MapYourCity Challenge

ai4eo.eu

AI4EO Challenges

esa

Φ-lab

The team behind the MapYourCity Challenge

esa
Φ-lab

Nicolas Longépé – Earth Observation Data Scientist @ESA Φ-Lab

NOVASPACE

Dennis Alexander Albrecht – Consultant @Novaspace

MindEarth

Mattia Marconcini – Senior Data Scientist & Founder @MindEarth

SINERGISE

Devis Peressutti – Senior Data Scientist @Sinergise/Planet

esa
Φ-lab

Nikolaos Dionelis – Research Fellow @ESA Φ-Lab

earthpulse!

Juan Pedro – Co-founder and CTO @Earth Pulse

MindEarth

Alessandra Feliciotti – Operations Manager & Project Manager @MindEarth

SINERGISE

Nika Oman Kadunc – Senior Data Scientist @Sinergise/Planet

Today’s agenda

1. Challenge introduction
2. Announcement of top-3 winners and associated short talks

1 Challenge introduction

The MapYourCity Challenge: Why building age?

1

What is building age?

- Refers to the time elapsed since a building was built
- A time capsule revealing architectural trends, construction techniques, and design philosophies

2

Why estimating it?

- Gain information about a structure’s safety and integrity
- Calculate renovation and preservation needs
- Support urban city planning

3

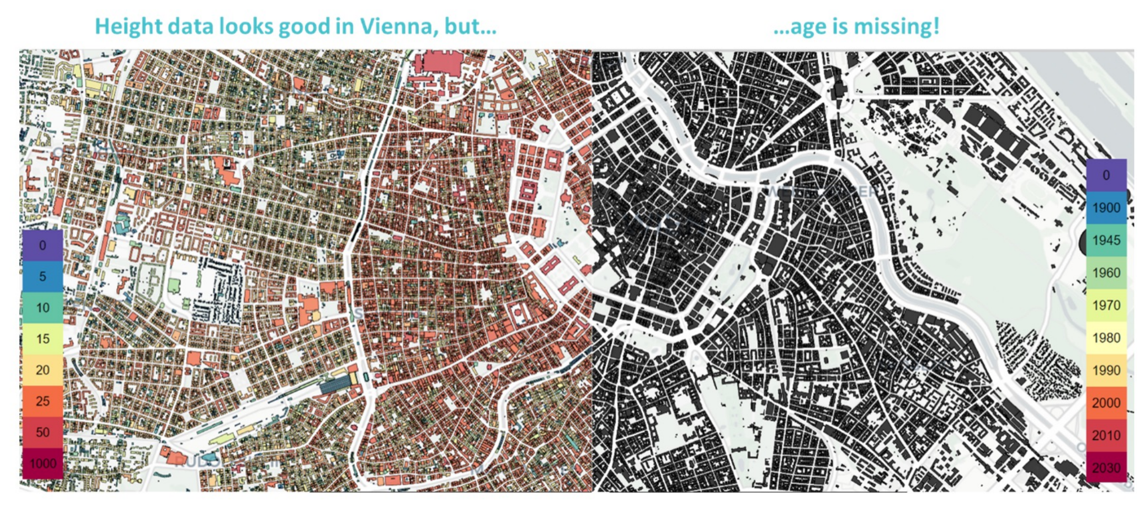
What are current problems?

- Building age monitoring is done mainly manually
- Time consuming and tedious
- Challenging to keep track due to swiftly changing cityscapes

The MapYourCity Challenge: Which task?

Merge perspectives!

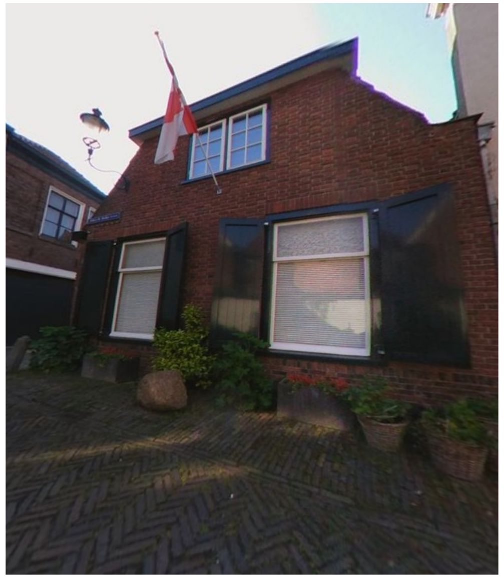
Help in defining the automation of building age detection



Height data looks good in Vienna, but...

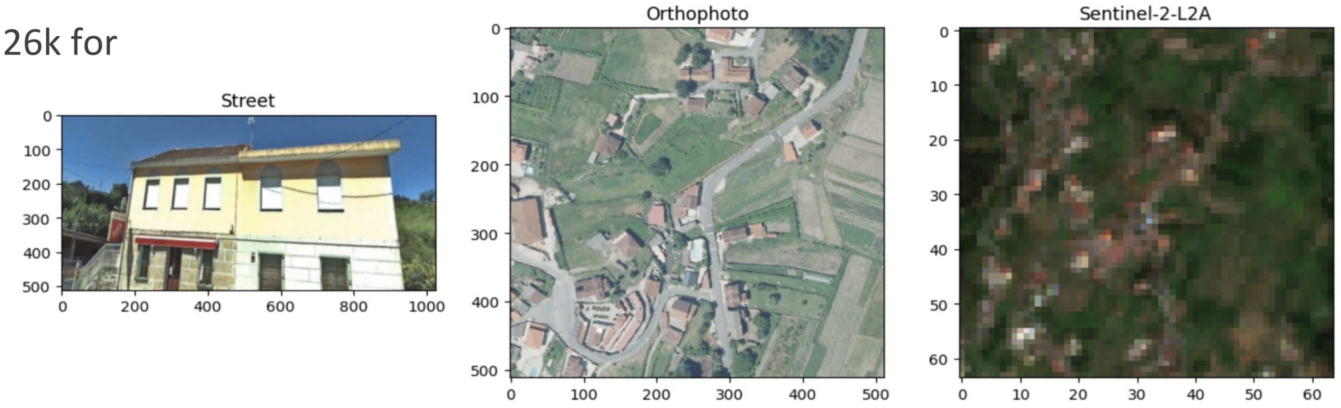
...age is missing!

