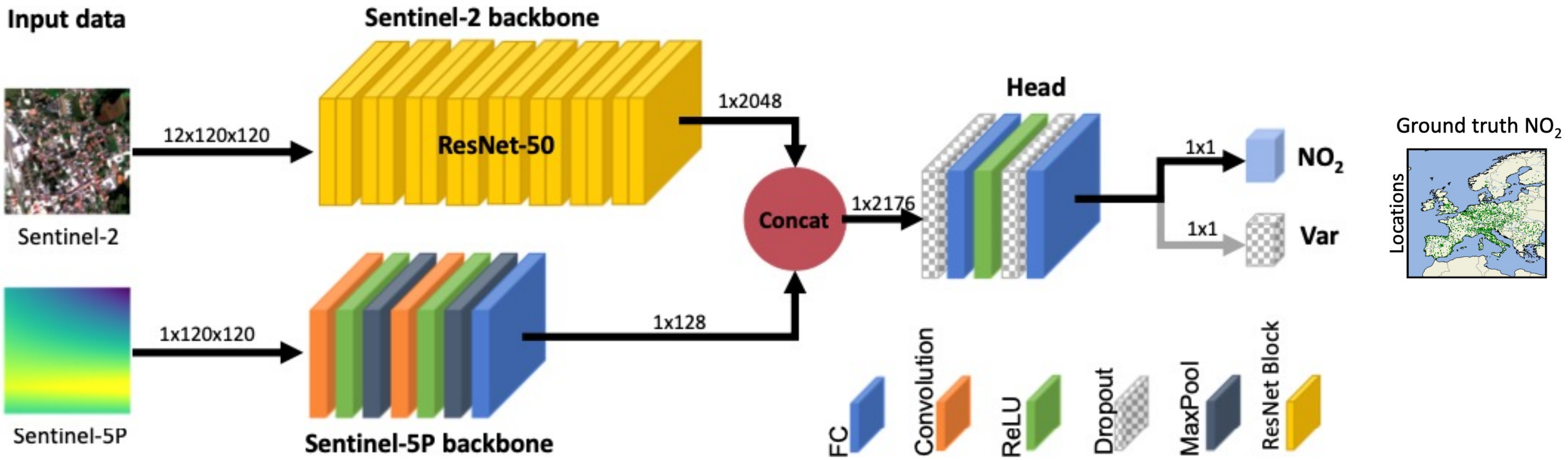
- esa **Sentinel-2**
  - Multi-spectral satellite imagery
  - 10 m resolution
- esa **Sentinel-5P**
  - Tropospheric NO_2 column density
  - 7x3.5 km resolution



L Scheibenreif, M Mommert, D Borth
Toward Global Estimation of Ground-Level NO 2 Pollution With Deep Learning and Remote Sensing,
IEEE Transactions on Geoscience and Remote Sensing (TGSRS), March 2022

Approach

Fusion: Separate Backbones + Shared Regression Head



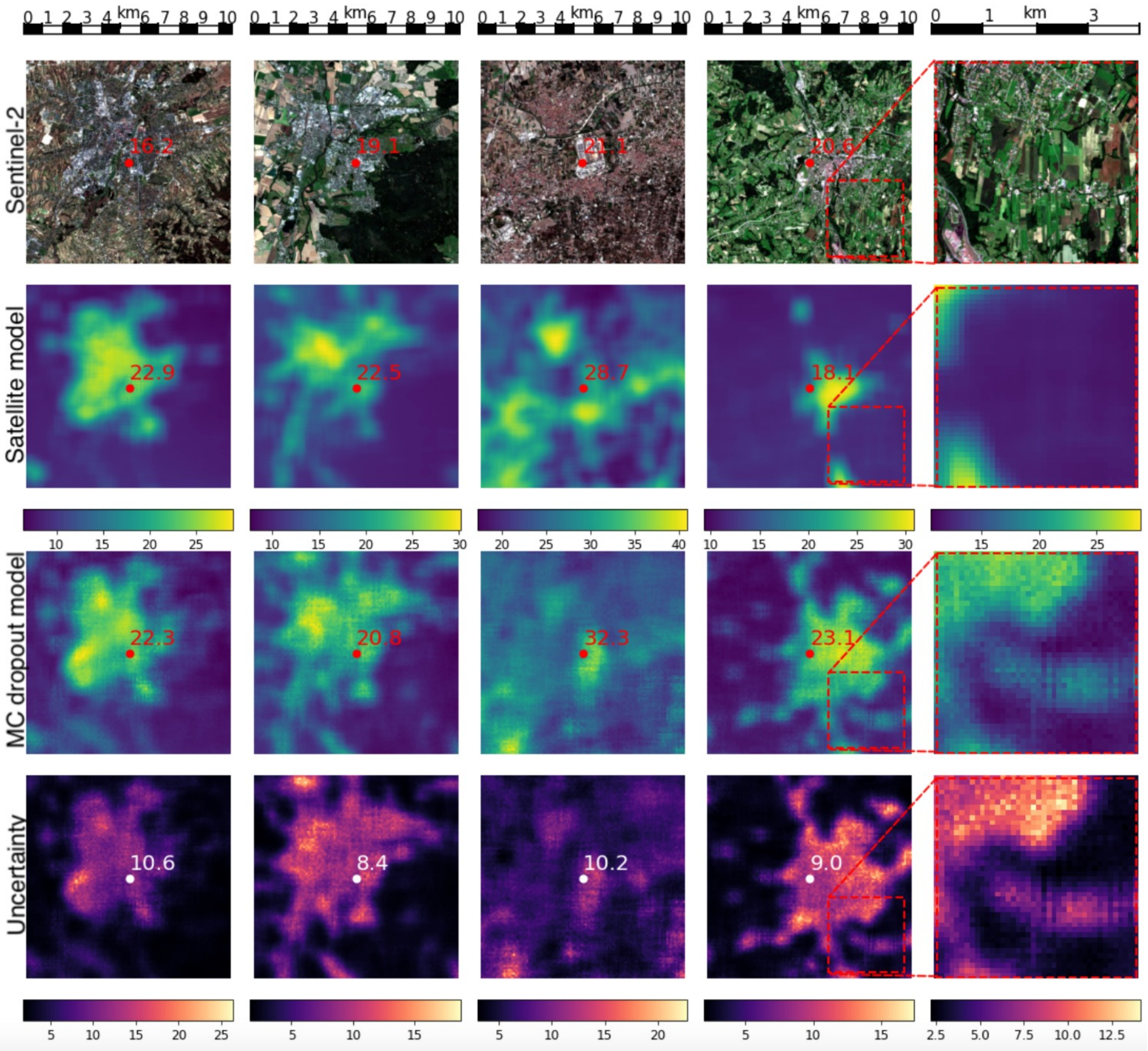
NN with dropout is mathematically equivalent to an approximation to the probabilistic deep Gaussian process

Y. Gal and Z. Ghahramani,
"**Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning**,"
in *International Conference on Machine Learning*. PMLR, 2016, pp. 1050–1059.

$$\mathbf{y}_i \sim N(f^\theta(\mathbf{x}_i), g^\theta(\mathbf{x}_i)^{-1})$$

$$L(\mathbf{x}, \mathbf{y}) = \frac{1}{2}(\mathbf{y} - f^\theta(\mathbf{x}))g^\theta(\mathbf{x})(\mathbf{y} - f^\theta(\mathbf{x}))^T - \frac{1}{2}\log \det\ g^\theta(\mathbf{x}) + \frac{D}{2}\log 2\pi$$

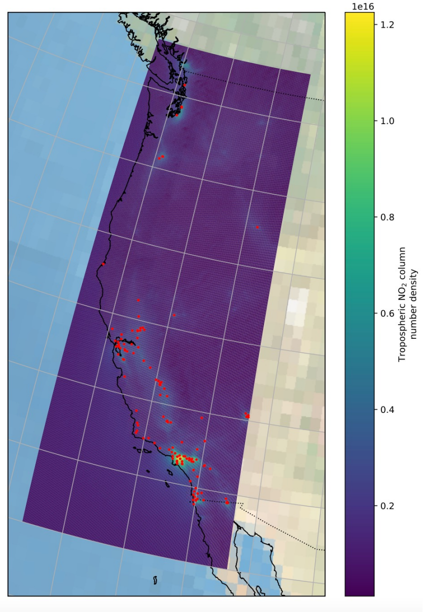
Results



in-distribution

US locations with high uncertainty

out-of-distribution data: CA



REGION	TIME-SPAN	MODEL NAME	UNC.-FILTER	OBSERVATIONS	R2-SCORE	MAE	MSE
EUROPE	2018-2020	SATELLITE	×	3087	0.65	5.18	48.01
EUROPE	2018-2020	MC DROPOUT	×	3087	0.60	5.52	50.85
EUROPE	2018-2020	MC DROPOUT	✓	3061	**0.66**	**4.99**	**45.25**
US	2018-2020	SATELLITE	×	91	0.22	7.87	95.66
US	2018-2020	MC DROPOUT	×	91	-2.44	11.39	422.29
US	2018-2020	MC DROPOUT	✓	86	**0.28**	**7.86**	**89.92**
US	QUARTERLY	SATELLITE	×	273	0.37	8.44	104.77
US	QUARTERLY	MC DROPOUT	×	273	-1.67	11.85	450.22
US	QUARTERLY	MC DROPOUT	✓	258	**0.42**	**8.26**	**98.85**
US	MONTHLY	SATELLITE	×	637	0.46	8.26	105.25
US	MONTHLY	MC DROPOUT	×	637	-1.23	11.73	434.25
US	MONTHLY	MC DROPOUT	✓	602	**0.48**	**8.23**	**102.38**

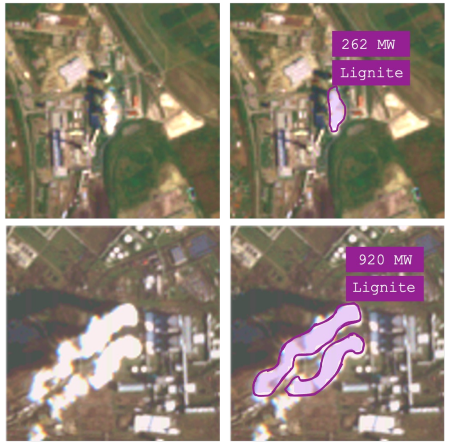
Multitask Learning Power Plant Greenhouse Gas Emissions Estimation

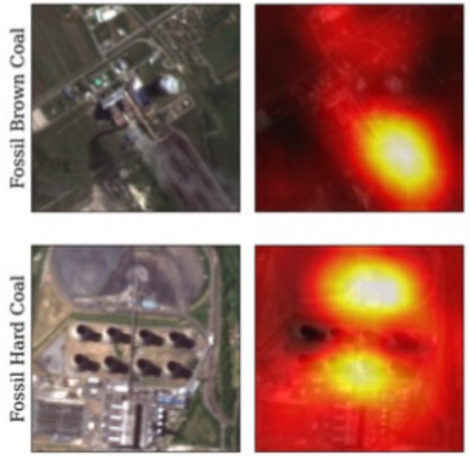
J Hanna, M Mommert, L Scheibenreif, D Borth
Multitask Learning for Estimating Power Plant Greenhouse Gas Emissions from Satellite Imagery,
NeurIPS Workshop on Tackling Climate Change with Machine Learning, 2021

Multitask Learning Power Plant Greenhouse Gas Emissions Estimation

- **Idea:**

  Estimation of power generation (and CO_2) as prediction of:
  - rate of power generation,
  - the type of fired fuel
  - plume footprint





- **Data**
  - esa Sentinel-2
  - entsoe Power Plant Metadata

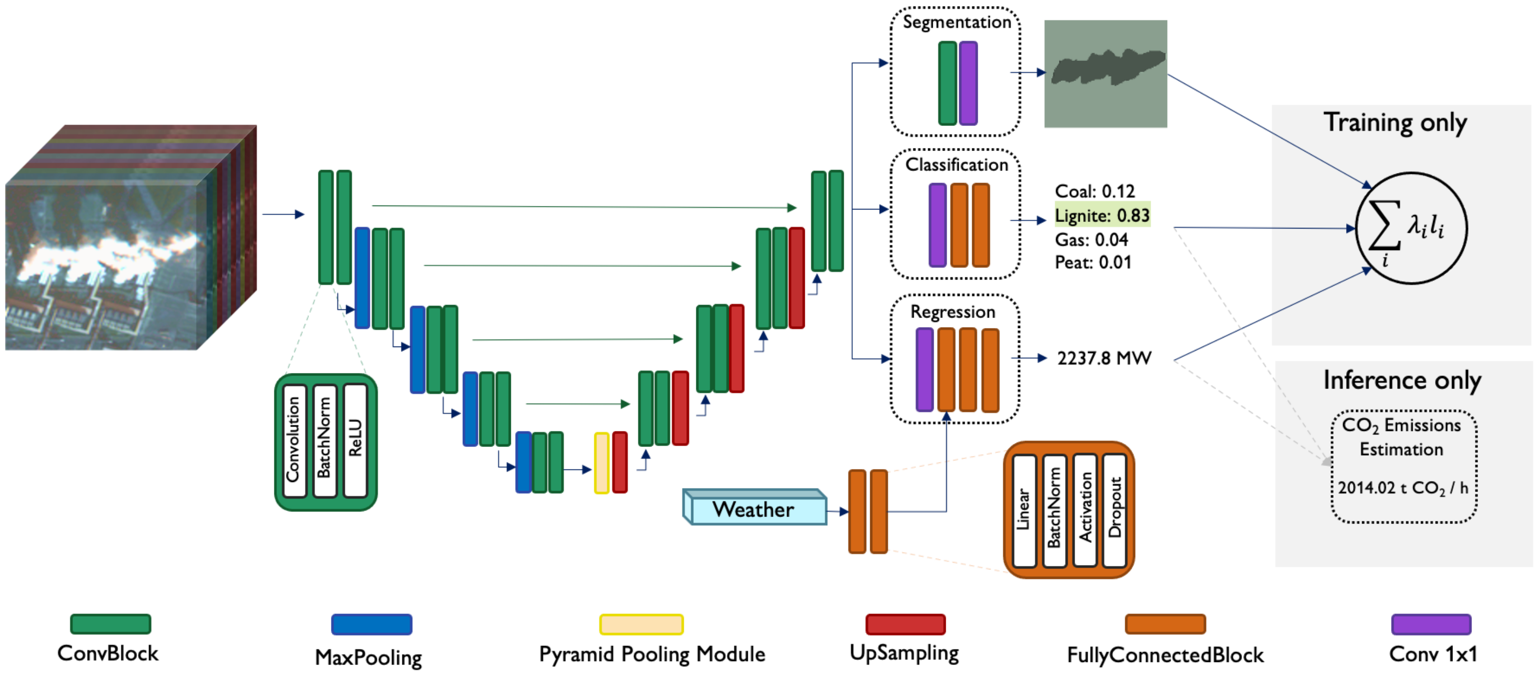
    (type of fuel, hourly power generation rate, max installed capacity, …)
  - ECMWF Environmental Variables

    (temperature at surface, relative humidity, wind norm and direction)

J Hanna, M Mommert, L Scheibenreif, D Borth
Multitask Learning for Estimating Power Plant Greenhouse Gas Emissions from Satellite Imagery,
NeurIPS Workshop on Tackling Climate Change with Machine Learning, 2021

Approach

Multi-task Learning



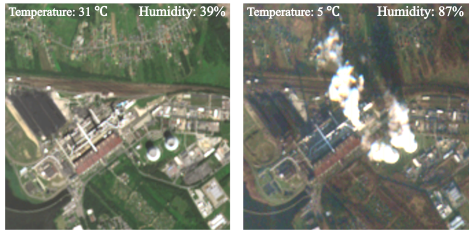
$$
\begin{aligned}
\mathcal{L}_{total} &= \sum_{t=1}^{T} \lambda_t l_t + \mu \|W\|_2 \\
&= \lambda_1 l_{seg} + \lambda_2 l_{reg} + \lambda_3 l_{cls} + \mu \|W\|_2
\end{aligned}
$$

$$
\begin{aligned}
\bar{\kappa}_t^{(\tau)} &= \alpha \kappa_t^{(\tau)} + (1-\alpha) \bar{\kappa}_t^{(\tau-1)} \\
\lambda_t^{(\tau)} &= \mathrm{FL}(\bar{\kappa}_t^{(\tau)}, \gamma_t) \\
&= -(1-\bar{\kappa}_t^{(\tau)})^{\gamma_t} log(\bar{\kappa}_t^{(\tau)})
\end{aligned}
$$

$$
\begin{aligned}
\kappa_{\mathrm{cls}} &= \frac{1}{N} \sum_{i}^{N} \mathbb{1}_{\{y_i = \hat{y}_i\}} \\
\kappa_{\mathrm{seg}} &= \frac{1}{N} \sum_{i}^{N} \mathbb{1}_{\{\mathrm{IoU}(y_i, \hat{y}_i) \geq T\}} \\
\kappa_{\mathrm{reg}} &= \frac{1}{N} \sum_{i}^{N} \mathbb{1}_{\{\mathrm{MAE}(y_i, \hat{y}_i) \leq T\}}
\end{aligned}
$$

Setup:

- shared backbone
- multiple heads
- dynamic task weighting



(a) Low Humidity, High Temperature (b) High Humidity, Low Temperature

Results

Single-task vs. Multi-task for RGB & Multispectral Setups

		Task Weights (λ_i) Segmentation	Regression	Classification	Segmentation IoU (%)	MAE	Regression R^2 (%)	MAPE (%)	Classification Accuracy (%)
RGB	Single	1	0	0	55 ± 2	-	-	-	-
		0	1	0	-	218 ± 21	55 ± 5	60 ± 2	-
		0	0	1	-	-	-	-	87 ± 1
	Multi	0.33	0.33	0.33	57 ± 1	232 ± 17	48 ± 2	61 ± 3	88 ± 1
		0.15	0.7	0.15	53 ± 1	202 ± 6	62 ± 5	53 ± 2	89 ± 1
		Dynamic Weighting			57 ± 1	178 ± 5	70 ± 4	50 ± 5	88 ± 1

