Audiovisual Self-Supervised Learning for Remote Sensing Data

Konrad Heidler[1,2], Lichao Mou[1,2], Di Hu[3], Pu Jin[1,2], Guangyao Li[3], Chuang Gan[4], Ji-Rong Wen[3], Xiao Xiang Zhu[1,2]

[1] German Aerospace Center (DLR)
[2] Technische Universität München (TUM)
[3] Renmin University of China
[4] MIT-IBM Watson AI Lab

26.05.2022

Audiovisual Learning in Remote Sensing

Idea

Audiovisual Learning in Remote Sensing

Idea

- Recent SSL methods rely on generating different “views” of the data

Audiovisual Learning in Remote Sensing

Idea

- Recent SSL methods rely on generating different “views” of the data
- Consider imagery and local audio as drastically different “views”

Audiovisual Learning in Remote Sensing

Idea

- Recent SSL methods rely on generating different “views” of the data
- Consider imagery and local audio as drastically different “views”
- Use Audiovisual SSL as pretraining for other tasks

Audiovisual Data in Remote Sensing

Existing Geo-Audio Datasets

Audiovisual Data in Remote Sensing

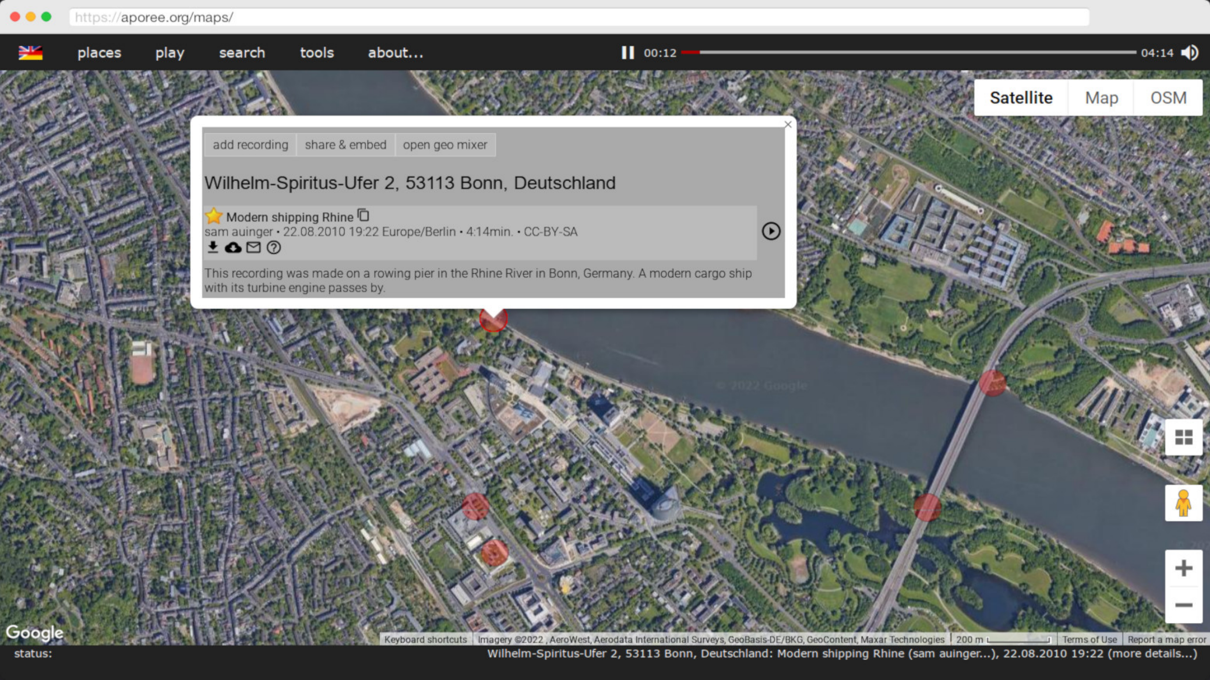
Existing Geo-Audio Datasets

- CVS [1] – Unspecified audio content (Freesound)

Audiovisual Data in Remote Sensing

Existing Geo-Audio Datasets

- CVS [1] – Unspecified audio content (Freesound)
- ADVANCE [2] – Small (∼5,000 samples)


https://aporee.org/maps/
places
play
search
tools
about...
00:12
04:14
Satellite
Map
OSM
add recording
share & embed
open geo mixer
Wilhelm-Spiritus-Ufer 2, 53113 Bonn, Deutschland
Modern shipping Rhine
sam auinger • 22.08.2010 19:22 Europe/Berlin • 4:14min. • CC-BY-SA
This recording was made on a rowing pier in the Rhine River in Bonn, Germany. A modern cargo ship with its turbine engine passes by.
Google
Keyboard shortcuts
Imagery ©2022 , AeroWest, Aerodata International Surveys, GeoBasis-DE/BKG, GeoContent, Maxar Technologies
200 m
Terms of Use
Report a map error
status:
Wilhelm-Spiritus-Ufer 2, 53113 Bonn, Deutschland: Modern shipping Rhine (sam auinger...), 22.08.2010 19:22 (more details...)