Some examples …



Rules and timeline

- Challenge runtime **02 April to 14 July** 2024
- Dataset with 3 multi-modal inputs, covering 35 k anonymized location (26k for training + 9k for testing) over 5 European countries and 19 cities
  - Streetview collected by MindEarth and Mapillary
  - Aerial top-view RGB images at 50 cm resolution
  - 12-bands Sentinel-2 images.
- Classification task with 7 classes extracted from Eubucco database
  - Bef. 1930, 1930-1945, 1946 – 1960, 1961 – 1976, 1977 – 1992, 1993 – 2006, aft. 2006
- Test dataset: from 4 anonymized cities with 50% of samples with all modalities and 50% with VHR RGB + S-2 imageries
- Evaluation based on Mean Producer Accuracy (mean of the diagonal of the confusion matrix)
- 123 teams registered, 30 active teams, over 300 submissions



2500 €

1500 €

1000 €

2 Announcement of the winning teams

And the third place goes to ….

#3: Caroline Arnold DKRZ (Germany)

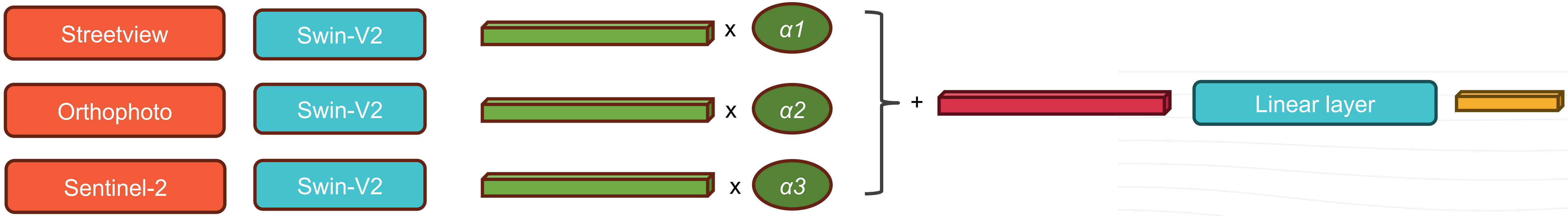
2500 €

1500 €

1000 €

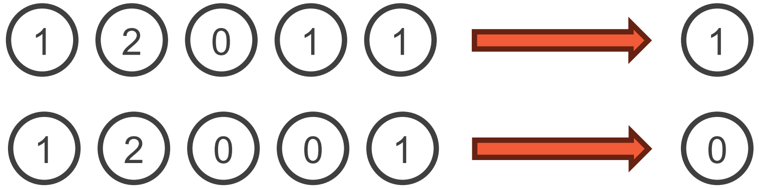
Robust Multi-Modal Model

- ✓ Each modality is encoded separately with pretrained SwinV2 vision transformer
- ✓ Calculate attention weight from embedding vector
- ✓ Weighted sum of all embedding vectors
- ✓ Classification head with 7 classes (building age)
- ✓ Works if one or two modalities are missing during inference



Building the Model

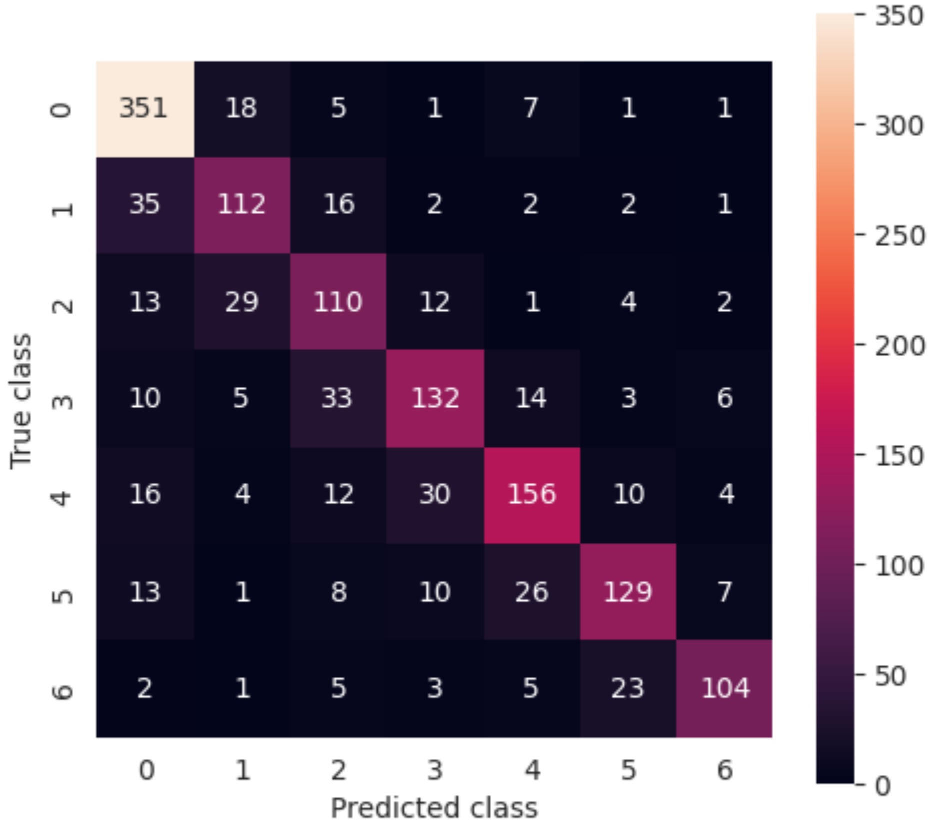
- 5-fold cross validation
  - Group by city ID
  - Stratify by class label distribution
- Inference
  - Use majority class of 5 folds
  - Tie: Use more probable class