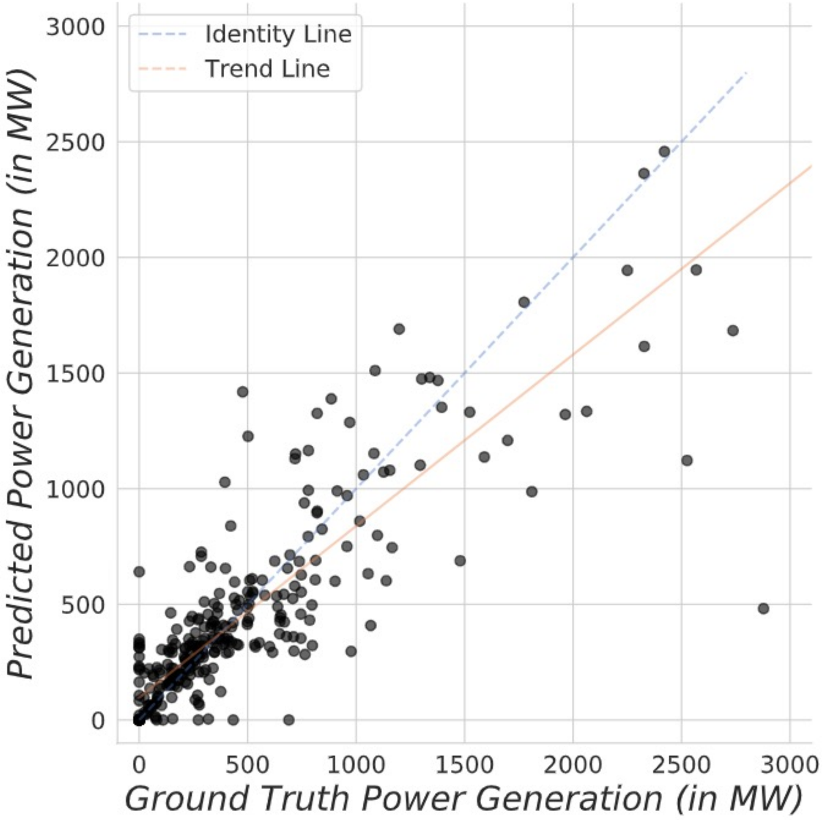
Results

Single-task vs. Multi-task for RGB & Multispectral Setups

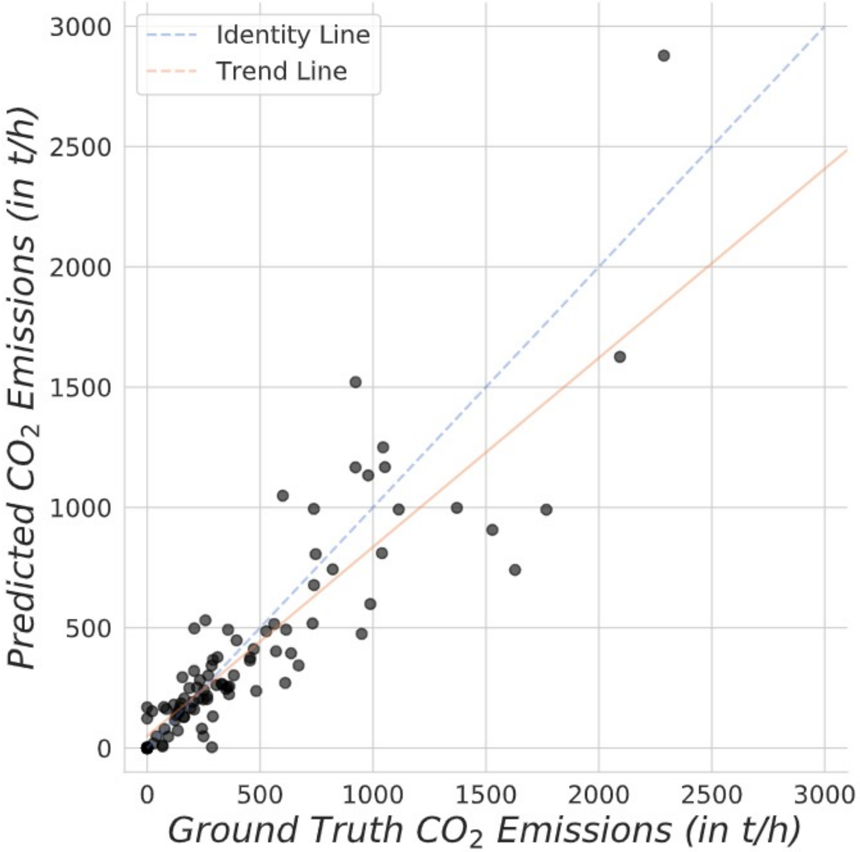
		Task Weights (λ_i) Segmentation	Regression	Classification	Segmentation IoU (%)	Regression MAE	R^2 (%)	MAPE (%)	Classification Accuracy (%)
RGB	Single	1	0	0	55 ± 2	-	-	-	-
		0	1	0	-	218 ± 21	55 ± 5	60 ± 2	-
		0	0	1	-	-	-	-	87 ± 1
	Multi	0.33	0.33	0.33	57 ± 1	232 ± 17	48 ± 2	61 ± 3	88 ± 1
		0.15	0.7	0.15	53 ± 1	202 ± 6	62 ± 5	53 ± 2	89 ± 1
		Dynamic Weighting			57 ± 1	178 ± 5	70 ± 4	50 ± 5	88 ± 1
Multispectral	Single	1	0	0	59 ± 1	-	-	-	-
		0	1	0	-	202 ± 20	65 ± 5	60 ± 1	-
		0	0	1	-	-	-	-	90 ± 1
	Multi	0.33	0.33	0.33	61 ± 1	194 ± 9	63 ± 5	57 ± 5	92 ± 1
		0.15	0.7	0.15	62 ± 1	181 ± 6	69 ± 3	56 ± 1	**94 ± 1**
		Dynamic Weighting			**64 ± 0**	**157 ± 4**	**78 ± 3**	**43 ± 5**	93 ± 1

Results

Power Generation Estimation



CO2 Estimation



Universität St.Gallen

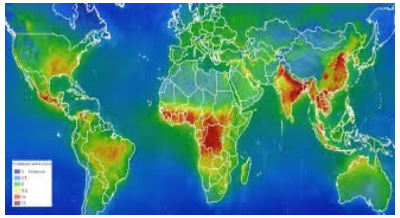
Self-supervised Learning

L Scheibenreif, M Mommert, D Borth
Contrastive Self-supervised Data Fusion for Satellite Imagery
Int. Society for Photogrammetry and Remote Sensing (ISPRS), 2022

L Scheibenreif, J Hanna, M Mommert, D Borth
Self-supervised Vision Transformer for Land-cover Segmentation and Classification
CVPR Earth Vision Workshop, 2022 - [Best Student Paper Award]

Overview

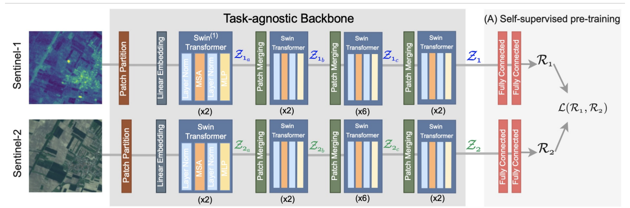
Shared-Backbones/Heads



Approach:
- Muilti-modal Fusion
- Multi-task Learning
- Auxiliary Tasks

Application
- NO2 estimation
- Power Production
- CO2 estimation

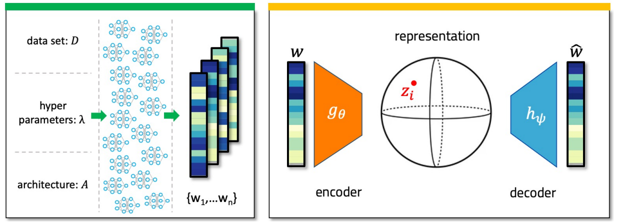
Self-supervised Learning



Approach:
- Contrastive Learning
- Augmentation free
- CNNs & Transformer

Application
- Land-use Classification
- Single-class / Multi-class
- Segmentation

Hyper-Representations



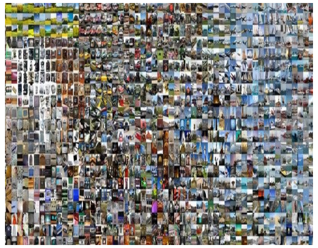
Approach:
- Contrastive Learning
- Model Zoos
- CNNs

Application
- Model analysis
- Sample unseen models
- Sparsificaiton

Universität St.Gallen

Self-supervised Learning

Setup



Large amounts of unlabeled data

→ Self-supervised Learning

Learn a good feature representation from self-supervised pretext tasks, i.e., labels are generated automatically

→ f_θ

feature encoder

pretext task

→ Supervised Learning (fine-tuning)

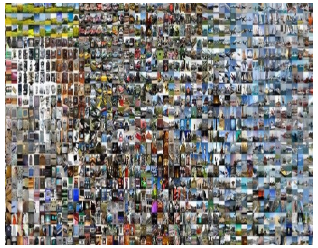
Attach a shallow network on the feature encoder; train the shallow network on the downstream task with small amount of labeled data

→ evaluate on the downstream task e.g. classification, regression, detection, …

downstream task

Self-supervised Learning

Setup



Large amounts of unlabeled data

→ Self-supervised Learning

Learn a good feature representation from self-supervised pretext tasks, i.e., labels are generated automatically

→ f_θ

feature encoder

→ Supervised Learning (fine-tuning)

Attach a shallow network on the feature encoder; train the shallow network on the downstream task with small amount of labeled data

→ evaluate on the downstream task e.g. classification, regression, detection, …

pretext task

downstream task

Goal

SSL aims to learn rich task-agnostic representations from raw unlabeled data suitable for many different downstream tasks

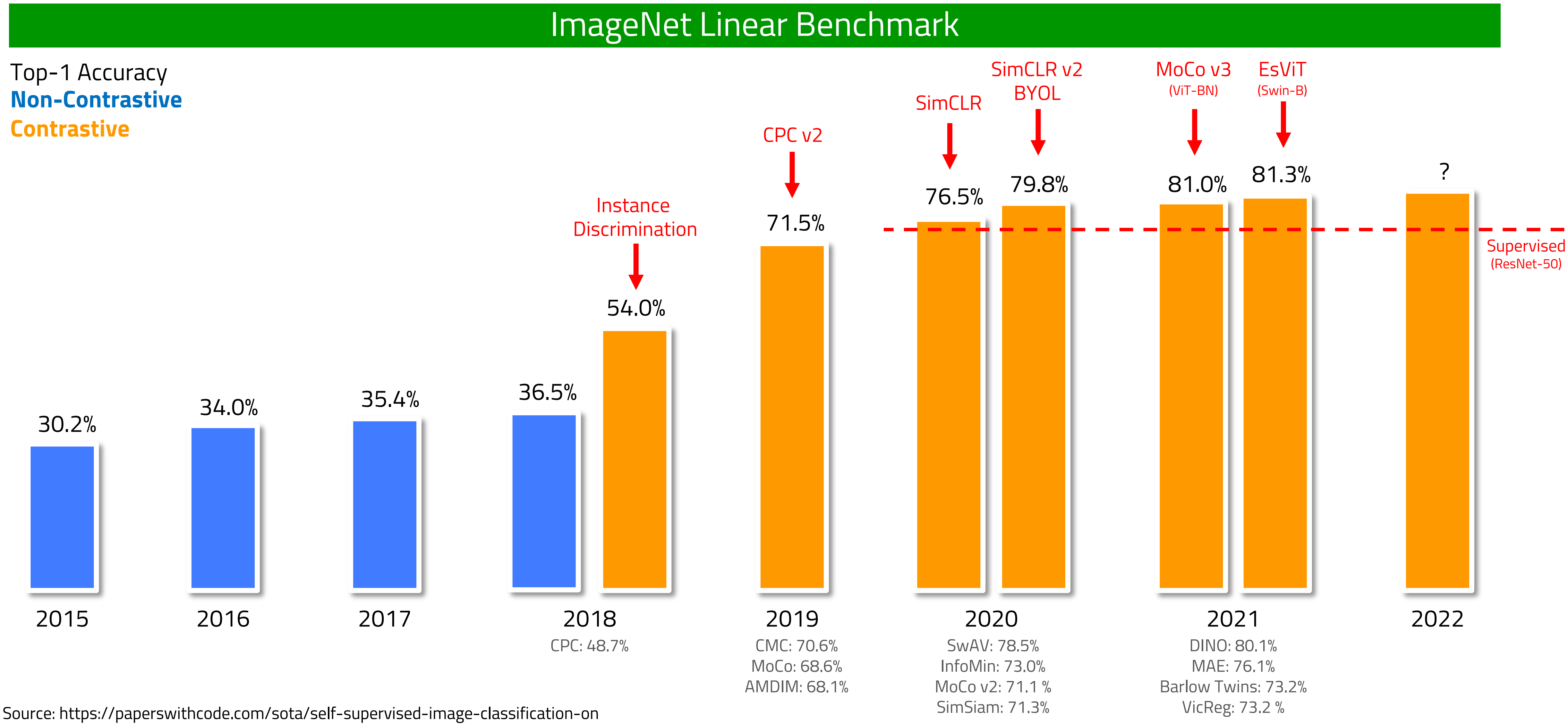
- We want to be able to train equally good models without too many labels
- We potentially might be able to generalize better because we have to learn more about the world

Evaluation

Evaluation of SSL (usually) focuses on the downstream task and not so much on the self-supervised learning task

- We don't care so much about the pretext task performance
- We want rich (disentangled) representations -> evaluation on downstream task fine-tuning a linear model with labels
- We want to know about generalizability of our rep-representation -> fine-tuning with non-linear model with labels