SoundingEarth Dataset

Radio Aporee:::Maps

- Crowd-sourced database of geo-tagged field recordings.

SoundingEarth Dataset

Radio Aporee:::Maps

- Crowd-sourced database of geo-tagged field recordings.
  Field Recording: Audio recording taken to capture the ambience of a scene.

SoundingEarth Dataset

Radio Aporee:::Maps

- Crowd-sourced database of geo-tagged field recordings.
  Field Recording: Audio recording taken to capture the ambience of a scene.
- Good spatial coverage, with focus on Europe.
- High quality audio.

SoundingEarth Dataset

Radio Aporee:::Maps

- Crowd-sourced database of geo-tagged field recordings.
  Field Recording: Audio recording taken to capture the ambience of a scene.
- Good spatial coverage, with focus on Europe.
- High quality audio.

Sounding Earth Dataset

- Download Aporee audio and convert to log-mel spectrograms
- Pair spectrograms with corresponding Google Earth imagery

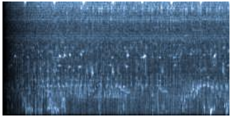
SoundingEarth Dataset

Lake Bunyonyi, Uganda

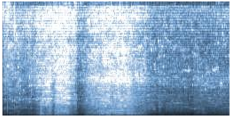
Tokyo, Japan

SoundingEarth Dataset

Lake Bunyonyi, Uganda

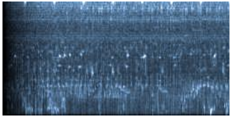


Tokyo, Japan

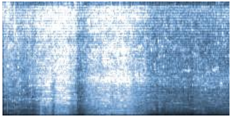


SoundingEarth Dataset

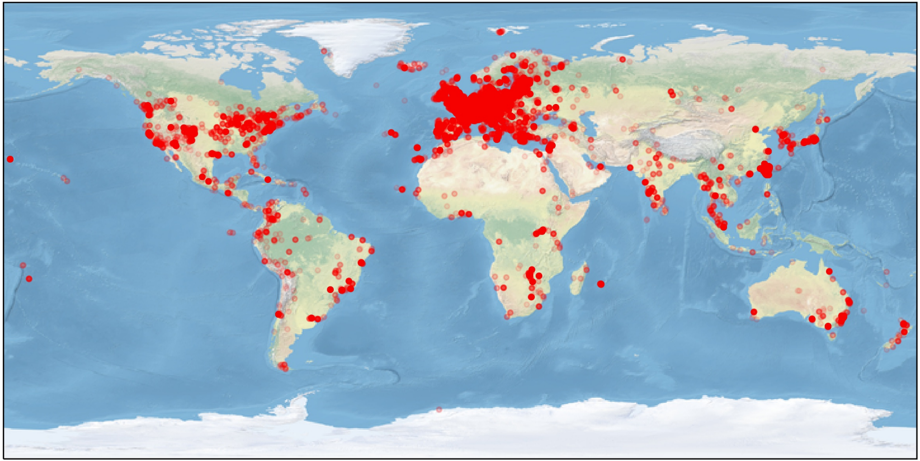
Lake Bunyonyi, Uganda



Tokyo, Japan



Spatial Distribution



SoundingEarth Dataset Overview

Per Sample

- Raw Audio (mp3)
- Log-mel Spectrogram ($128 \times \mathrm{T}$)
- Google Earth Imagery ($1024 \times 1024 \times 3$)

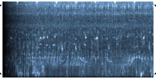
Dataset

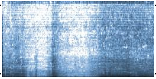
- 50,545 samples
- $\sim$3500 hours of audio.