- Train two distinct models
  - Samples with all modalities: Streetview + Orthophotos, SwinV2 base transformer
  - Samples with only top modalities: Orthophotos + Sentinel-2, SwinV2 small transformer
- Train with all data first, then finetune with country-specific samples
- Data augmentation: flips, color jitter, vegetation / building index (S-2)
- Cross Entropy Loss, Adam optimizer, lr = 10E-5, weight decay = 5E-3

Optimization and Evaluation

- Development set with country distribution matching the test set – 1000 samples
- Mean average precision (MAP)
  - ○ 0.7236 for samples with streetview
  - ○ 0.5976 for samples without streetview
- Confusion matrix
  - ○ Neighboring classes are sometimes confused
  - ○ Overall satisfying



True class \ Predicted class	0	1	2	3	4	5	6
0	351	18	5	1	7	1	1
1	35	112	16	2	2	2	1
2	13	29	110	12	1	4	2
3	10	5	33	132	14	3	6
4	16	4	12	30	156	10	4
5	13	1	8	10	26	129	7
6	2	1	5	3	5	23	104

https://github.com/crlna16/ai4eo-map-your-city

AI4EO Challenges

esa

Φ-lab

And the second place goes to ….

#2: Tran Hoang Ba, Axelspace (Japan)

2500 €

1500 €

1000 €

AI4EO Challenges

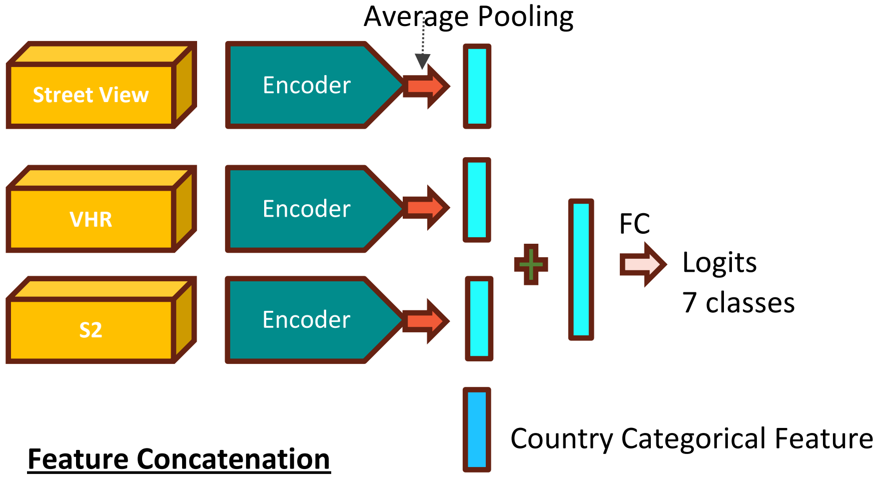
esa

Φ-lab

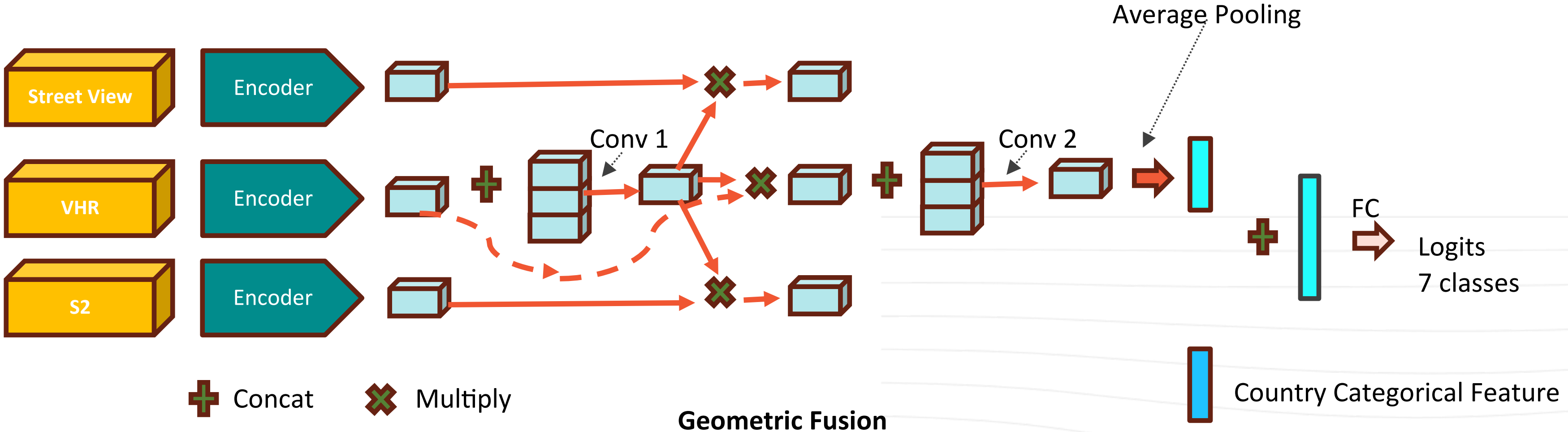
Ba Tran
Computer Vision Software Engineer
@ Axelspace