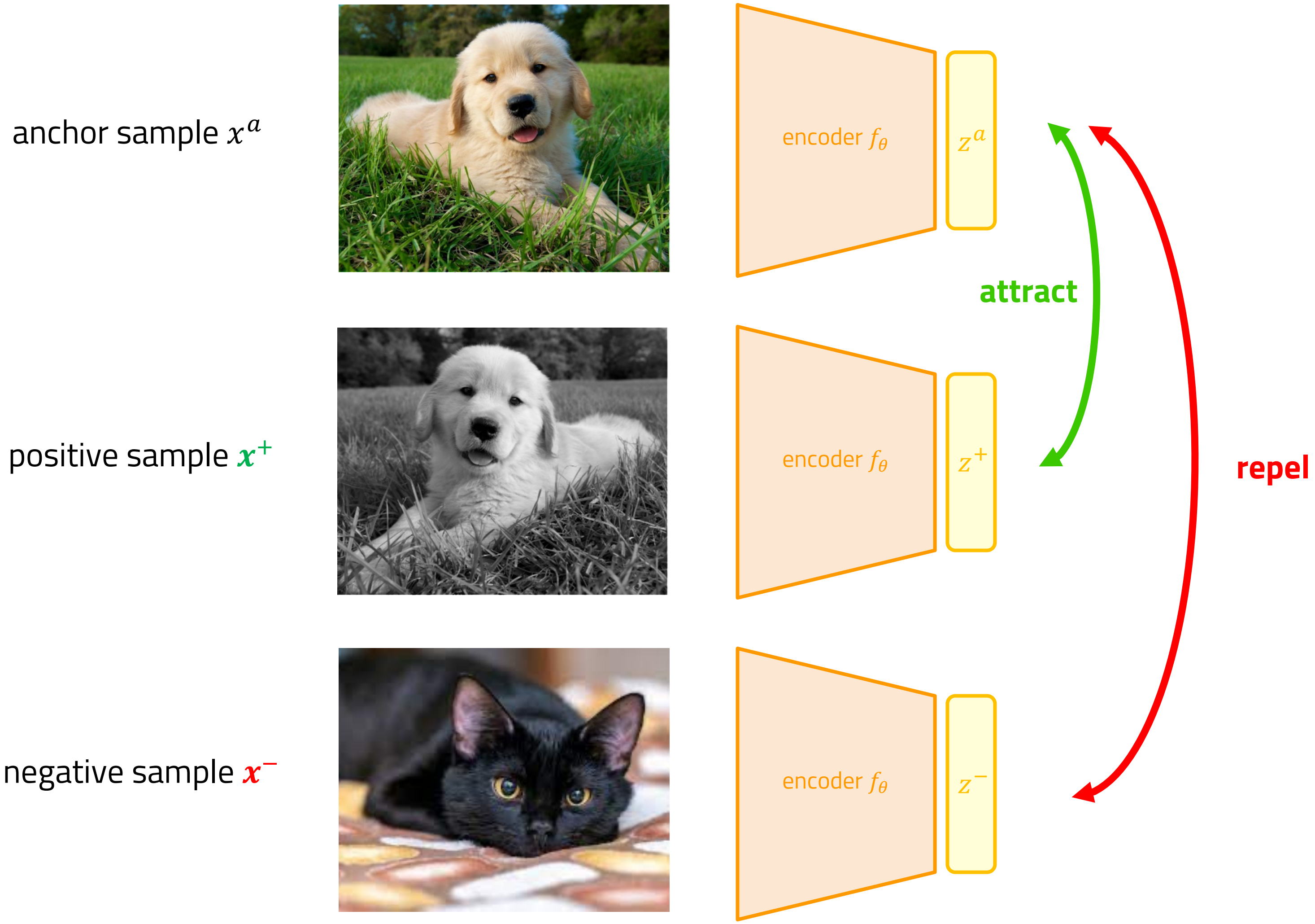
“Evolution” of SSL Approaches



Source: https://paperswithcode.com/sota/self-supervised-image-classification-on

Universität St.Gallen

Contrastive Learning

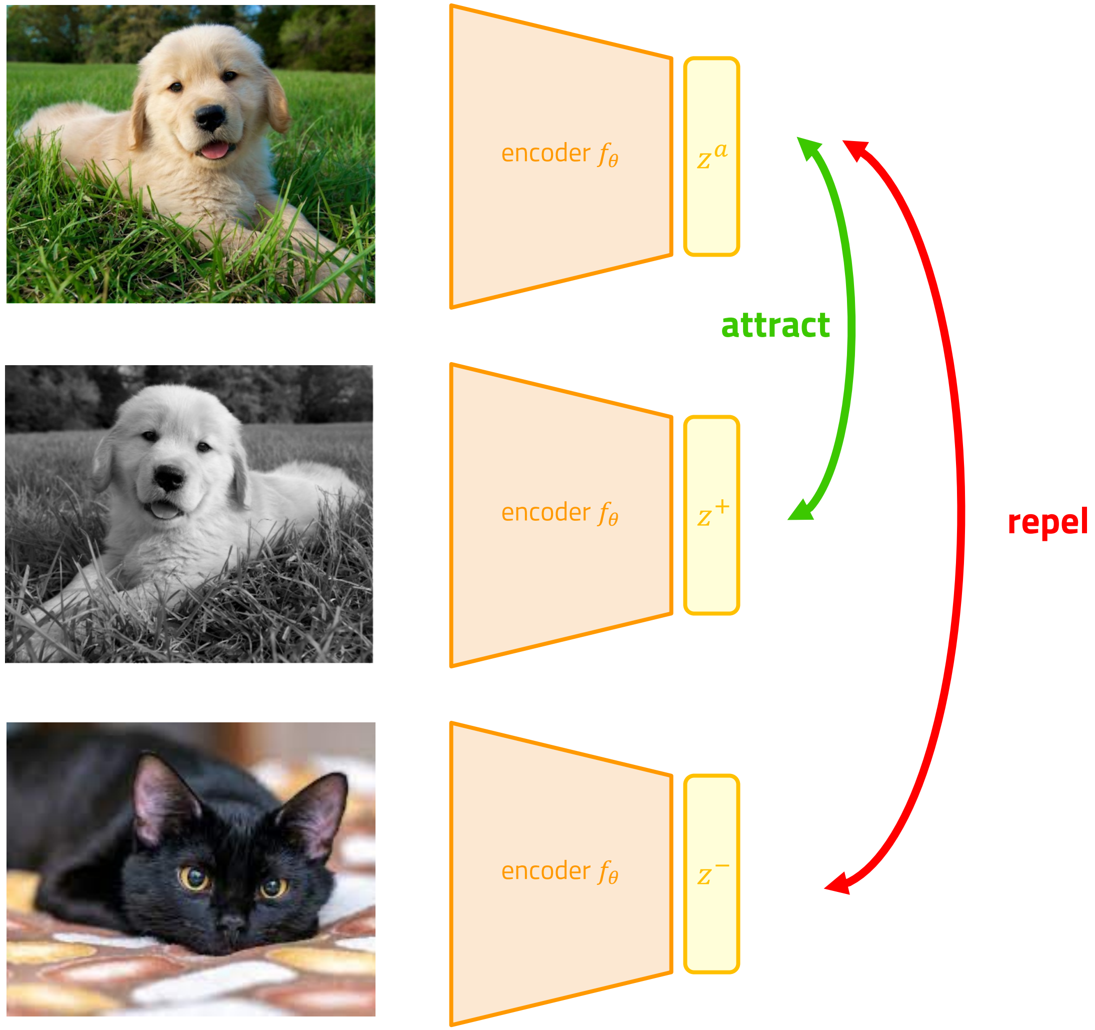


Contrastive Learning

anchor sample x^a

positive sample x^+

negative sample x^-



Setup

Contrast is being defined in latent space i.e., the **embedding** vector of the image after a forward-pass through an (the same) **encoder** f_θ.

Since we have now vectors representing sample we have to quantify "attract" and "repel" and include this into a loss.

Design Decisions:

1. Select encoder
2. Select similarity / distance (metric)
3. Define a proper loss function

Negatives:
"Hard negatives" are important to learn contrast, but might be drawn from the same class

Learning Objective

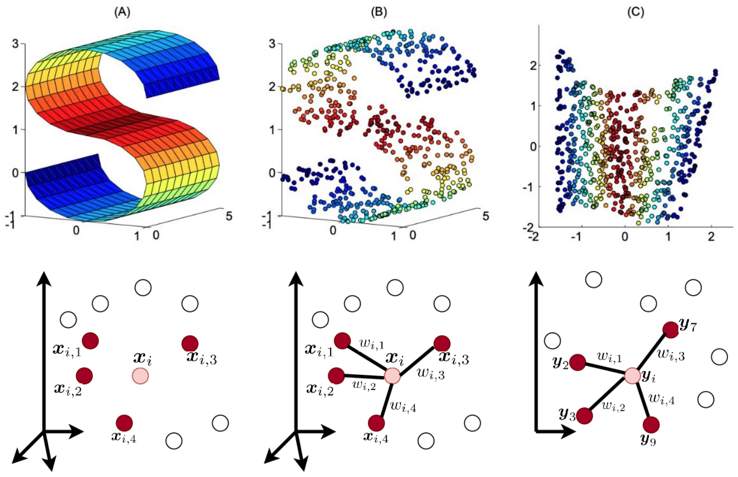
Historical

Precursor of this type of learning objective comes from two disciplines:

- **Multiple Instance Learning**
- **Metric Learning**

with ideas inspired by:

- Multidimensional scaling (MDS) [MDS; Cox et al. 1994]
- Locally linear embedding (LLE) [LLE; Roweis et al. 2000]



InfoNCE / NT-Xent loss

Given an anchor, one positive and N-1 negative samples $\{\mathbf{x}^a, \mathbf{x}^+, \mathbf{x}_1^-, \ldots, \mathbf{x}_{N-1}^-\}$:

$$\mathcal{L}^{\text{NT-Xent}} = -\log \frac{\overbrace{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\boxed{\tau})}^{\text{attract}}}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} \underbrace{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\boxed{\tau})}_{\text{repel}}}$$

temperature term

Loss functions

- Contrastive loss
  [Chopra et al. 2005]
- Triplet loss
  [Schroff et al. 2015; FaceNet]
- Lifted structured loss
  [Song et al. 2015]
- N-pair loss
  [Sohn 2016]
- InfoNCE loss
  [van den Oord, et al. 2018]
- NT-Xent loss
  [Chen et al., 2020]

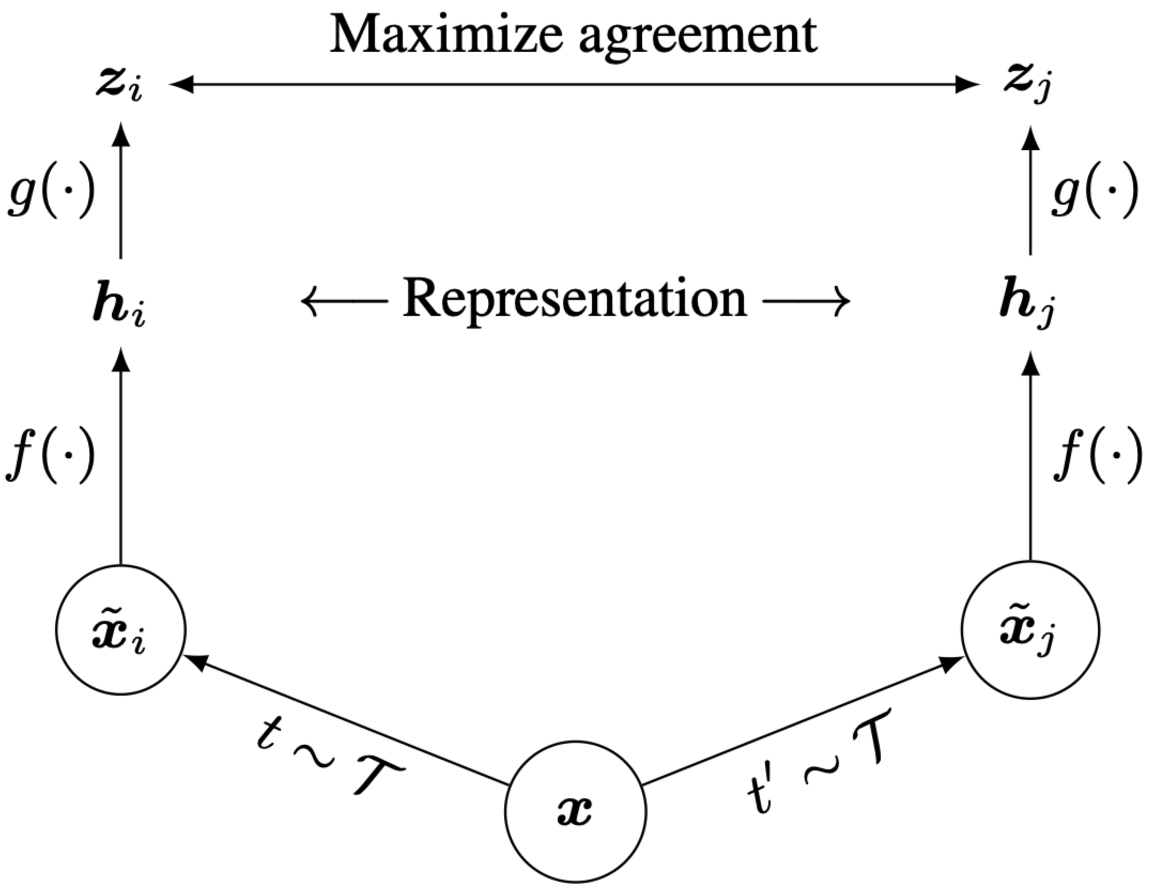
Loss calculation is done within the mini batch
i.e., batch size is a limiting factor for sample size
as it related directly to the GPU or TPU memory!

Contrastive Self-supervised Data Fusion for Satellite Imagery

L Scheibenreif, M Mommert, D Borth
Contrastive Self-supervised Data Fusion for Satellite Imagery
Int. Society for Photogrammetry and Remote Sensing (ISPRS), 2022

Contrastive Self-supervised Data Fusion for Satellite Imagery

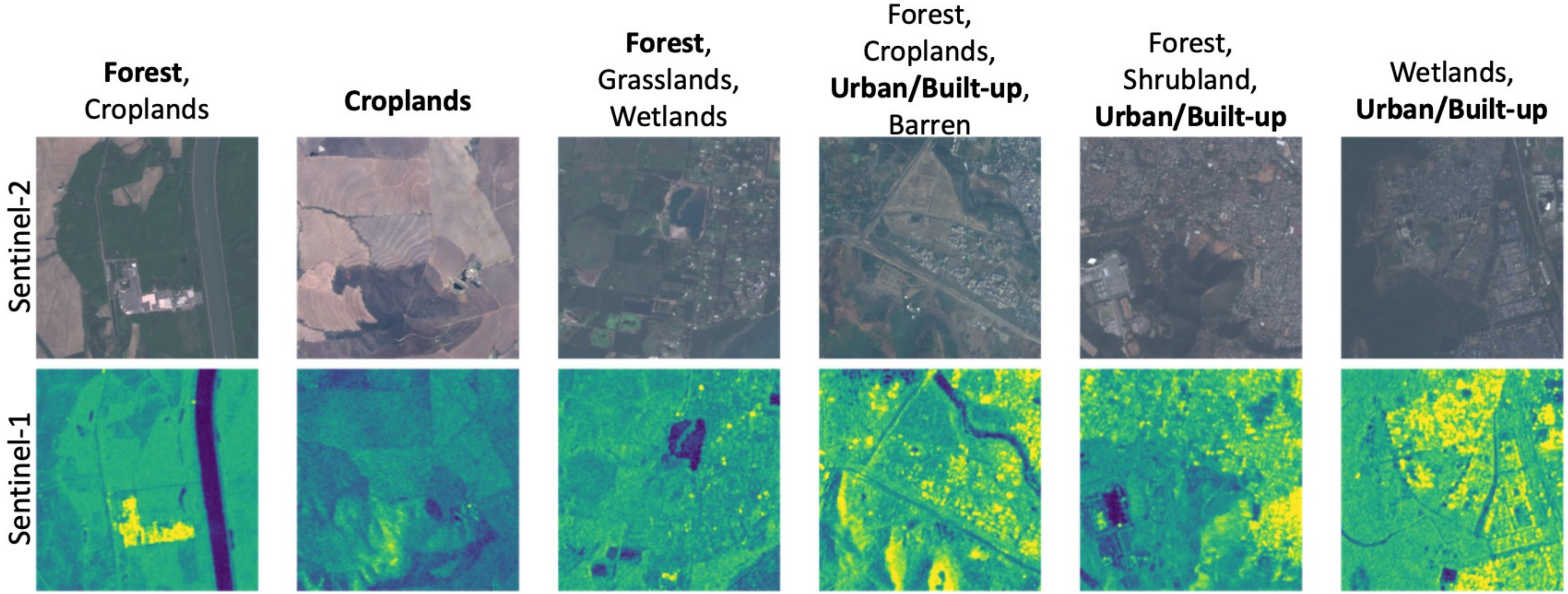
- Contrastive SSL yields great performance on natural images (e.g., SimCLR)
- Based on multiple views of same instance
- In natural images, multiple views are generated with **random augmentations**
- In remote sensing, unlabeled data is abundant, but less labeled data
- What could multiple views be in remote sensing and earth observation?



L Scheibenreif, M Mommert, D Borth
Contrastive Self-supervised Data Fusion for Satellite Imagery
Int. Society for Photogrammetry and Remote Sensing (ISPRS), 2022

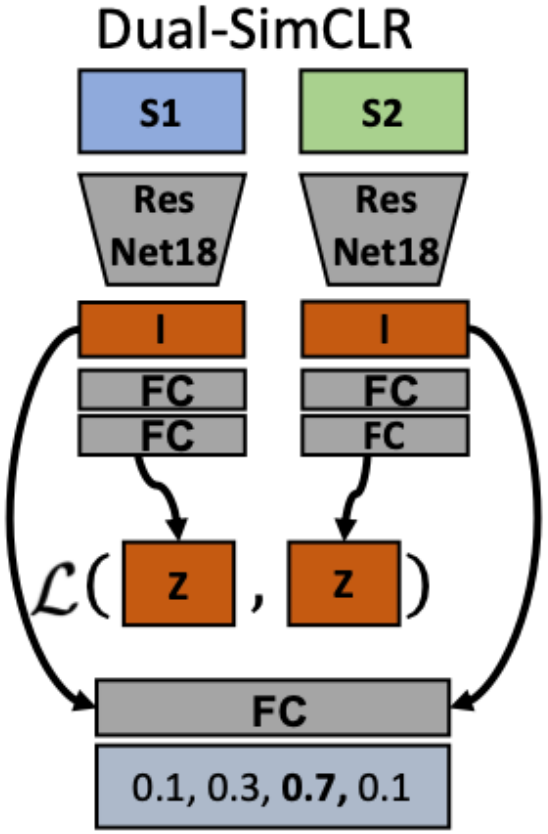
Contrastive SSL in Satellite Imagery

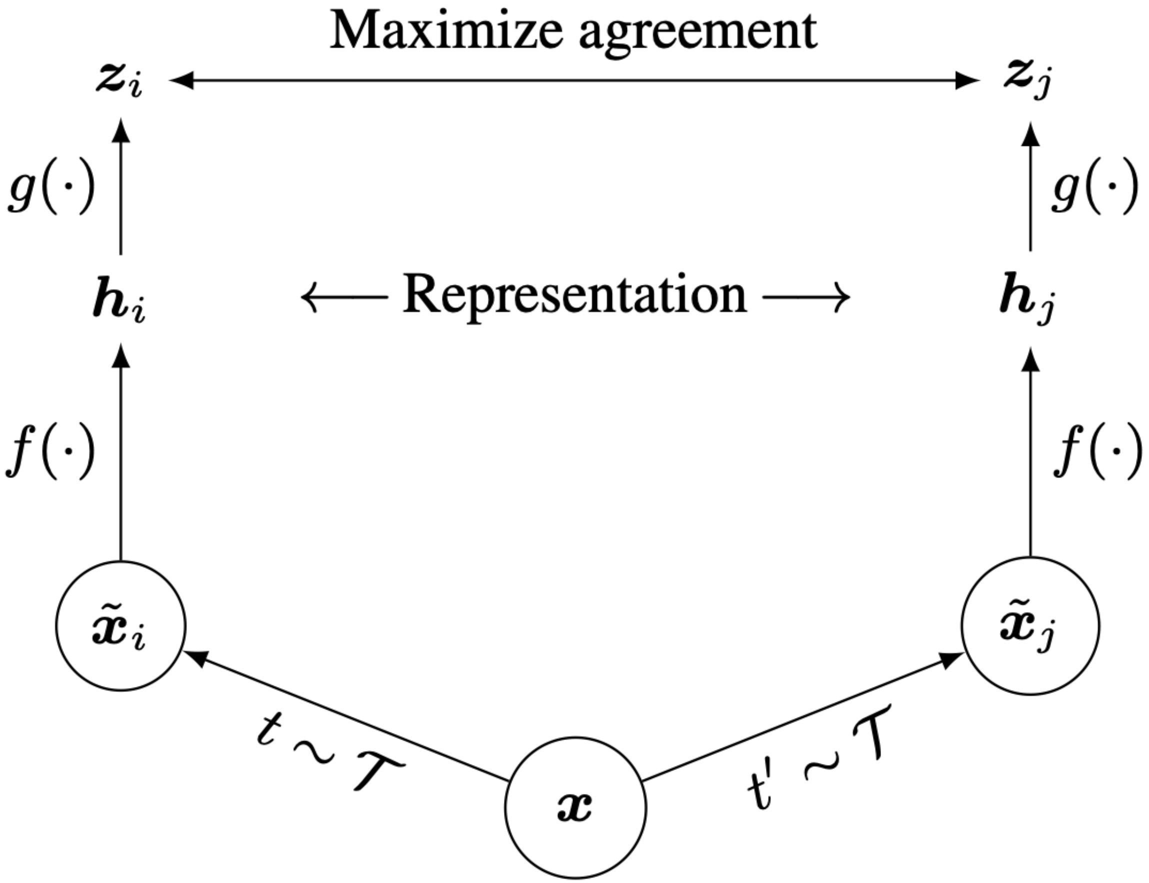
In satellite imagery, there are multiple views of the same location



Schmitt, M., Hughes, L. H., Qiu, C., & Zhu, X. X.
SEN12MS--A Curated Dataset of Georeferenced Multi-Spectral Sentinel-1/2 Imagery for Deep Learning and Data Fusion.
arXiv preprint arXiv:1906.07789, 2019