Multi-Modal Self-Supervised Learning

Task Formulation

Find/train embedding functions f_{image} and f_{audio}, such that

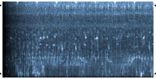
$f_{\text{image}}\Big($  $\Big) \approx f_{\text{audio}}\Big($  $\Big)$ for corresponding pairs, and

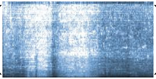
$f_{\text{image}}\Big($  $\Big) \not\approx f_{\text{audio}}\Big($  $\Big)$ for unrelated pairs

Multi-Modal Self-Supervised Learning

Task Formulation

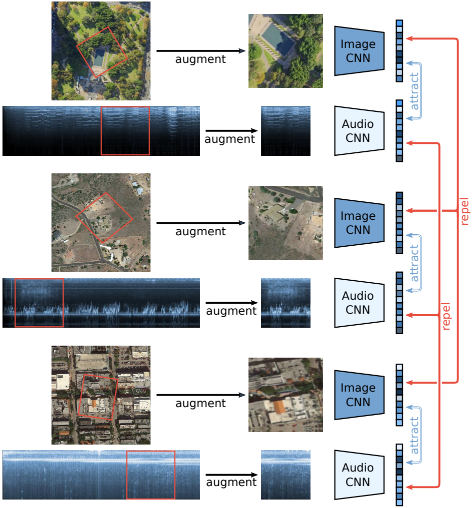
Find/train embedding functions f_{image} and f_{audio}, such that

$f_{\text{image}}\Big($  $\Big) \approx f_{\text{audio}}\Big($  $\Big)$ for corresponding pairs, and

$f_{\text{image}}\Big($  $\Big) \not\approx f_{\text{audio}}\Big($  $\Big)$ for unrelated pairs

$\rightarrow$ Choose ResNets as prototypes for f_{image} and f_{audio}.

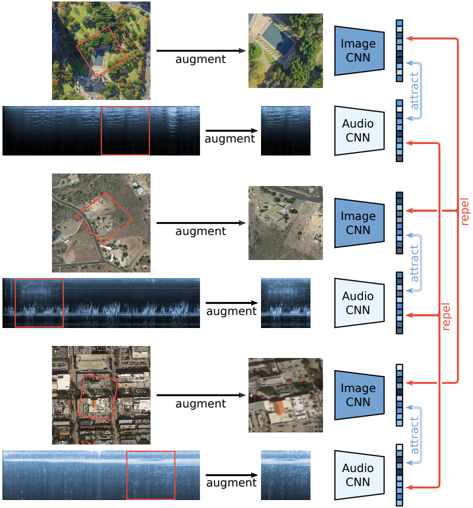
Framework


augment
augment
augment
augment
augment
augment
Image CNN
Audio CNN
Image CNN
Audio CNN
Image CNN
Audio CNN
attract
attract
attract
repel
repel

Framework

Data Augmentation

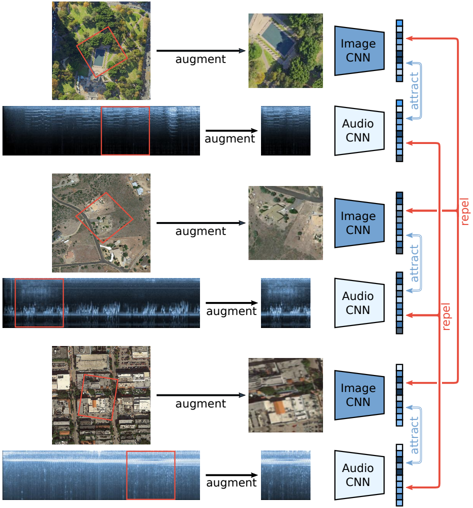
- Imagery: Flips, Rotations, Crops
- Spectrograms: Temporal Crops



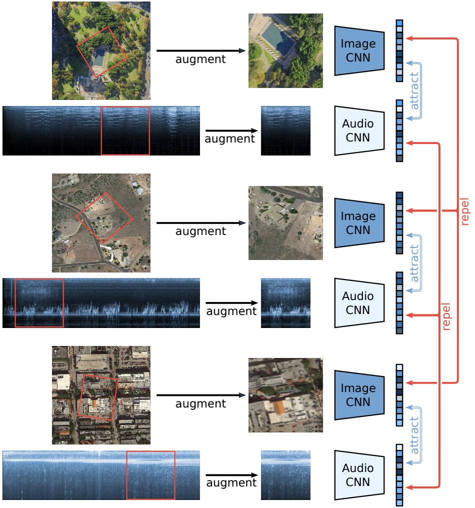
Framework

Training Steps

- Forward Pass: Calculate Embeddings
- Calculate Loss
- Backpropagate to update f_{image} and f_{audio}



Framework



Loss Function

- Pull corresponding embeddings together
- Push other embeddings apart
- Evaluate multiple loss functions