The approach

- Two types of classification model: **type I** (training and inference on full modalities), **type II** (training on full modalities but inference on only 2 top view modalities)
- Country categorical feature (one-hot representation) was used in both **type I**, and **type II**
- For **type I**, 3 inputs from 3 modalities are fetched into 3 encoders of the same feature extractor models and fused (late fusion) by 2 methods:
  - Feature Concatenation: Features after average pooling are concatenated together with country categorical feature, and then fetched into the final FC layer
  - Geometric Fusion: Features before average pooling are fused by method proposed in [1], then fetched into average pooling and later concatenated with country categorical feature.
- 4 models were trained for **type I**, and ensemble of 4 models was used as the final model
  - Encoders: efficientnetv2_s, mobilevitv2_150, efficientnetv2_b3
  - 2 models used Feature Concatenation, 2 used Geometric Fusion approach



Feature Concatenation

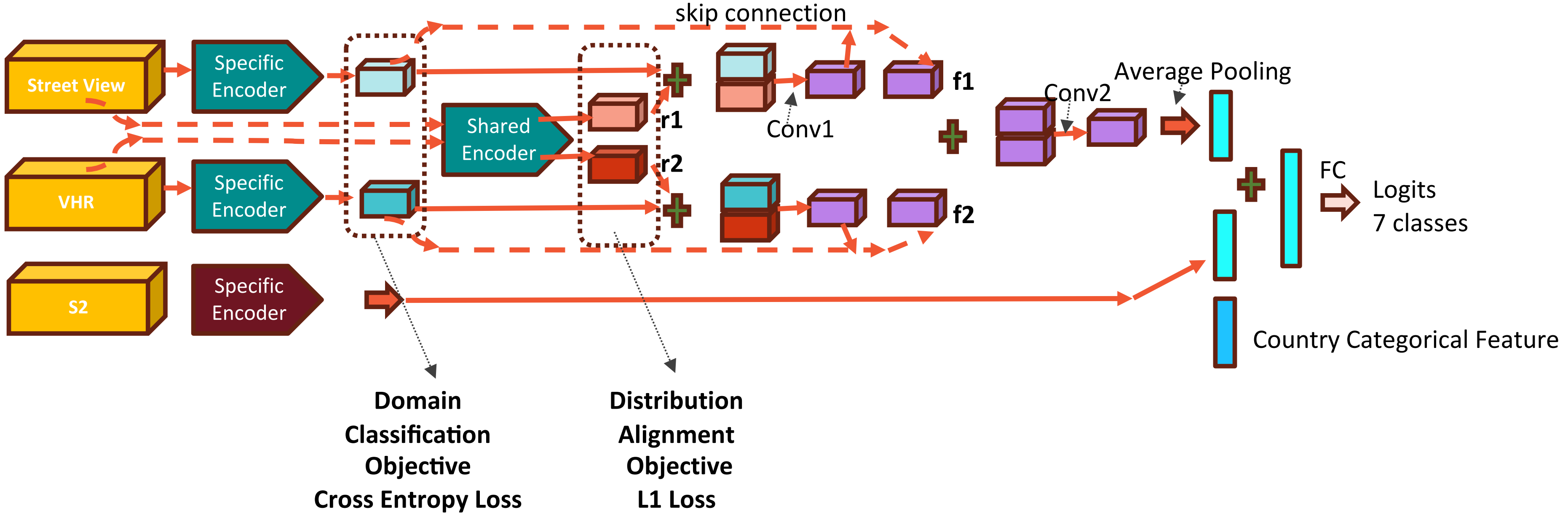


Geometric Fusion

[1] Chen, Boan, et al. "Multi-modal fusion of satellite and street-view images for urban village classification based on a dual-branch deep neural network." International Journal of Applied Earth Observation and Geoinformation 109 (2022): 102794.

Approach for type II

- For type II, two important techniques were used to address the problem of missing modality
  - Input Dropout: During training, street-view images were randomly replaced with all zero images with p = 0.5
  - Shared-Specific Feature Modelling: each modality input was assigned with a specific separate encoder (3 specific encoders), but street-view and VHR modality inputs were further assigned a shared encoder (1 share encoder). Street-view feature and VHR feature extracted by shared encoder were aligned by Distribution Alignment Objective (L1 Loss for pairwise feature similarity) so that during inference with missing modality, VHR feature extracted by shared encoder would be used as street-view feature from specific encoder. Street-view and VHR modality used the same type of encoder for shared and specific encoder, while S2 modality used a different type (a shallower one).
- Only feature concatenation approach was used for type II
- 2 models were trained for type II:
  - Model 1: efficientnetv2_b1 was used for street-view and VHR encoder (shared and specific), efficientnetv2_b0 was used for S2.
  - Model 2: efficientnetv2_b2 was used for street-view and VHR encoder (shared and specific), efficientnetv2_b0 was used for S2.