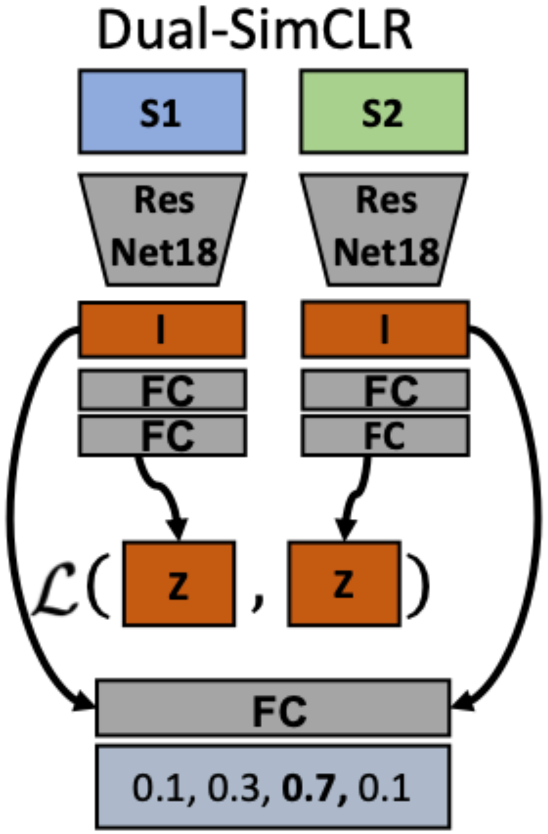
Approach: “Dual-SimCLR”

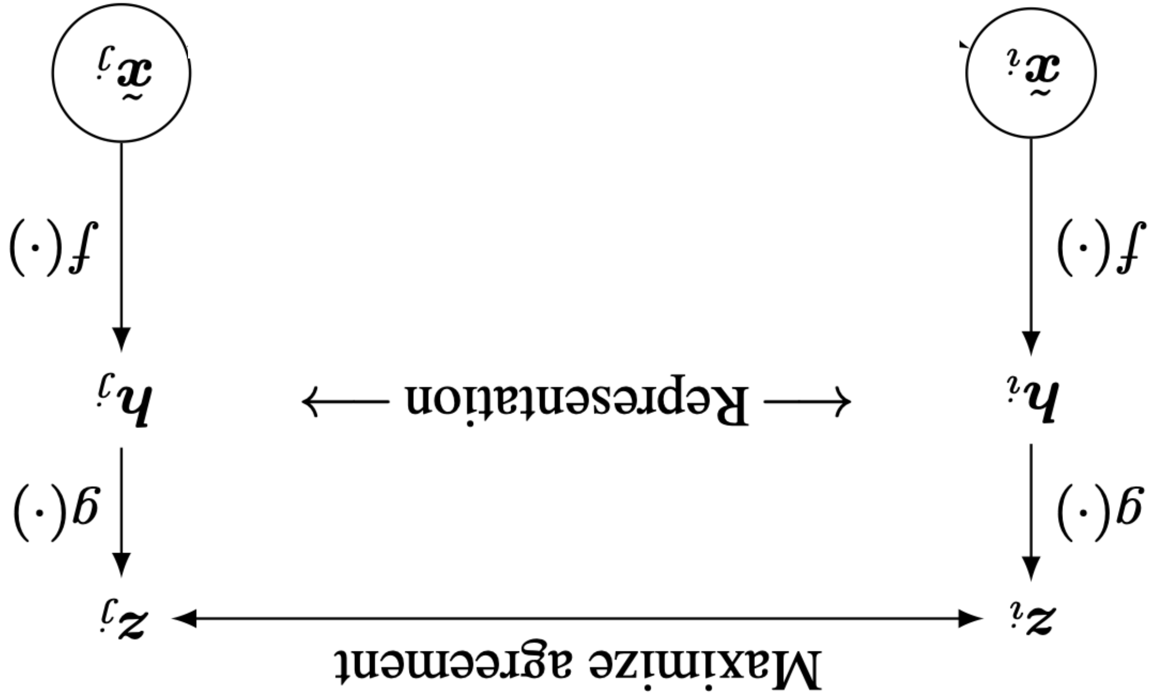




Universität St.Gallen

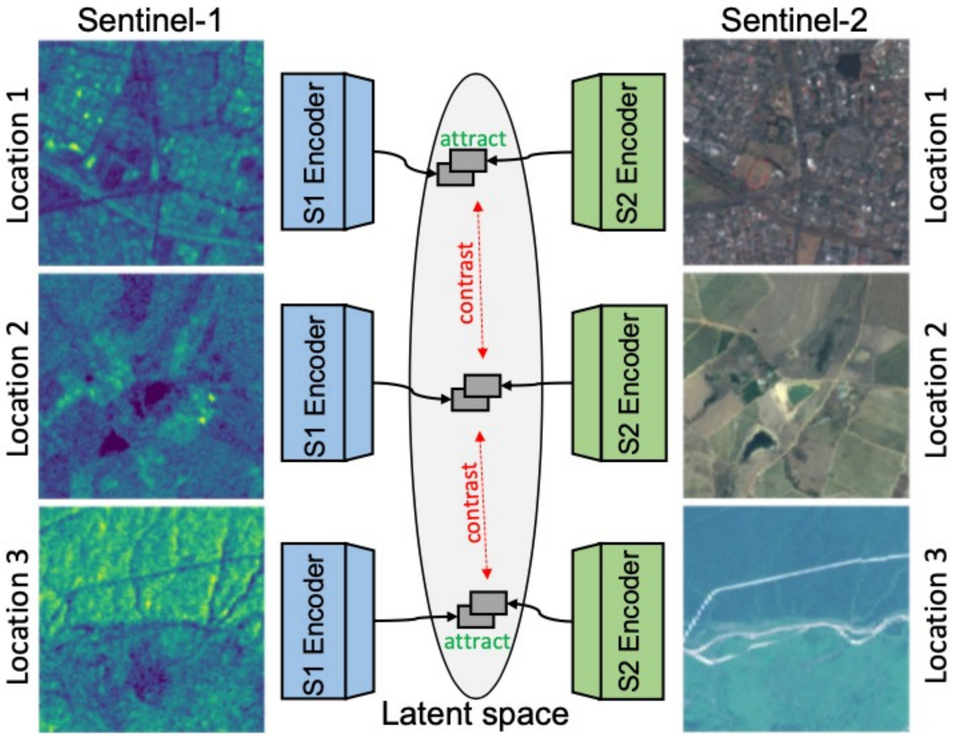
Approach: “Dual-SimCLR”



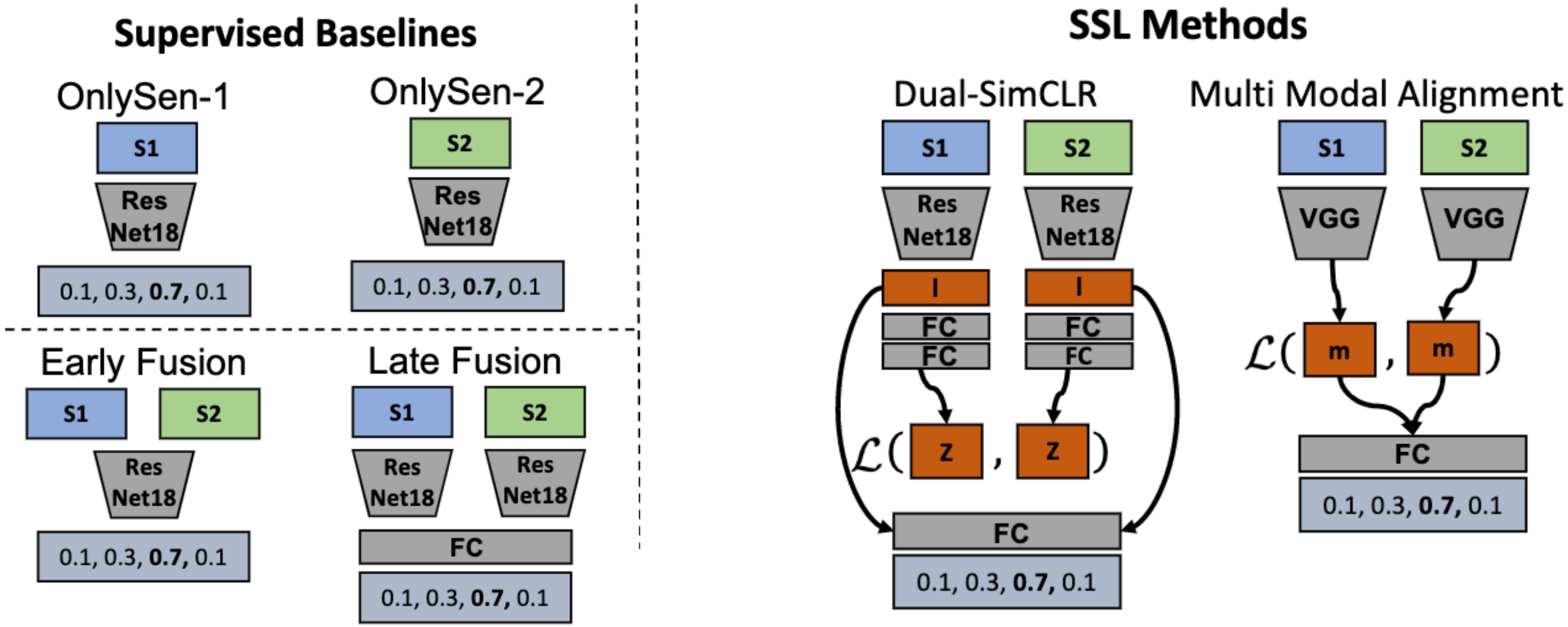


Approach: “Dual-SimCLR”

- SSL contrast on pairs of Sentinel-1/2 images for the same location
  - – SEN12MS dataset
- Supervised training on different downstream tasks:
  - – Single-label classification
  - – Multi-label classification
  - – DFC2020 dataset
  - – EuroSAT



Experimental Setup



Universität St.Gallen

Results

Single-label classification

Accuracy (%)	Forest	Shrubland	Grassl.	Wetl.	Cropl.	Urban	Barren	Water	**Average**	**OA**
OnlySen-1	80 ± 15	57 ± 2	18 ± 17	0 ± 0	75 ± 10	67 ± 9	58 ± 2	97 ± 2	57 ± 3	62 ± 1
OnlySen-2	43 ± 26	78 ± 12	45 ± 29	11 ± 6	59 ± 9	62 ± 5	61 ± 18	96 ± 6	57 ± 6	62 ± 5
EarlyFusion	60 ± 12	66 ± 37	62 ± 8	1 ± 1	66 ± 10	73 ± 6	66 ± 18	99 ± 0	62 ± 4	66 ± 2
LateFusion	62 ± 23	76 ± 14	51 ± 18	1 ± 2	64 ± 11	71 ± 5	75 ± 9	100 ± 1	62 ± 4	65 ± 3

fine-tuning to DFC2020 dataset

Universität St.Gallen

Results

Single-label classification

Accuracy (%)	Forest	Shrubland	Grassl.	Wetl.	Cropl.	Urban	Barren	Water	Average	OA
OnlySen-1	80 ± 15	57 ± 2	18 ± 17	0 ± 0	75 ± 10	67 ± 9	58 ± 2	97 ± 2	57 ± 3	62 ± 1
OnlySen-2	43 ± 26	78 ± 12	45 ± 29	11 ± 6	59 ± 9	62 ± 5	61 ± 18	96 ± 6	57 ± 6	62 ± 5
EarlyFusion	60 ± 12	66 ± 37	62 ± 8	1 ± 1	66 ± 10	73 ± 6	66 ± 18	99 ± 0	62 ± 4	66 ± 2
LateFusion	62 ± 23	76 ± 14	51 ± 18	1 ± 2	64 ± 11	71 ± 5	75 ± 9	100 ± 1	62 ± 4	65 ± 3
SimCLR (RGB)	11 ± 12	69 ± 13	45 ± 14	3 ± 3	66 ± 22	26 ± 23	77 ± 14	99 ± 1	49 ± 3	58 ± 4
D-SimCLR	78 ± 11	84 ± 6	62 ± 10	10 ± 6	63 ± 3	84 ± 4	82 ± 7	99 ± 0	$\mathbf{70 \pm 2}$	$\mathbf{70 \pm 1}$
MMA	68 ± 17	89 ± 5	53 ± 13	8 ± 9	71 ± 7	80 ± 6	81 ± 7	100 ± 0	69 ± 2	69 ± 1

fine-tuning to DFC2020 dataset

Results

Single-label classification

Accuracy (%)	Forest	Shrubland	Grassl.	Wetl.	Cropl.	Urban	Barren	Water	**Average**	**OA**
OnlySen-1	80 ± 15	57 ± 2	18 ± 17	0 ± 0	75 ± 10	67 ± 9	58 ± 2	97 ± 2	57 ± 3	62 ± 1
OnlySen-2	43 ± 26	78 ± 12	45 ± 29	11 ± 6	59 ± 9	62 ± 5	61 ± 18	96 ± 6	57 ± 6	62 ± 5
EarlyFusion	60 ± 12	66 ± 37	62 ± 8	1 ± 1	66 ± 10	73 ± 6	66 ± 18	99 ± 0	62 ± 4	66 ± 2
LateFusion	62 ± 23	76 ± 14	51 ± 18	1 ± 2	64 ± 11	71 ± 5	75 ± 9	100 ± 1	62 ± 4	65 ± 3
SimCLR (RGB)	11 ± 12	69 ± 13	45 ± 14	3 ± 3	66 ± 22	26 ± 23	77 ± 14	99 ± 1	49 ± 3	58 ± 4
D-SimCLR	78 ± 11	84 ± 6	62 ± 10	10 ± 6	63 ± 3	84 ± 4	82 ± 7	99 ± 0	**70 ± 2**	**70 ± 1**
MMA	68 ± 17	89 ± 5	53 ± 13	8 ± 9	71 ± 7	80 ± 6	81 ± 7	100 ± 0	69 ± 2	69 ± 1

Multi-label classification

F1 Score (%)	Forest	Shrubland	Grassl.	Wetl.	Cropl.	Urban	Barren	Water	**Average**	**O-F1**
OnlySen-1	69 ± 2	46 ± 6	29 ± 5	8 ± 8	68 ± 7	81 ± 3	60 ± 8	96 ± 1	57 ± 2	62 ± 2
OnlySen-2	37 ± 20	51 ± 14	43 ± 20	23 ± 18	76 ± 2	79 ± 6	63 ± 10	94 ± 2	58 ± 3	63 ± 2
EarlyFusion	48 ± 10	53 ± 7	45 ± 13	13 ± 11	69 ± 5	84 ± 4	71 ± 4	94 ± 1	60 ± 3	62 ± 3
LateFusion	56 ± 6	45 ± 11	33 ± 9	18 ± 24	64 ± 3	69 ± 16	53 ± 15	96 ± 1	54 ± 7	61 ± 5

fine-tuning to DFC2020 dataset

Universität St.Gallen

Results

Single-label classification

Accuracy (%)	Forest	Shrubland	Grassl.	Wetl.	Cropl.	Urban	Barren	Water	Average	OA
OnlySen-1	80 ± 15	57 ± 2	18 ± 17	0 ± 0	75 ± 10	67 ± 9	58 ± 2	97 ± 2	57 ± 3	62 ± 1
OnlySen-2	43 ± 26	78 ± 12	45 ± 29	11 ± 6	59 ± 9	62 ± 5	61 ± 18	96 ± 6	57 ± 6	62 ± 5
EarlyFusion	60 ± 12	66 ± 37	62 ± 8	1 ± 1	66 ± 10	73 ± 6	66 ± 18	99 ± 0	62 ± 4	66 ± 2
LateFusion	62 ± 23	76 ± 14	51 ± 18	1 ± 2	64 ± 11	71 ± 5	75 ± 9	100 ± 1	62 ± 4	65 ± 3
SimCLR (RGB)	11 ± 12	69 ± 13	45 ± 14	3 ± 3	66 ± 22	26 ± 23	77 ± 14	99 ± 1	49 ± 3	58 ± 4
D-SimCLR	78 ± 11	84 ± 6	62 ± 10	10 ± 6	63 ± 3	84 ± 4	82 ± 7	99 ± 0	$\mathbf{70 \pm 2}$	$\mathbf{70 \pm 1}$
MMA	68 ± 17	89 ± 5	53 ± 13	8 ± 9	71 ± 7	80 ± 6	81 ± 7	100 ± 0	69 ± 2	69 ± 1

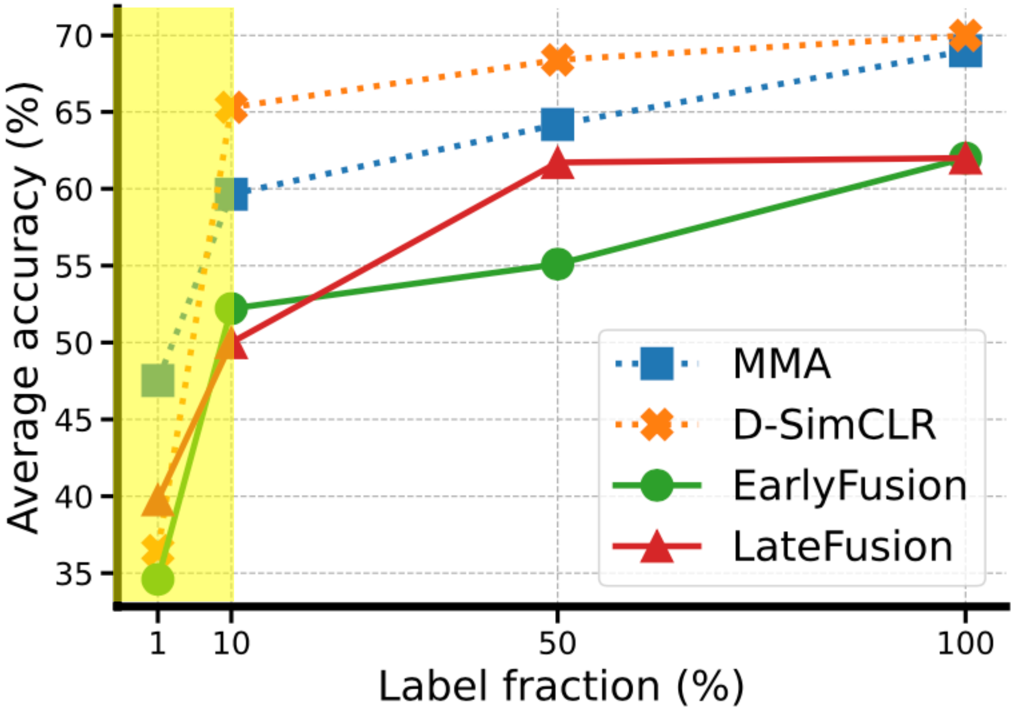
Multi-label classification

F1 Score (%)	Forest	Shrubland	Grassl.	Wetl.	Cropl.	Urban	Barren	Water	Average	O-F1
OnlySen-1	69 ± 2	46 ± 6	29 ± 5	8 ± 8	68 ± 7	81 ± 3	60 ± 8	96 ± 1	57 ± 2	62 ± 2
OnlySen-2	37 ± 20	51 ± 14	43 ± 20	23 ± 18	76 ± 2	79 ± 6	63 ± 10	94 ± 2	58 ± 3	63 ± 2
EarlyFusion	48 ± 10	53 ± 7	45 ± 13	13 ± 11	69 ± 5	84 ± 4	71 ± 4	94 ± 1	60 ± 3	62 ± 3
LateFusion	56 ± 6	45 ± 11	33 ± 9	18 ± 24	64 ± 3	69 ± 16	53 ± 15	96 ± 1	54 ± 7	61 ± 5
SimCLR (RGB)	3 ± 4	49 ± 11	24 ± 16	10 ± 8	63 ± 24	40 ± 36	49 ± 15	73 ± 6	39 ± 10	49 ± 6
D-SimCLR	62 ± 2	61 ± 3	53 ± 7	31 ± 2	72 ± 3	87 ± 0	77 ± 1	96 ± 1	$\mathbf{67 \pm 1}$	$\mathbf{69 \pm 1}$
MMA	58 ± 5	57 ± 5	35 ± 8	10 ± 6	77 ± 3	89 ± 1	73 ± 5	97 ± 0	62 ± 2	66 ± 1

fine-tuning to DFC2020 dataset

Results

Ablation on labeled dataset size



fine-tuning to DFC2020 dataset

Self-supervised Vision Transformer for Land-cover Segmentation and Classification