Loss Function

SSL Loss Functions

- Triplet Loss: $L_{\text{triplet}} = \max\{\, \mathrm{d}_{\text{false}} - \mathrm{d}_{\text{true}} + 1, 0\,\}$
- Contrastive Loss: $L_{\text{contrastive}} = \frac{\exp(\mathrm{sim}(z_i, z_j)/\tau)}{\sum_{k=1,k\neq i}^{2N} \exp(\mathrm{sim}(z_i, z_k)/\tau)}$

Loss Function

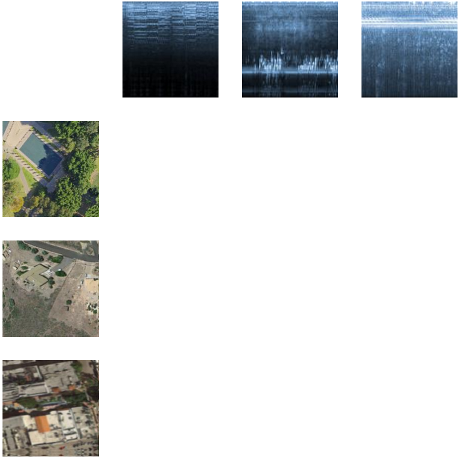
SSL Loss Functions

- Triplet Loss: $L_{\text{triplet}} = \max\{\,\mathrm{d}_{\text{false}} - \mathrm{d}_{\text{true}} + 1, 0\,\}$
- Contrastive Loss: $L_{\text{contrastive}} = \frac{\exp(\mathrm{sim}(z_i, z_j)/\tau)}{\sum_{k=1,k\neq i}^{2N} \exp(\mathrm{sim}(z_i, z_k)/\tau)}$

Observations

- Contrastive Loss requires large batch sizes to work well
- Triplet Loss is “wasteful”
- Implement batch-all Triplet Loss

Loss Function



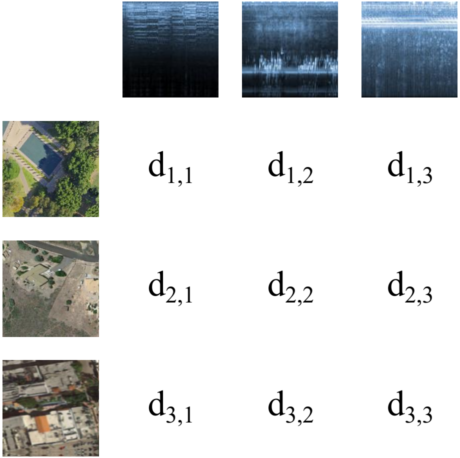
Batch-all Triplet Loss

Calculate Triplet Loss for all possible triplets in a batch

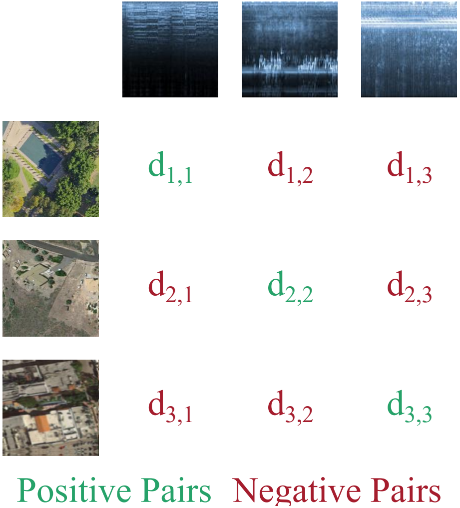
Loss Function

Batch-all Triplet Loss

Calculate Triplet Loss for all possible triplets in a batch



Loss Function



Batch-all Triplet Loss

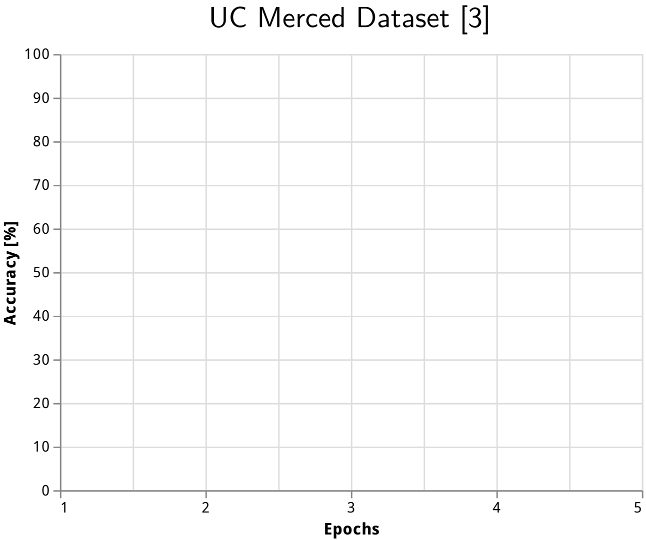
Calculate Triplet Loss for all possible triplets in a batch