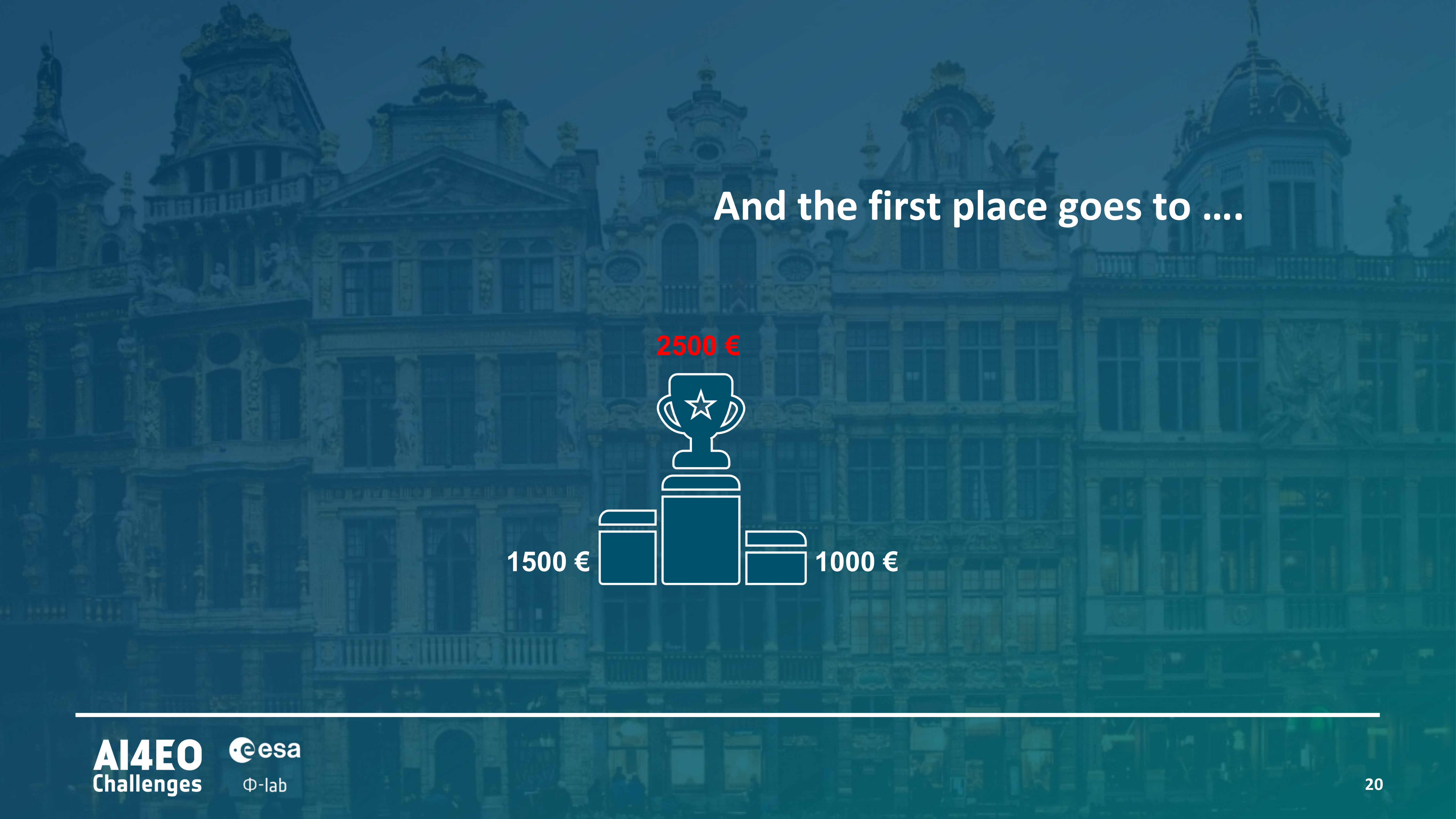


At Inference, with the missing of Street-view modality, f1 = r2

[2] Wang, Hu, et al. "Multi-modal learning with missing modality via shared-specific feature modelling." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

Optimization Parameters/Techniques

- pytorch, pytorch lightning framework, timm library for collection of encoders
- Single GPU (GeoForce GTX 1080Ti)
- Batch size of 5 (~6) with accumulated grad batches of 8 -> Effective batch sizes of 40 (~48)
- Optimizer: AdamW with lr of 1e-4, weight decay of 1e-3
- Scheduler: CosineAnnealingLR with eta min of 5e-5
- Epochs: 30 epochs without Early Stopping
- Loss: Focal Loss with Label Smoothing (smoothing coefficient of 0.1)
- Train dataset was divided into stratified 10 folds. Each model would leave out 1 different fold for validation and remaining 9 folds for training.
- Data processing and augmentation:
  - Street-view and VHR inputs were resized to 512x512 and normalized to [0,1]. S2 inputs were clipped by maximum values of 10000, and then divided by 10000 to be normalized to [0,1]
  - Albumentations library
  - Street-view input was augmented with: ShiftScaleRotate(-5,5), ColorJitter, AdvancedBlur, HorizontalFlip, CoarseDropout, GridDropout, Spatter
  - VHR: ShiftScaleRotate(-180,180), ColorJitter, AdvancedBlur, Flip, CoarseDropout
  - S2: Rotation(-180,180), Flip
- Test time Augmentation:
  - **Type I**: Horizontal Flip for Street-view, (Horizontal Flip, Vertical Flip, both Horizontal Vertical Flip) for VHR -> 2x4 = 8 TTA patterns -> 4 models x 8 TTA patterns = total of 32 patterns for ensembling
  - **Type II**: (Horizontal Flip, Vertical Flip, both Horizontal Vertical Flip) for VHR -> 2 models x 4 TTA patterns = total of 8 patterns for ensembling



And the first place goes to ….