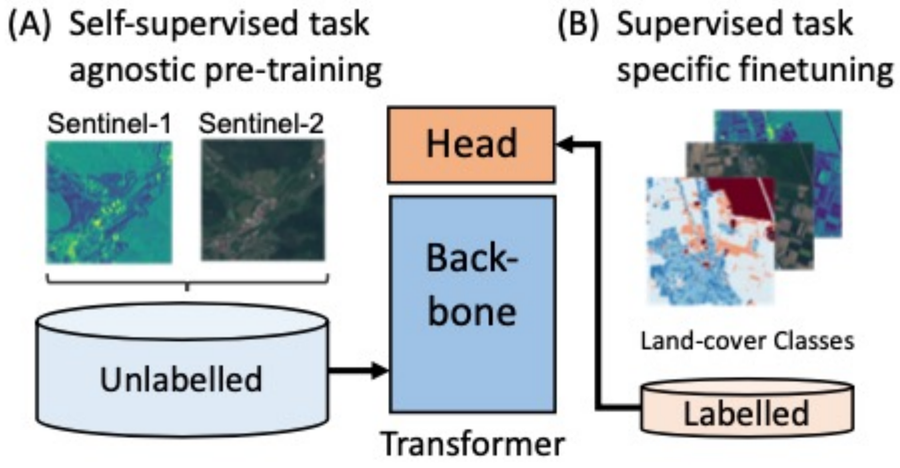
L Scheibenreif, J Hanna, M Mommert, D Borth
Self-supervised Vision Transformer for Land-cover Segmentation and Classification
CVPR Earth Vision Workshop, 2022

Self-supervised Vision Transformer for Land-cover Segmentation and Classification

- Transformer models are state-of-the-art in NLP [Otter 2020]
- show great potential in Computer Vision [Dosovitskiy 2020]
- struggle on small datasets

- Self-supervised learning (SSL) contributes to success of Transformers in NLP

We adapt contrastive SSL to remote sensing data for pre-training of <u>Vision Transformers</u> and extend downstream tasks to <u>segmentation</u>



  - Self-supervised pre-training of large encoders
  - Finetuning of small heads for downstream tasks

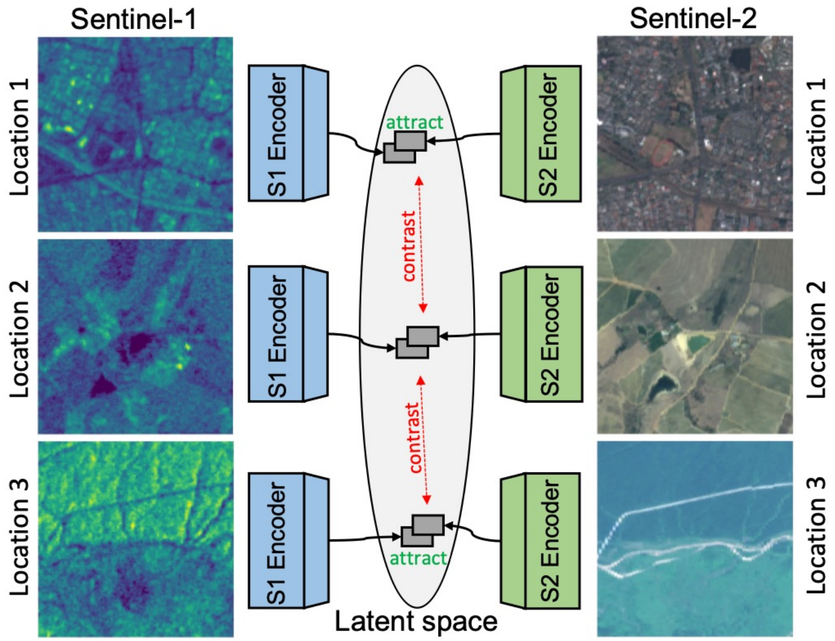
- SSL Related Work:
  Acquire multiple views as co-located measurements
  [Manas 2021], [Saha 2021], [Chen, 2021]

L Scheibenreif, J Hanna, M Mommert, D Borth
Self-supervised Vision Transformer for Land-cover Segmentation and Classification
CVPR Earth Vision Workshop, 2022

Data & Setup

Self-supervised pre-training

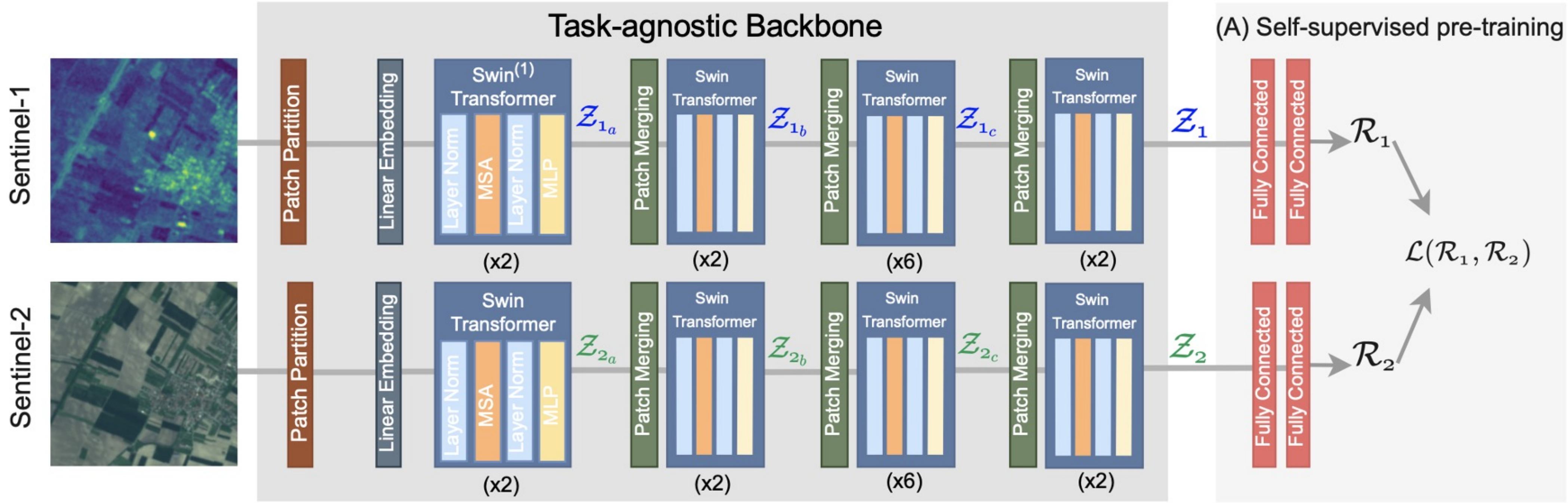
- Co-located Sentinel-1/2 image pairs
- SEN12MS dataset [Schmitt 2019]
- Low-resolution land cover labels are ignored



Land-cover classification downstream tasks

- Dataset from Data Fusion Contest (DFC2020) [Yokoya 2020]
- **Task 1:** Single- and multilabel classification
- **Task 2:** Segmentation

Self-supervised Pre-training

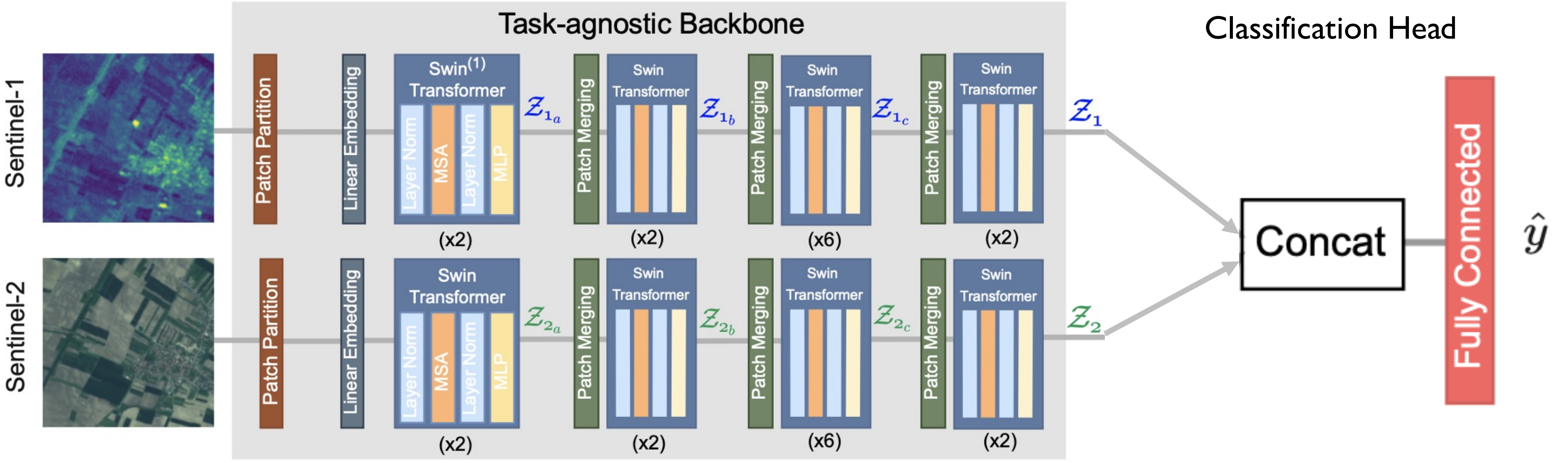


1. Encode Sentinel-1/2 images with distinct encoders
2. Compute contrastive loss on projected representations

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\mathrm{sim}(\mathcal{R}_i, \mathcal{R}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} \exp(\mathrm{sim}(\mathcal{R}_i, \mathcal{R}_k)/\tau)}$$

Swin Transformer: [Liu 2021]

Classification

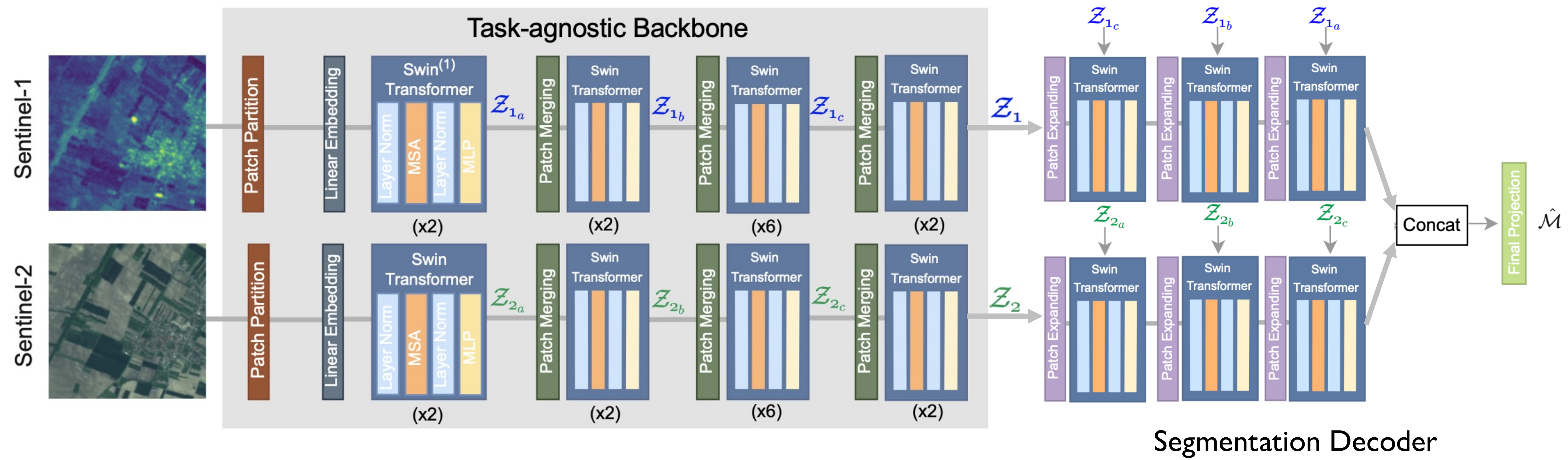


1. Encode Sentinel-1/2 images with distinct encoders
2. Compute contrastive loss on projected representations
3. Replace projection head by downstream task specific head

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\mathrm{sim}(\mathcal{R}_i, \mathcal{R}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(\mathcal{R}_i, \mathcal{R}_k)/\tau)}$$

Swin Transformer: [Liu 2021]

Segmentation

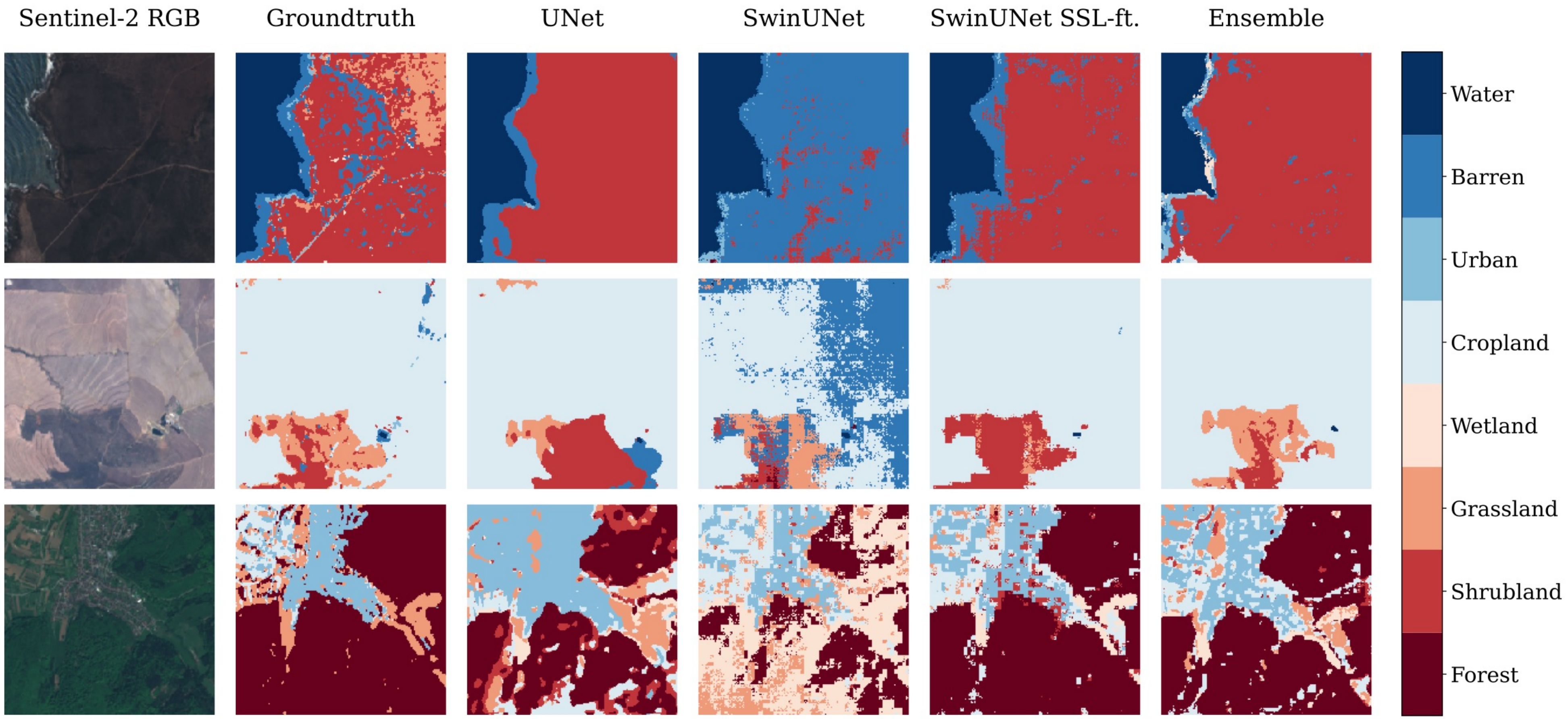


1. Encode Sentinel-1/2 images with distinct encoders
2. Compute contrastive loss on projected representations
3. Replace projection head by downstream task specific module

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\mathrm{sim}(\mathcal{R}_i, \mathcal{R}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(\mathcal{R}_i, \mathcal{R}_k)/\tau)}$$