Downstream: Aerial Image Classification

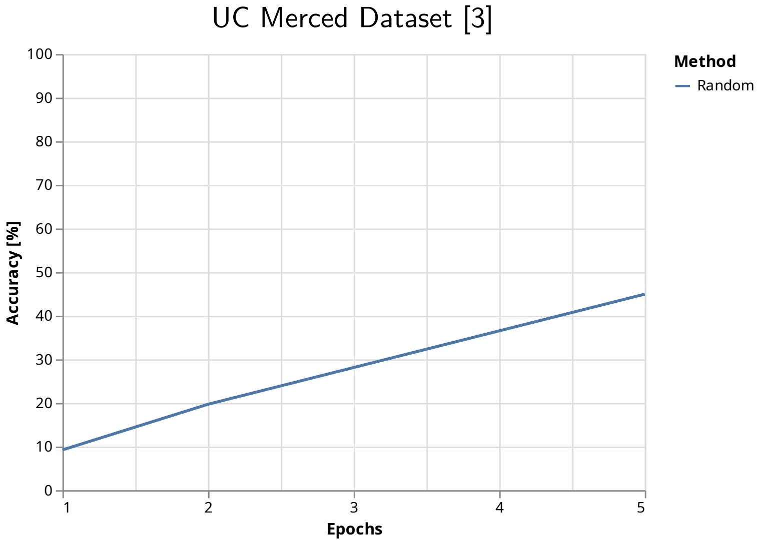
Experiment

- Compare different pre-training methods
- Fine-tune on different datasets
- Architecture: Add classification head (FC Layer) to ResNet-50 backbones
- Just using imagery, no audio