2500 €

1500 €

1000 €

#1: Eric Park and Hagai Raja Sinulingga, TelePIX (South Korea)

2500 €

with 2 different solutions sharing the 1st prize

1500 €

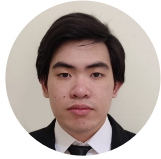
1000 €

AI4EO Challenges

esa Φ-lab

How We Approach the 2nd Best Solution

Hagai Raja Sinulingga
AI Engineer@TelePIX



Steve Andreas Immanuel
AI Engineer@TelePIX

We tested combinations of SOTA methods which summarized as follow:
("used" denotes the approach integrated into the final submission; the rest were attempted but didn't pan out)

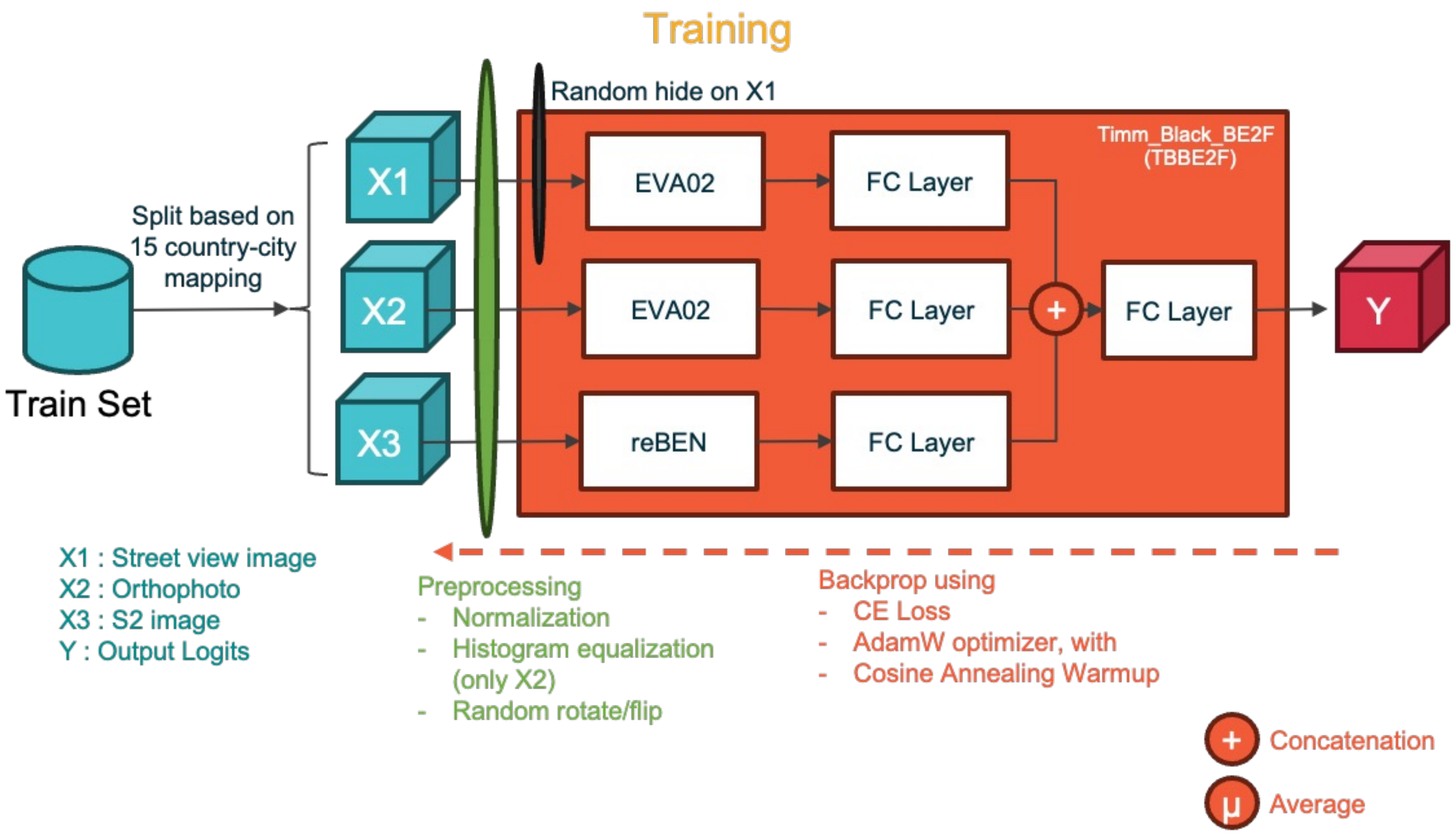
- Data Splitting for result validation:
  - Split train set into train and val with ratio 80:20 (for all dataset and split per country)
  - K-Fold but with city as the fold for each country (used)
- Preprocessing: Normalization (used), Histogram equalization (used), Center-crop, Random Rotate/Flip (used), random hide Street-view (used), Generative top-view image using GAN/Diffusion model, Segmentation
- Encoder selection: SeResNext, MobileNet, EfficientNet, ViT (from CLIP, Model-Soup, RVSA, Satlas, SSMAE, Alpha-CLIP), EVA02 (used), reBEN (used), and Hiera
- Data specific pre-training: Contrastive Learning, Masked Image Modelling
- Feature fusion: Input Level, Feature Level, Prediction Level (used)
- Loss function: CrossEntropyLoss (used), OrdinalCELoss, LabelSmoothing, WeightedCELoss, FocalLoss
- Prediction approach: Classification (used), Regression
- Ensemble technique: Max Confidence, Mean Confidence (used), Confidence Thresholding, Multi-stage Ensemble
- Hyper-parameter tuning: Bayesian optimization using Optuna (learning rate, batch size, lr decay factor, etc) (used)

What We Learn from Our Experiments

These are the key factors to increase performance:

- Powerful encoders: bigger is better and it is important to choose the right pre-trained model
- Random hide street-view: during training we randomly (p: 0.5) turn the street-view image to completely black (0 value) so the model could learn the significant features from the Orthophoto and S2
- Split based on country: reduce the variability since city in the same country looks similar but different with city on another country
- Ensemble technique: increase robustness of the final prediction
- Preprocessing: increase robustness and training stability