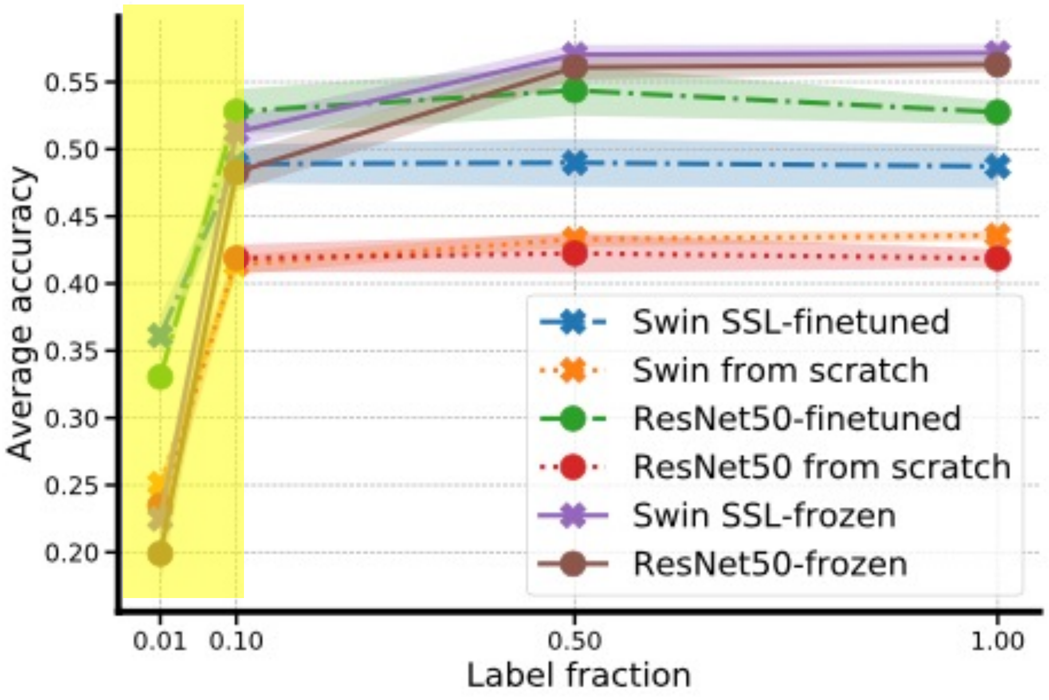
Swin Transformer: [Liu 2021]

Results

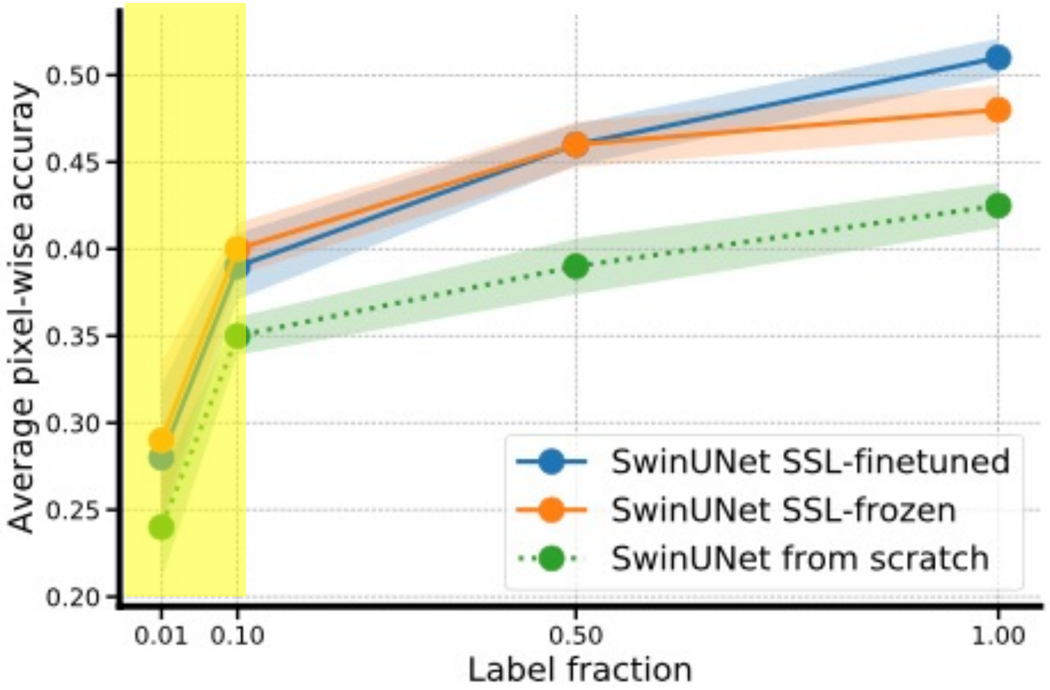

Sentinel-2 RGB
Groundtruth
UNet
SwinUNet
SwinUNet SSL-ft.
Ensemble
Water
Barren
Urban
Cropland
Wetland
Grassland
Shrubland
Forest

Results

Classification



Segmentation



SSL pre-training and 10-20% of labeled data outperform fully supervised training

Universität St.Gallen

Hyper-Representation Learning

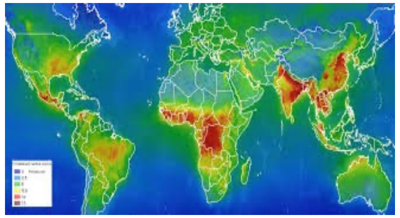
K Schürholt, D Kostadinov, D Borth
Self-Supervised Representation Learning on Neural Network Weights for Model Characteristic Prediction
Neural Information Processing Systems (NeurIPS), 2021

K Schürholt, B Knyazev, X Giró-i-Nieto, D Borth
Hyper-Representations as Generative Models: Sampling Unseen Neural Network Weights
Neural Information Processing Systems (NeurIPS), 2022

K Schürholt, D Taskiran, B Knyazev, X Giró-i-Nieto, D Borth
Model Zoos: A Dataset of Diverse Populations of Neural Network Models
Neural Information Processing Systems (NeurIPS), 2022
[Google Research Scholar Award 2022]

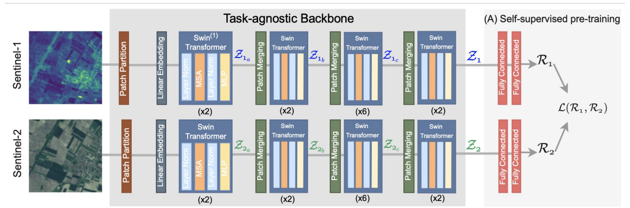
Overview

Shared-Backbones/Heads	Self-supervised Learning	Hyper-Representations



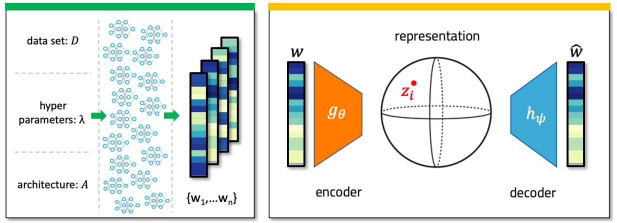
Approach:
- Muilti-modal Fusion
- Multi-task Learning
- Auxiliary Tasks

Application
- NO2 estimation
- Power Production
- CO2 estimation



Approach:
- Contrastive Learning
- Augmentation free
- CNNs & Transformer

Application
- Land-use Classification
- Single-class / Multi-class
- Segmentation



Approach:
- Contrastive Learning
- Model Zoos
- CNNs

Application
- Model analysis
- Sample unseen models
- Sparsificaiton

Neural Network Training

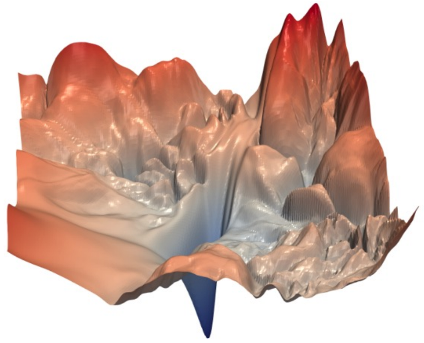
Neural Networks are successfully applied on multiple domains

Loss surface and optimization problem of Neural Networks are non-convex

Goodfellow, Vinyals, Saxe; ICLR 2015; *Qualitatively characterizing neural network optimization problems*

Dauphin et al.; NeurIPS 2014; *Identifying and attacking the saddle point problem in high-dimensional non-convex optimization*

LeCun, Bengion, Hinton; Nature 2015; *Deep Learning*



Li et al.; NeurIPS 2018; *Visualizing the Loss Landscape of Neural Nets*

Neural Network training optimization is high dimensional

Brown et al.; 2020; *Language Models are Few-Shot Learners*

Larsen et al.; ICML 2021; *How many degrees of freedom do we need to train deep networks: a loss landscape perspective*

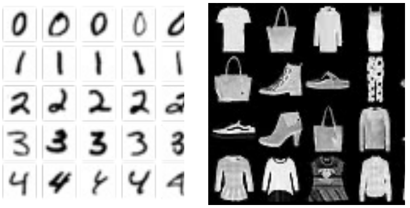
Neural Network training is sensitive to hyperparameters and random initialization

Hanin, Rolnick; NeurIPS 2018; *How to Start Training: The Effect of Initialization and Architecture*

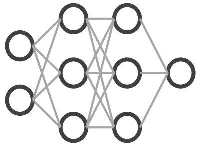
**We want to better understand the relation
between properties of NN models and their solution in weight space**

Investigating Populations of NN Models

Dataset



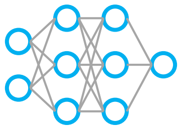
Architecture



Hyperparamerters

- Optimizer
- Activation
- Initialization Method
- Learning Rate
- L2-Regularization

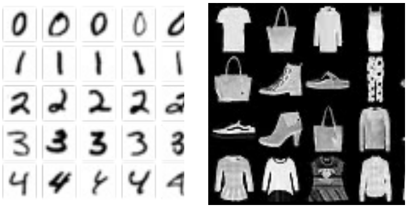
Model



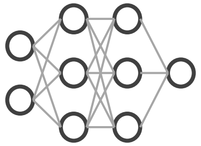
Investigating Populations of NN Models

Hypothesis:
1. Neural Networks populate a structure in weight space
2. That structure contains information on properties and generating factors of the models

Dataset



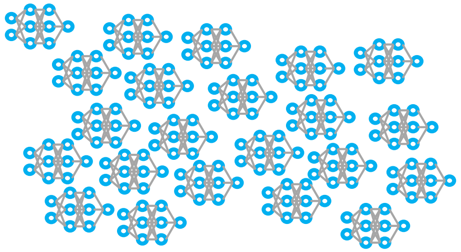
Architecture



Hyperparamerters

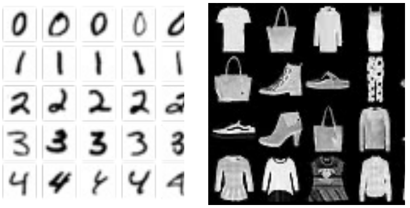
- Optimizer
- Activation
- Initialization Method
- Learning Rate
- L2-Regularization

Model Population

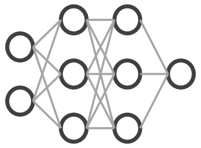


Investigating Populations of NN Models

Dataset



Architecture



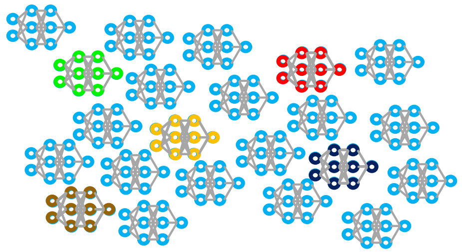
Hyperparamerters

- Optimizer
- Activation
- Initialization Method
- Learning Rate
- L2-Regularization

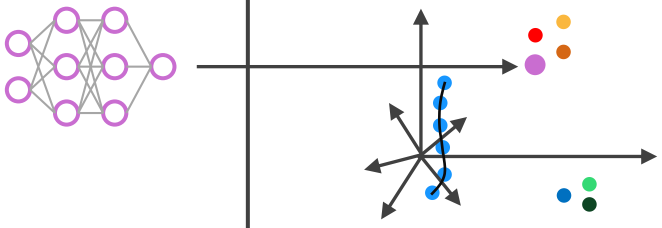
Hypothesis:

1. Neural Networks populate a structure in weight space
2. That structure contains information on properties and generating factors of the models

Model Population



Representation Space

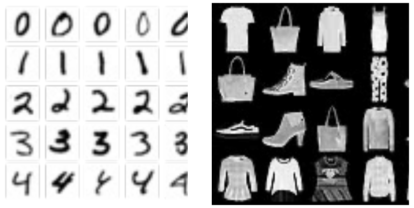


Goal: Learn meaningful representations of populations of Neural Network models

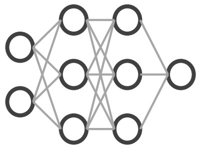
Investigating Populations of NN Models

Universität St.Gallen

Dataset



Architecture



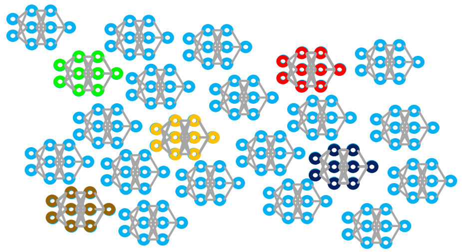
Hyperparamerters

- Optimizer
- Activation
- Initialization Method
- Learning Rate
- L2-Regularization

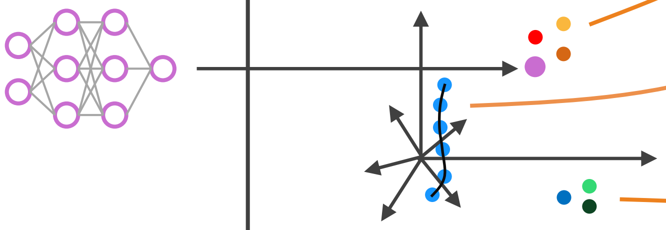
Hypothesis:

1. Neural Networks populate a structure in weight space
2. That structure contains information on properties and generating factors of the models

Model Population



Representation Space

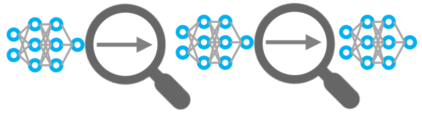


Goal: Learn meaningful representations of populations of Neural Network models

Model Analysis

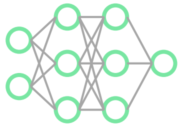
versioning, diagnostics, ...

Learning Dynamics



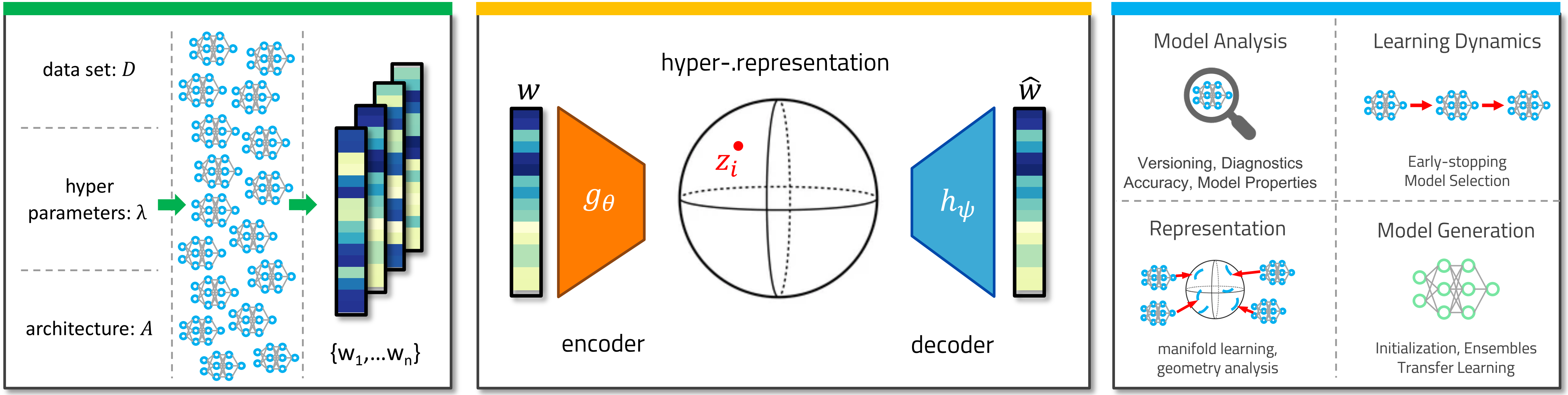
early-stopping, model selection, ...

Model Generation



initialization, transfer-learning, meta-learning, ...

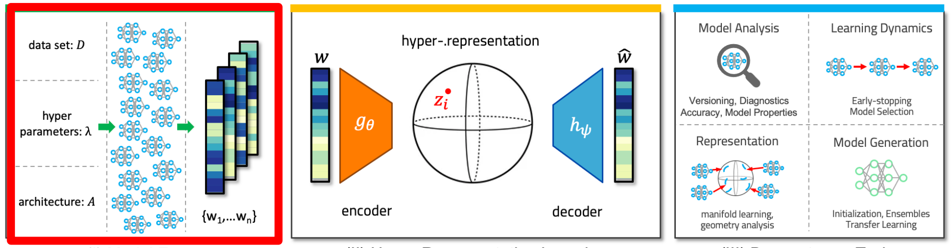
Approach



(I) Model Zoos

(II) Hyper-Representation Learning

(III) Down-stream Tasks



(I) Model Zoos (II) Hyper-Representation Learning (III) Down-stream Tasks

Model Zoos

Datasets:

- MNIST, F-MNIST, SVHN, USPS, STL, CIFAR10, CIFAR100, Tiny ImageNet, EuroSAT

Architectures

- **CNN:** 2464 paramters (ours)
- **CNN:** 4970 paramters (Unterthiner et al., 2020)
- **ResNet-18:** 11 million parameters (He, 2015)

Hyperparamters

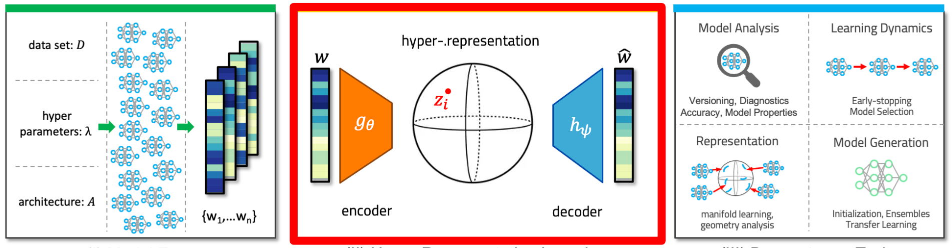
- Seed, activation, initialization method, learning rate, regularization, ...