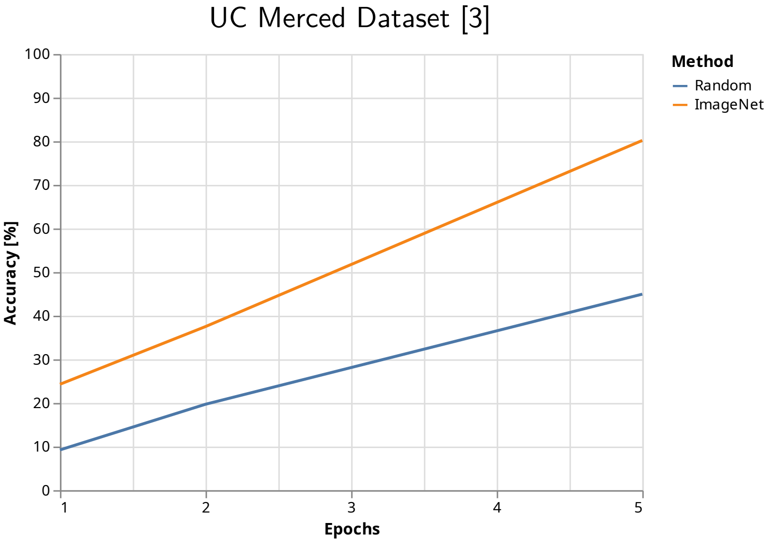
Downstream: Aerial Image Classification



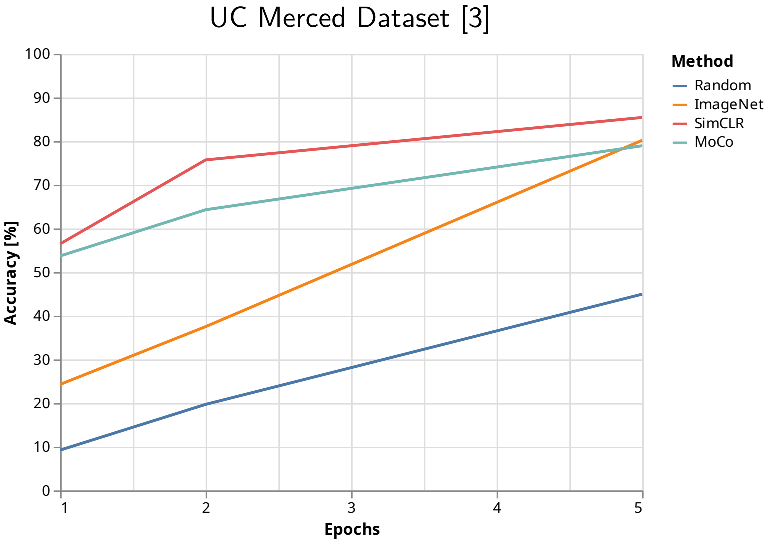
Downstream: Aerial Image Classification



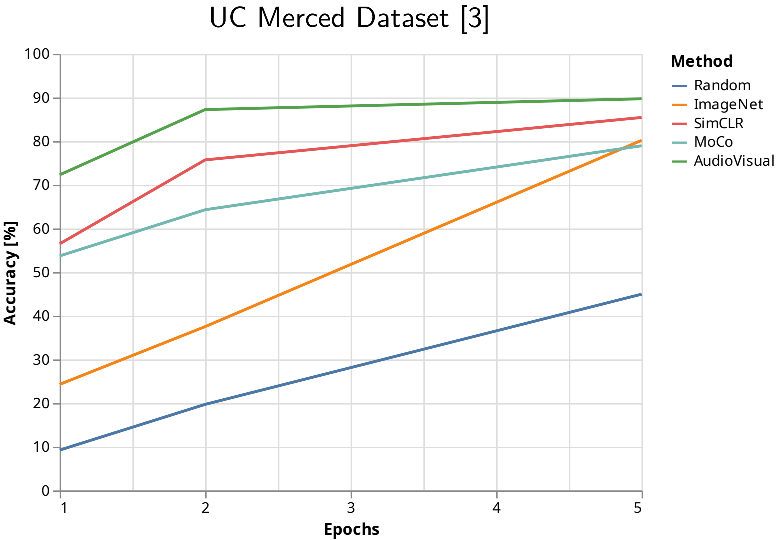
Downstream: Aerial Image Classification



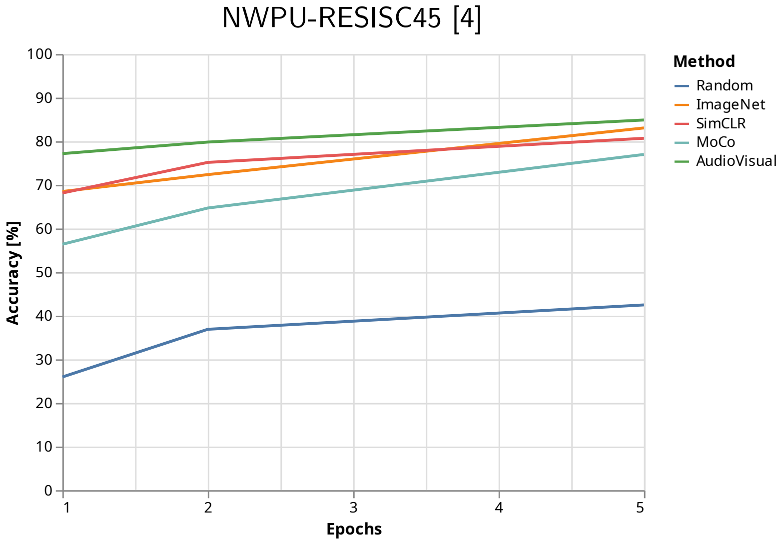
Downstream: Aerial Image Classification



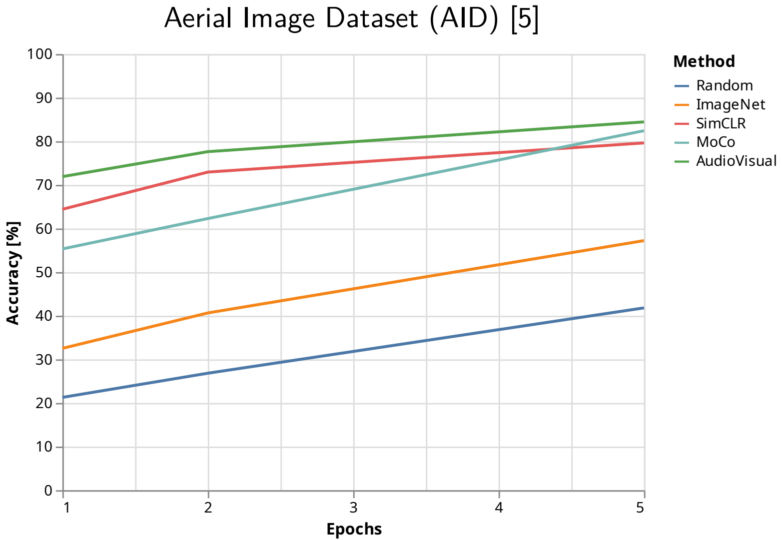
Downstream: Aerial Image Classification



Downstream: Aerial Image Classification



Downstream: Aerial Image Classification



Downstream: Aerial Image Segmentation

Experiment

- Different Application: Image Segmentation
- Fine-tune DeepLabv3+ with pre-trained ResNet-50 backbones
- Dataset: DeepGlobe 2018 [6]
- Just using imagery, no audio