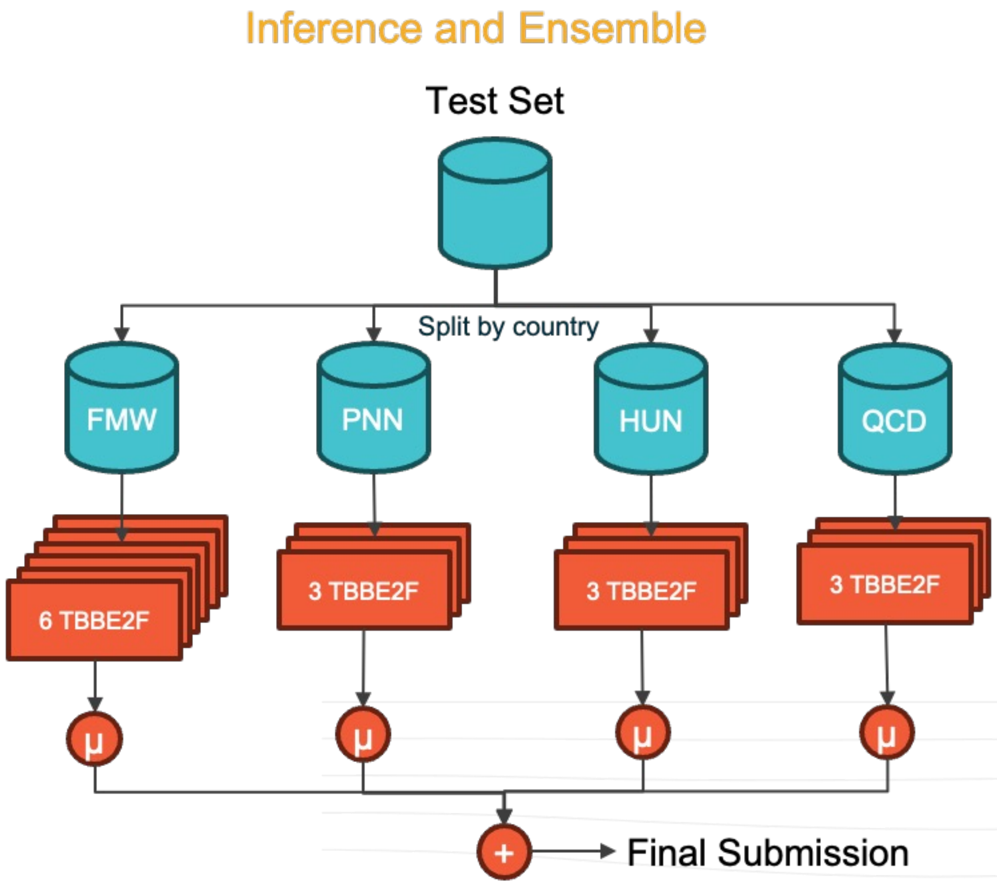
Overall Flow of Our Final Submission Experiment



X1 : Street view image
X2 : Orthophoto
X3 : S2 image
Y : Output Logits

Preprocessing
- Normalization
- Histogram equalization (only X2)
- Random rotate/flip

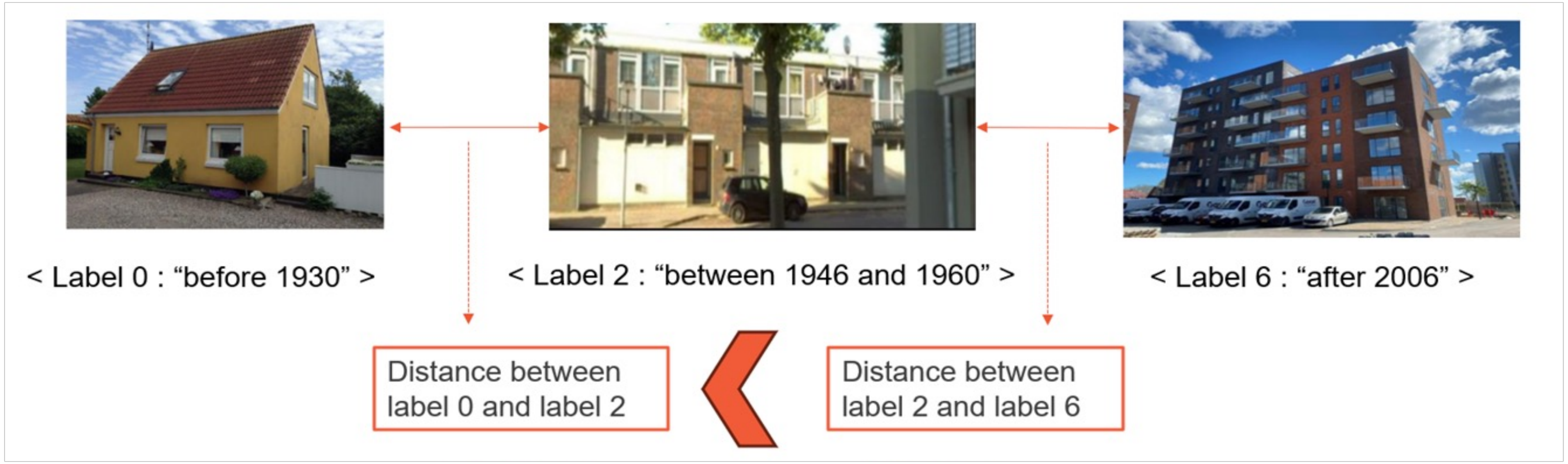
Backprop using
- CE Loss
- AdamW optimizer, with
- Cosine Annealing Warmup