- **More than 50k neural networks**
- **2.6 million model states**
- **Sparsified Model Twins**

all models are open source: www.modelzoos.cc

Our Zoos	Data	Architecture	Samples
Tetris-Seed	Tetris	MLP (100 params.)	75k
Tetris-Hyp	Tetris	MLP (100 params.)	217.5k
MNIST-Seed	MNIST	CNN (2464 params.)	50k
F-MNIST-Seed	F-MNIST	CNN (2464 params.)	50k
MNIST-Hyp-1-Fix-Seed	MNIST	CNN (2464 params.)	~57.6k
MNIST-Hyp-1-Rand-Seed	MNIST	CNN (2464 params.)	~57.6k
MNIST-Hyp-5-Fix-Seed	MNIST	CNN (2464 params.)	~64k
MNIST-Hyp-5-Rand-Seed	MNIST	CNN (2464 params.)	~64k

Zoos from Unterthiner et al., 2020	Data	Architecture	Samples
MNIST-Hyp	MNIST	CNN (4970 params.)	270k
F-MNIST-Hyp	F-MNIST	CNN (4970 params.)	270k
CIFAR-Hyp	CIFAR10	CNN (4970 params.)	270k
SVHN-Hyp	SVHN	CNN (4970 params.)	270k



(I) Model Zoos (II) Hyper-Representation Learning (III) Down-stream Tasks

NN Weights Augmentations

Augmentations:

- increase number of training samples
- Encode inductive bias

Erasing & Noise:

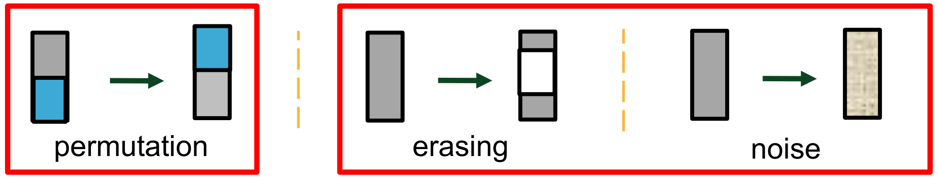
- Adaptations from computer vision

Permutation Augmentation:

- Leverages symmetries in weight space
- Proof: equivalence holds forward & backward
- Scales with faculty of # neurons/kernels
- Fully-connected and convolutional layers

- Full Details are in the appendix of our paper

Augmentations



Assumptions

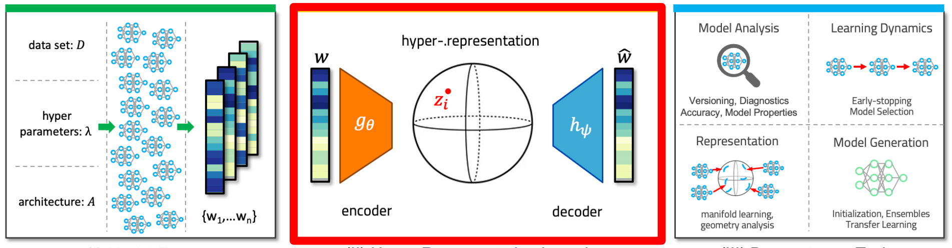
$$(\mathbf{P}^l)^{\mathrm{T}}\mathbf{P}^l = \mathbf{I}, \qquad \mathbf{P}^l\sigma(\mathbf{n}^l) = \sigma(\mathbf{P}^l\mathbf{n}^l),$$

Forward pass

$$\begin{aligned}
\mathbf{n}^{l+1} &= \mathbf{W}^{l+1}\,\mathbf{I}\,\sigma(\mathbf{W}^l\,\mathbf{a}^{l-1} + \mathbf{b}^l) + \mathbf{b}^{l+1} \\
&= \mathbf{W}^{l+1}\,(\mathbf{P}^l)^{\mathrm{T}}\,\mathbf{P}^l\,\sigma(\mathbf{W}^l\,\mathbf{a}^{l-1} + \mathbf{b}^l) + \mathbf{b}^{l+1} \\
&= \mathbf{W}^{l+1}\,(\mathbf{P}^l)^{\mathrm{T}}\,\sigma(\mathbf{P}^l\,\mathbf{W}^l\,\mathbf{a}^{l-1} + \mathbf{P}^l\,\mathbf{b}^l) + \mathbf{b}^{l+1} \\
&= \hat{\mathbf{W}}^{l+1}\,\sigma(\hat{\mathbf{W}}^l\,\mathbf{a}^{l-1} + \hat{\mathbf{b}}^l) + \mathbf{b}^{l+1},
\end{aligned}$$

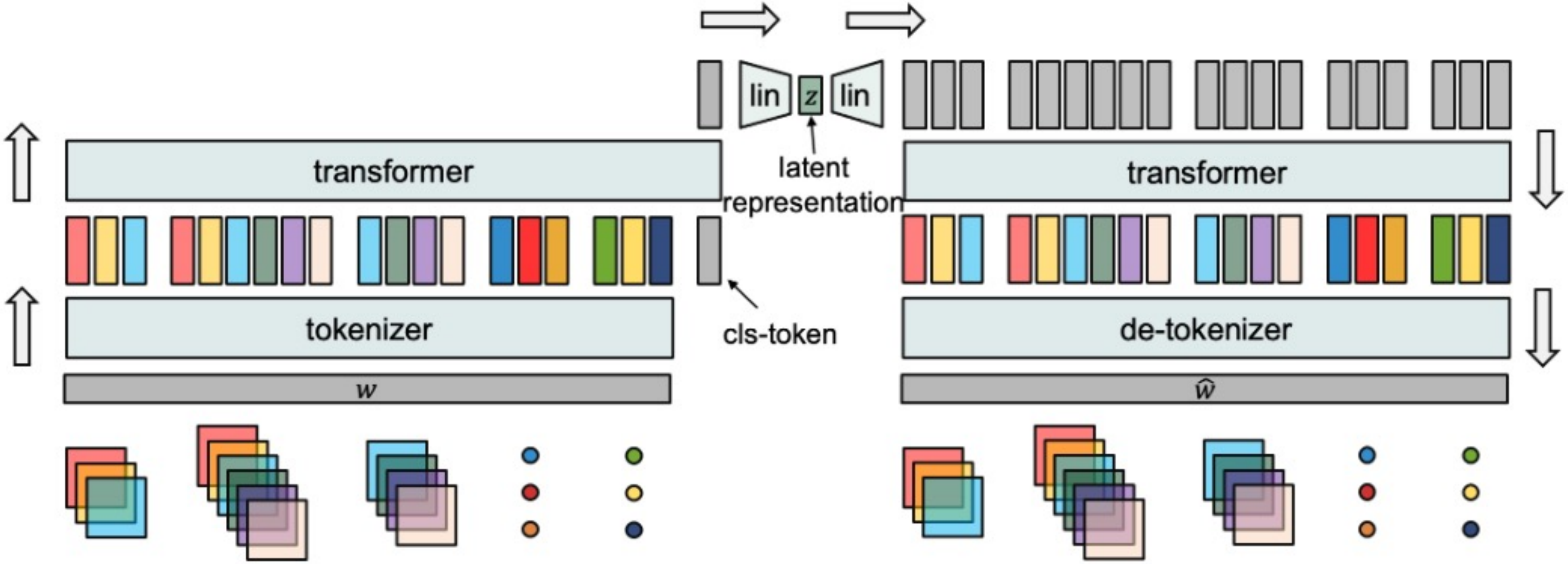
Backward pass

$$\begin{aligned}
(\mathbf{P}^l\mathbf{W}^l)_{\text{new}} &= \mathbf{P}^l\mathbf{W}^l - \alpha\mathbf{P}^l\nabla_{\mathbf{W}^l}\mathcal{L} \\
&= \mathbf{P}^l\mathbf{W}^l - \alpha\mathbf{P}^l\delta^l(\mathbf{a}^{l-1})^{\mathrm{T}} \\
&= \mathbf{P}^l\mathbf{W}^l - \alpha\mathbf{P}^l\left[(\mathbf{W}^{l+1})^{\mathrm{T}}\delta^{l+1} \odot \sigma'(\mathbf{n}^l)\right]\,(\mathbf{a}^{l-1})^{\mathrm{T}} \\
&= \mathbf{P}^l\mathbf{W}^l - \alpha\left[(\mathbf{W}^{l+1}\mathbf{P}^{\mathrm{T}})^{\mathrm{T}}\delta^{l+1} \odot \sigma'(\mathbf{P}^l\mathbf{n}^l)\right]\,(\mathbf{a}^{l-1})^{\mathrm{T}} \\
&= \mathbf{P}^l\mathbf{W}^l - \alpha[(\mathbf{W}^{l+1}(\mathbf{P}^l)^{\mathrm{T}})^{\mathrm{T}}\delta^{l+1} \odot \sigma'(\mathbf{P}^l\mathbf{W}^l\mathbf{a}^{l-1} + \mathbf{P}^l\mathbf{b}^l)](\mathbf{a}^{l-1})^{\mathrm{T}}. \\
(\hat{\mathbf{W}}^l)_{\text{new}} &= \hat{\mathbf{W}}^l - \alpha\left[(\hat{\mathbf{W}}^{l+1})^{\mathrm{T}}\delta^{l+1} \odot \sigma'(\hat{\mathbf{W}}^l\mathbf{a}^{l-1} + \hat{\mathbf{b}}^l)\right]\,(\mathbf{a}^{l-1})^{\mathrm{T}}\square
\end{aligned}$$



(I) Model Zoos (II) Hyper-Representation Learning (III) Down-stream Tasks

Autoencoding Transformer

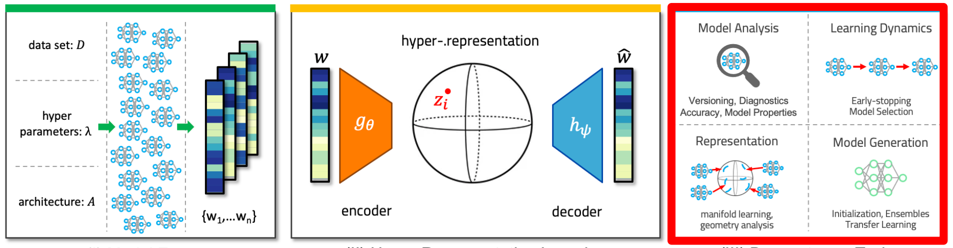


$$\mathcal{L}_c = \sum_{(i,j)} -\log \frac{\exp(\text{sim}(\bar{\mathbf{z}}_i, \bar{\mathbf{z}}_j)/\tau}{\sum_{k=1}^{2M_B} \mathbb{I}_{k \neq i} \exp(\text{sim}(\bar{\mathbf{z}}_i, \bar{\mathbf{z}}_j)/\tau}$$

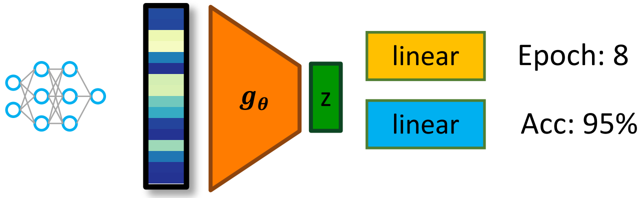
$$\mathcal{L}_{MSE} = \frac{1}{M} \sum_{i=1}^{M} \|\mathbf{w}_i - h_\psi(g_\theta(\mathbf{w}_i))\|_2^2.$$

$$\mathcal{L}_{c+} = \sum_i -\log\left(\exp(\text{sim}(\bar{\mathbf{z}}_i^j, \bar{\mathbf{z}}_i^k))/\tau\right) = \sum_i -\text{sim}(\bar{\mathbf{z}}_i^j, \bar{\mathbf{z}}_i^k) + \log(\tau).$$

Experiment Results



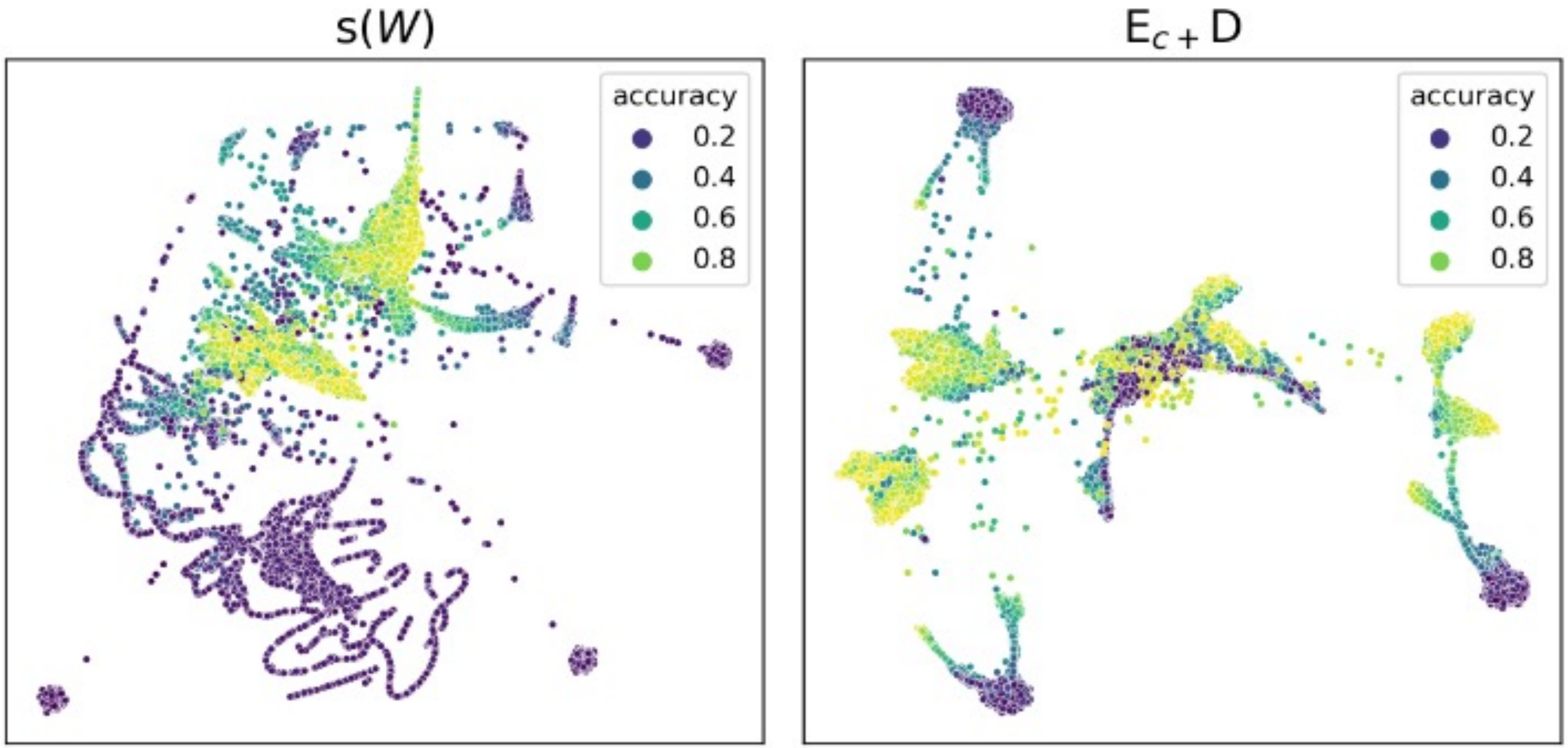
(I) Model Zoos (II) Hyper-Representation Learning (III) Down-stream Tasks



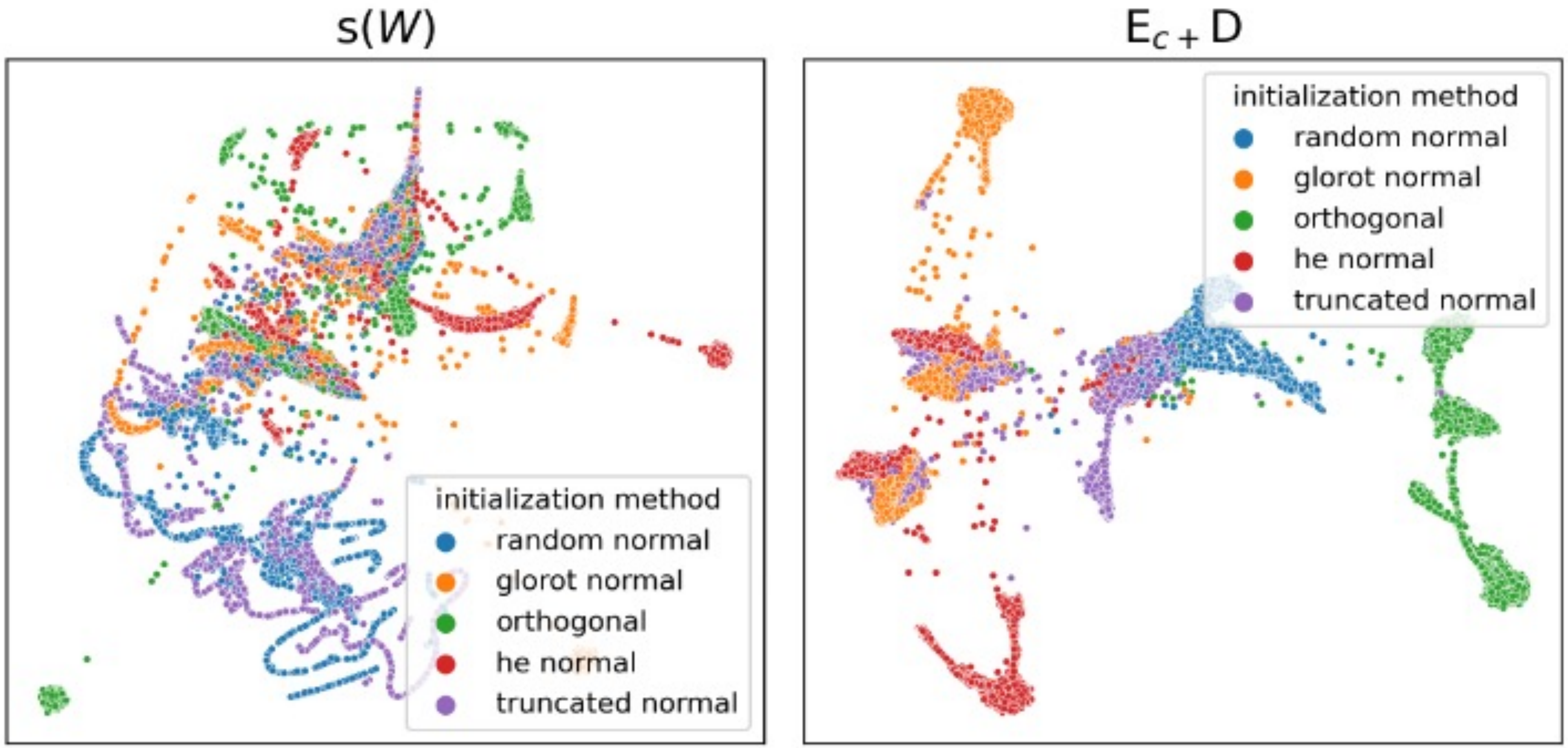
	MNIST-HYP			FASHION-HYP			CIFAR10-HYP			SVHN-HYP		
	W	s(W)	E_cD	W	s(W)	E_cD	W	s(W)	E_cD	W	s(W)	E_cD
EPH	25.8	33.2	**50.0**	26.6	34.6	**51.3**	25.7	30.3	**53.3**	22.8	37.8	**52.6**
ACC	74.7	81.5	**94.9**	70.9	78.5	**96.2**	76.4	82.9	**92.7**	80.5	82.1	**91.1**