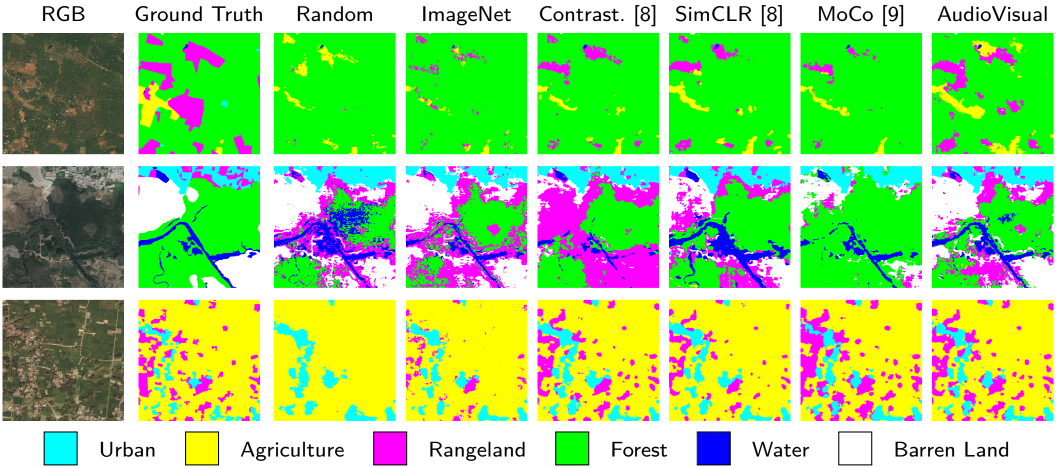
Downstream: Aerial Image Segmentation

	ResNet-18		ResNet-50	
Weights	OA	mIoU	OA	mIoU
Random				
ImageNet				
Tile2Vec [7]				
Contrastive [8]				
SimCLR [8]				
MoCo [9]				
AudioVisual				

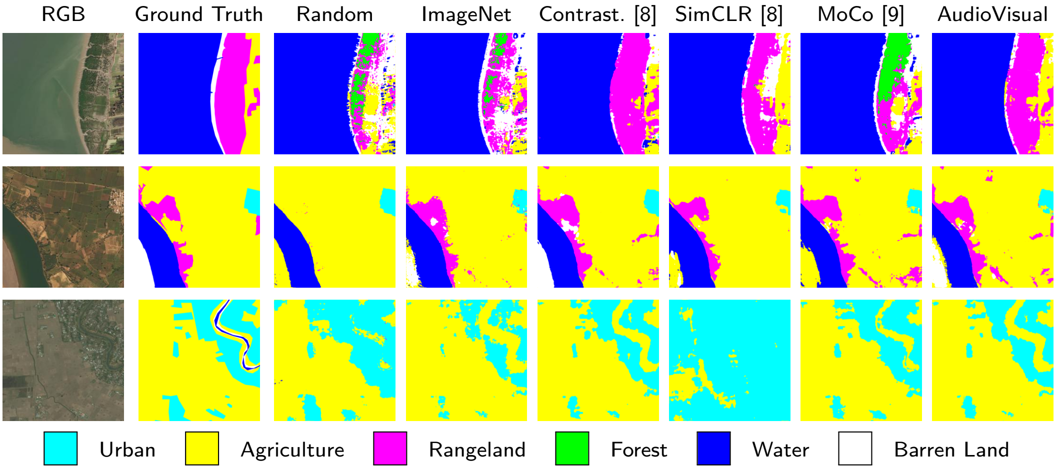
Downstream: Aerial Image Segmentation

	ResNet-18		ResNet-50	
Weights	OA	mIoU	OA	mIoU
Random	81.09	55.38	80.81	54.42
ImageNet	83.27	61.95	82.27	59.31
Tile2Vec [7]	80.50	56.93	—	—
Contrastive [8]	85.25	64.85	86.06	**68.46**
SimCLR [8]	85.65	66.15	83.80	63.97
MoCo [9]	84.79	65.28	85.07	66.17
AudioVisual	**86.11**	**67.07**	**86.58**	67.87

Downstream: Aerial Image Segmentation



Downstream: Aerial Image Segmentation



Downstream: Audiovisual Scene Classification

Experiment

- ADVANCE Dataset [2]: Audiovisual dataset with scene labels
- Linear Evaluation Protocol
- Compare against supervised baseline

Downstream: Audiovisual Scene Classification

Data Used	Audio F_1	Image F_1	Audio + Image F_1
Supervised Baseline [2]			
Ours (ResNet-18)			
Ours (ResNet-50)			

Downstream: Audiovisual Scene Classification

Data Used	Audio F_1	Image F_1	Audio + Image F_1
Supervised Baseline [2]	28.99		
Ours (ResNet-18)	37.69		
Ours (ResNet-50)	39.01		

Downstream: Audiovisual Scene Classification

Data Used	Audio F_1	Image F_1	Audio + Image F_1
Supervised Baseline [2]	28.99	72.85	
Ours (ResNet-18)	37.69	86.92	
Ours (ResNet-50)	39.01	83.84	

Downstream: Audiovisual Scene Classification

Data Used	Audio F_1	Image F_1	Audio + Image F_1
Supervised Baseline [2]	28.99	72.85	74.58
Ours (ResNet-18)	37.69	86.92	89.50
Ours (ResNet-50)	39.01	83.84	88.83

Conclusion

- Additional modalities like audio are beneficial for SSL

Conclusion

- Additional modalities like audio are beneficial for SSL
- The more different the modalities, the better?

Conclusion

- Additional modalities like audio are beneficial for SSL
- The more different the modalities, the better?

Conclusion

- Additional modalities like audio are beneficial for SSL
- The more different the modalities, the better?

Dataset / Code / Pre-trained Models

`https://github.com/khdlr/SoundingEarth`

References i

[1] Tawfiq Salem et al. "A Multimodal Approach to Mapping Soundscapes". In: *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*. IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium. Valencia: IEEE, July 2018, pp. 3477–3480.

[2] Di Hu et al. "Cross-Task Transfer for Geotagged Audiovisual Aerial Scene Recognition". In: *ECCV 2020* (2020), p. 17.

[3] Yi Yang and Shawn D. Newsam. "Bag-of-Visual-Words and Spatial Extensions for Land-Use Classification". In: *18th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems, ACM-GIS 2010, November 3-5, 2010, San Jose, CA, USA, Proceedings*. Ed. by Divyakant Agrawal et al. ACM, 2010, pp. 270–279.

[4] Gong Cheng, Junwei Han, and Xiaoqiang Lu. "Remote Sensing Image Scene Classification: Benchmark and State of the Art". In: *Proc. IEEE* 105.10 (Oct. 2017), pp. 1865–1883.

References ii

[5] G. Xia et al. "AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification". In: *IEEE Trans. Geosci. Remote Sens.* 55.7 (July 2017), pp. 3965–3981.

[6] Ilke Demir et al. "DeepGlobe 2018: A Challenge to Parse the Earth through Satellite Images". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. June 2018, pp. 172–17209. arXiv: `1805.06561 [cs]`.

[7] Neal Jean et al. "Tile2Vec: Unsupervised Representation Learning for Spatially Distributed Data". In: *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, the Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, the Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 2019, pp. 3967–3974.

[8] Ting Chen et al. "A Simple Framework for Contrastive Learning of Visual Representations". In: *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 1597–1607.

References iii

[9] Kaiming He et al. "Momentum Contrast for Unsupervised Visual Representation Learning". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. IEEE, 2020, pp. 9726–9735.

Backup Slides

Loss Function Ablation

		Naive TL		Contrastive Loss		Batch-all TL	
Benchmark	Metric	RN-18	RN-50	RN-18	RN-50	RN-18	RN-50
UC Merced Land Use [3]	Acc.	5.14	77.43	86.48	88.19	**90.19**	**89.71**
NWPU-RESISC45 [4]	Acc.	76.11	72.15	80.65	82.41	**81.71**	**84.88**
AID [5]	Acc.	78.70	75.64	77.18	81.08	**81.78**	**84.44**
DeepGlobe Land Cover [6]	Acc.	83.96	85.40	80.72	85.96	**86.11**	**86.58**
	mIoU	63.14	65.18	57.26	67.28	**67.07**	**67.87**
ADVANCE [2]	F_1	88.51	87.61	79.42	80.84	**89.46**	**88.83**