the higher -> the better
R^2 for regression downstream tasks

Embedding Homogeneity

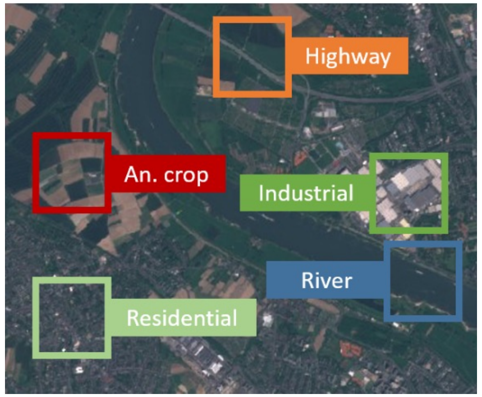


Embedding Homogeneity

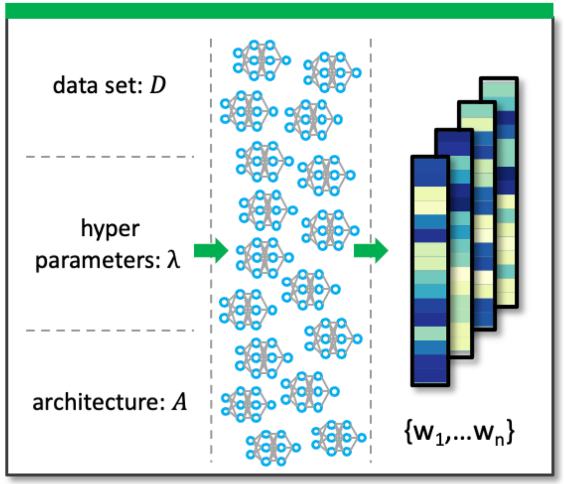


EuroSAT Model Zoo & Sparsified Twins

EuroSAT - Dataset

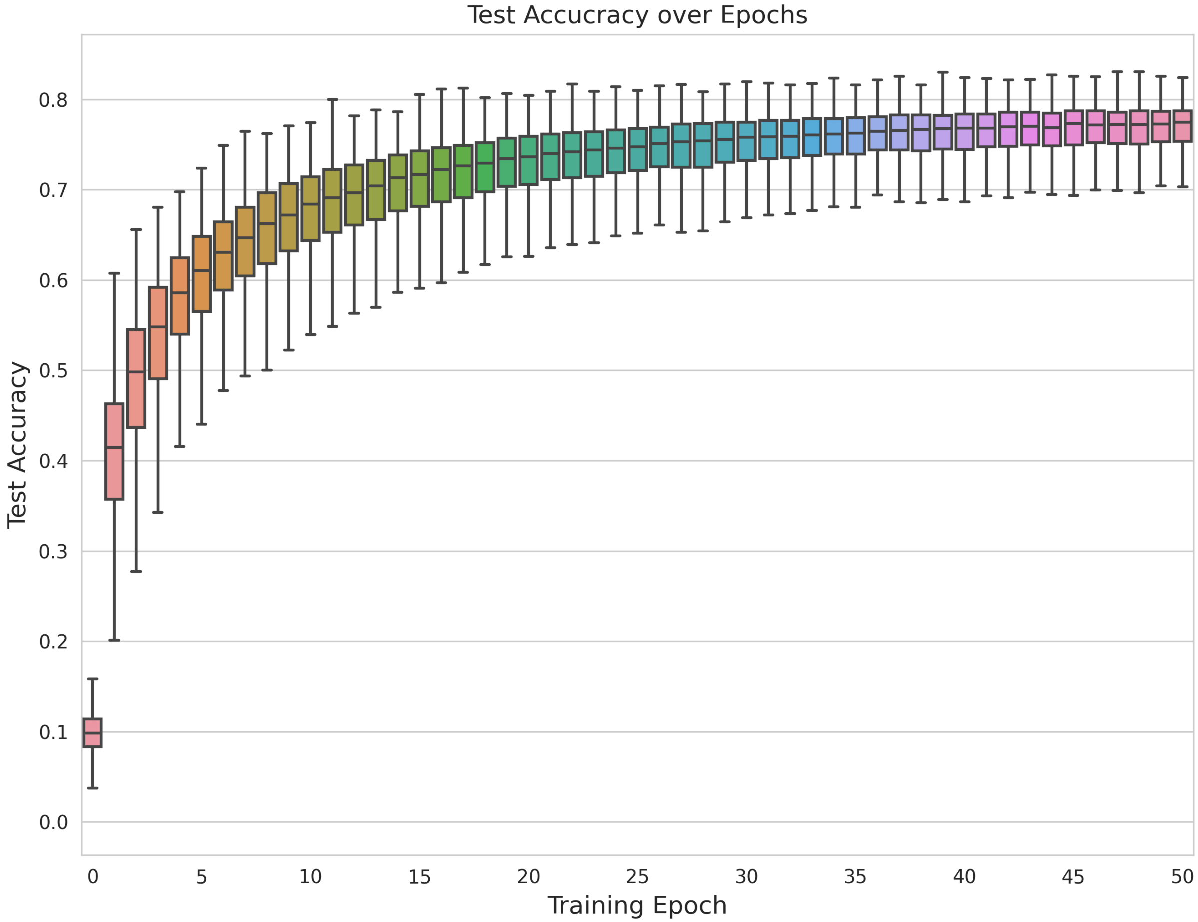


Patch-based Classification



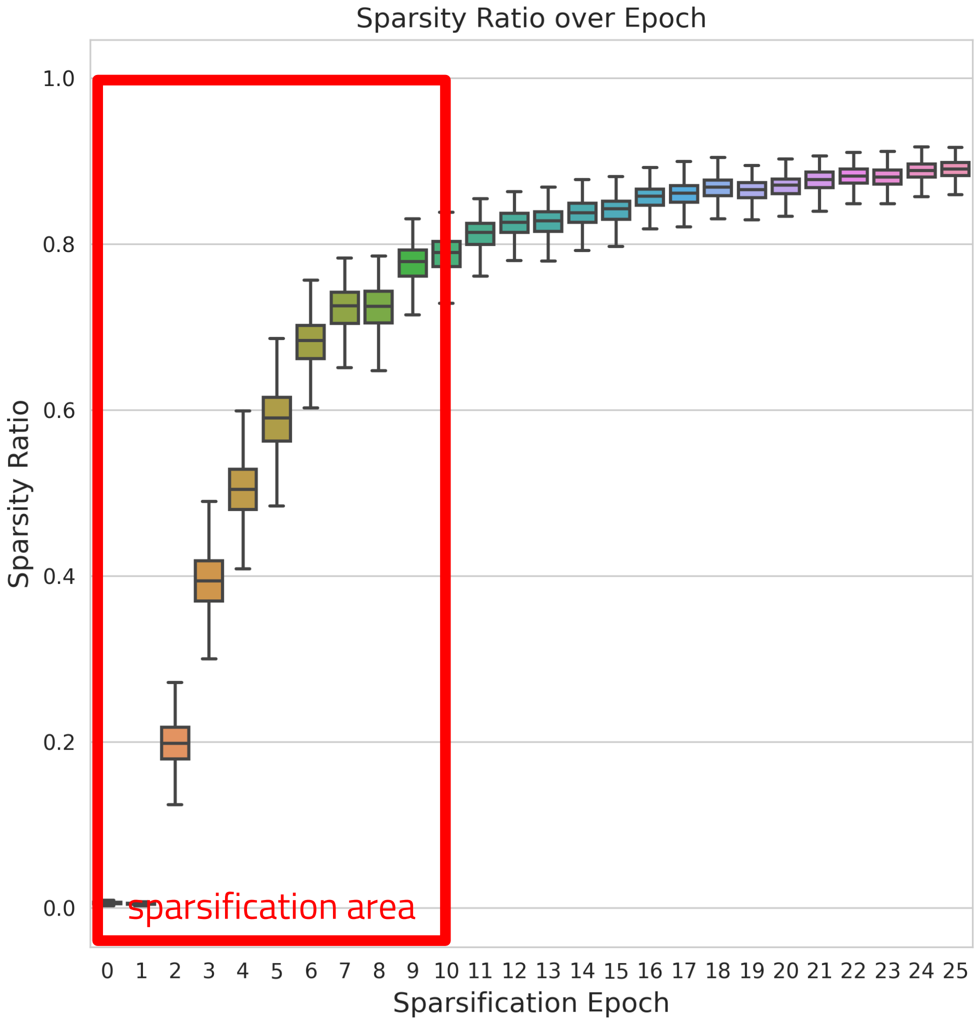
(I) Model Zoos

EuroSAT Model Zoo

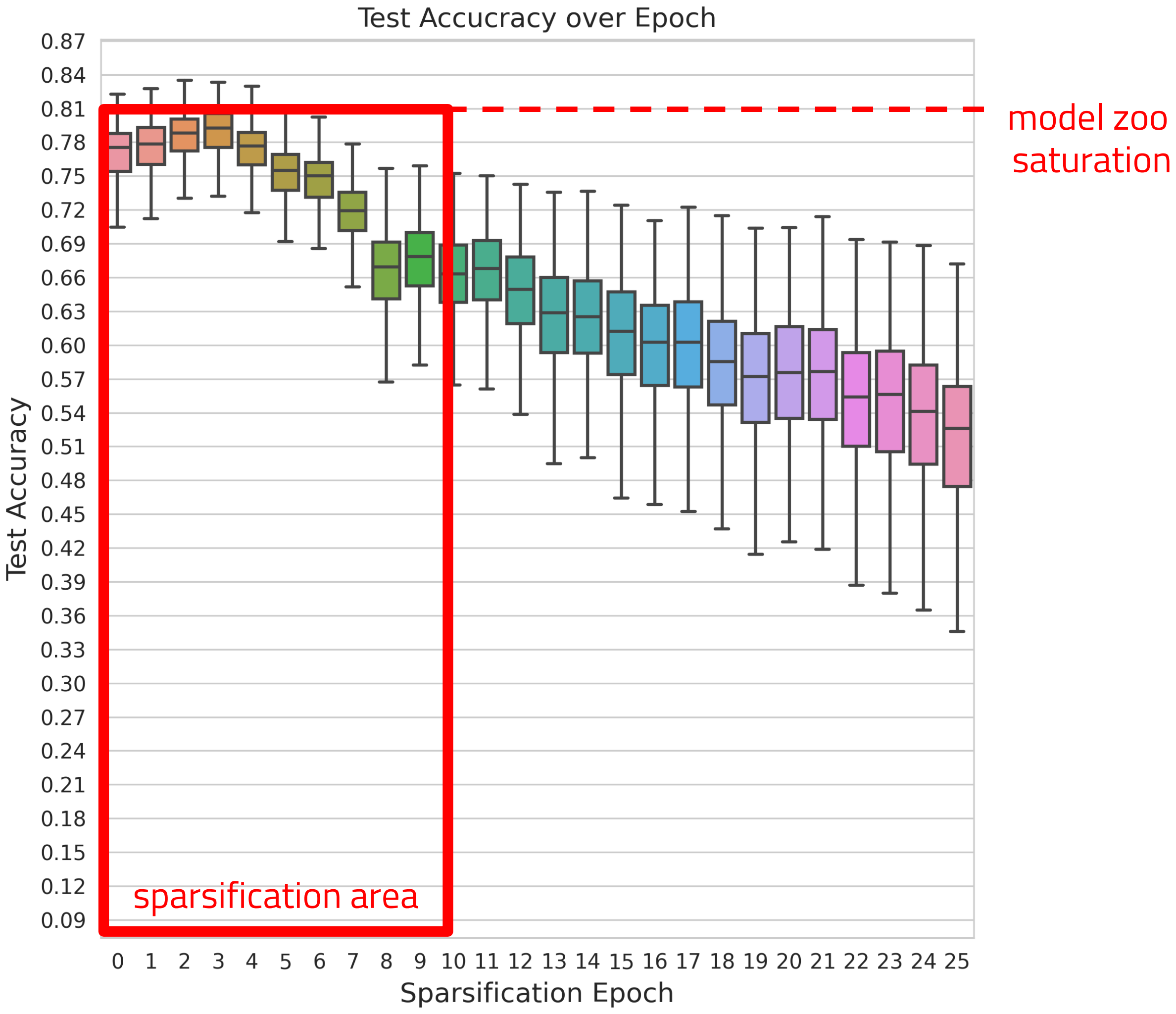


EuroSAT Model Zoo & Sparsified Twins

Sparsity Ratio



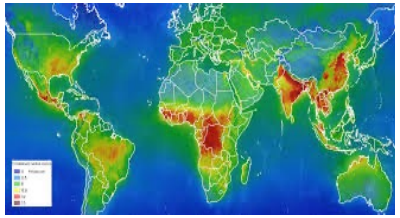
Test Accuracy



Universität St.Gallen

Overview

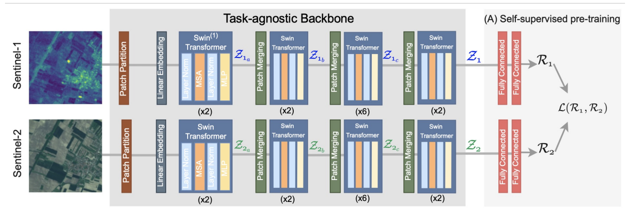
Shared-Backbones/Heads



Approach:
- Muilti-modal Fusion
- Multi-task Learning
- Auxiliary Tasks

Application
- NO2 estimation
- Power Production
- CO2 estimation

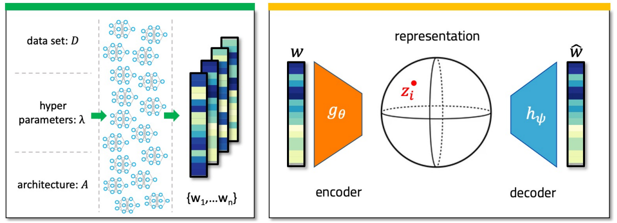
Self-supervised Learning



Approach:
- Contrastive Learning
- Augmentation free
- CNNs & Transformer

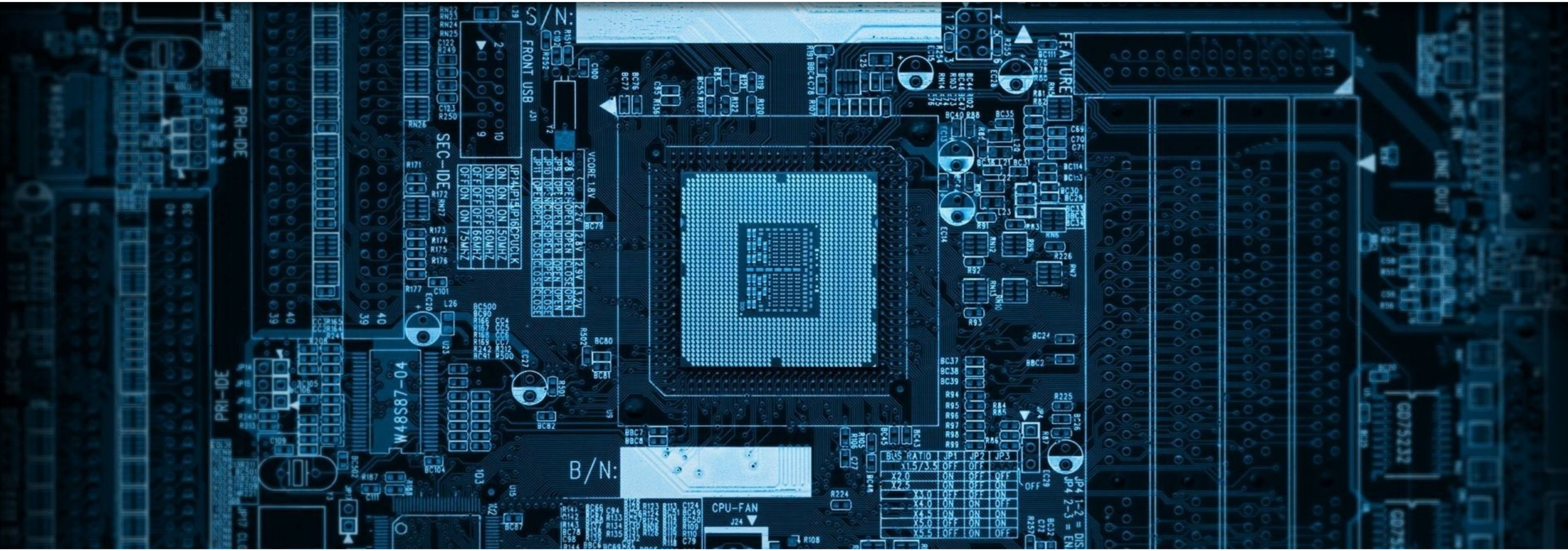
Application
- Land-use Classification
- Single-class / Multi-class
- Segmentation

Hyper-Representations



Approach:
- Contrastive Learning
- Model Zoos
- CNNs

Application
- Model analysis
- Sample unseen models
- Sparsificaiton


S/N:
FRONT USB
SEC-IDE
PRI-IDE
W48S87-04
B/N:
CPU-FAN
FEATURE
LINE OUT

Questions?