The approach for 1st Best Solution

JaeWan Park (Eric Park)
AI Engineer@TelePIX

Label Correlation



< Label 0 : “before 1930” >

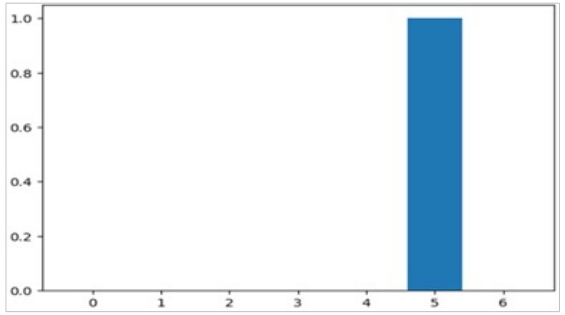
< Label 2 : “between 1946 and 1960” >

< Label 6 : “after 2006” >

Distance between label 0 and label 2

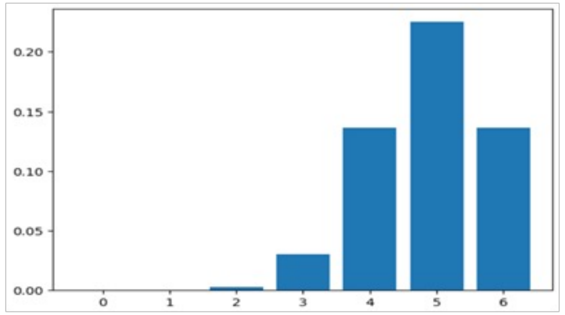
Distance between label 2 and label 6

- Reflecting the correlation according to the distances held by the labels in the learning process was expected to be a key approach for the solution.



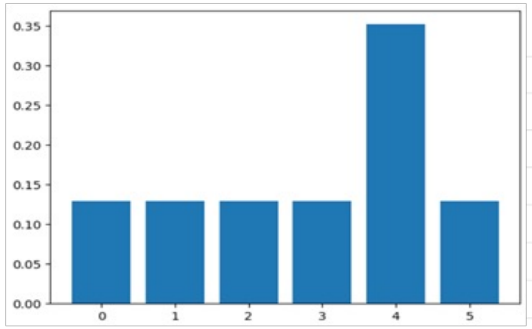
< Original Hard Label >

PB score increased



< Gaussian Distributed Soft Label >

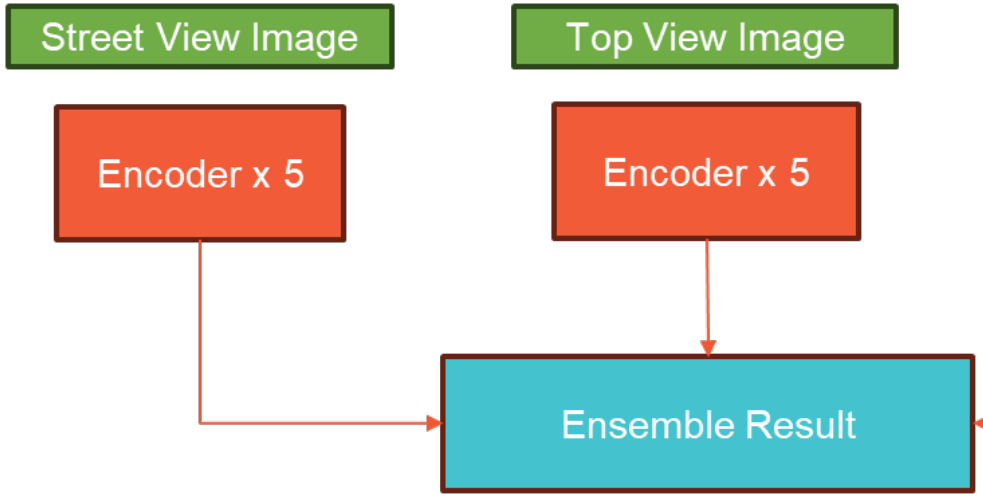
PB score increased



< Label Smoothed Soft Label >

The approach for 1st Best Solution

Architecture



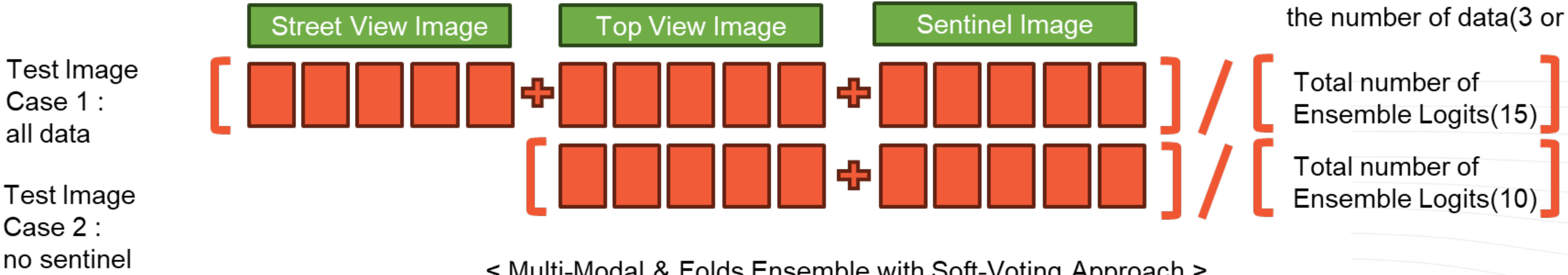
Design Points

1. **Simpicity & Flexibility** : Since experiments must be conducted in a cross-validation structure, a structure that allows for repeated experiments and flexible tests.

2. **Ensemble Intergration** : The test-dataset cases are divided into two, the ensemble layer is removed so that it can be used for both cases.

Ensemble : Multi Data & Folds Soft-Voting

Total number of logits =
the number of data(3 or 2) * the number of folds(5)


Street View Image
Top View Image
Sentinel Image
Test Image Case 1 : all data
Total number of Ensemble Logits(15)
Test Image Case 2 : no sentinel
Total number of Ensemble Logits(10)

< Multi-Modal & Folds Ensemble with Soft-Voting Approach >

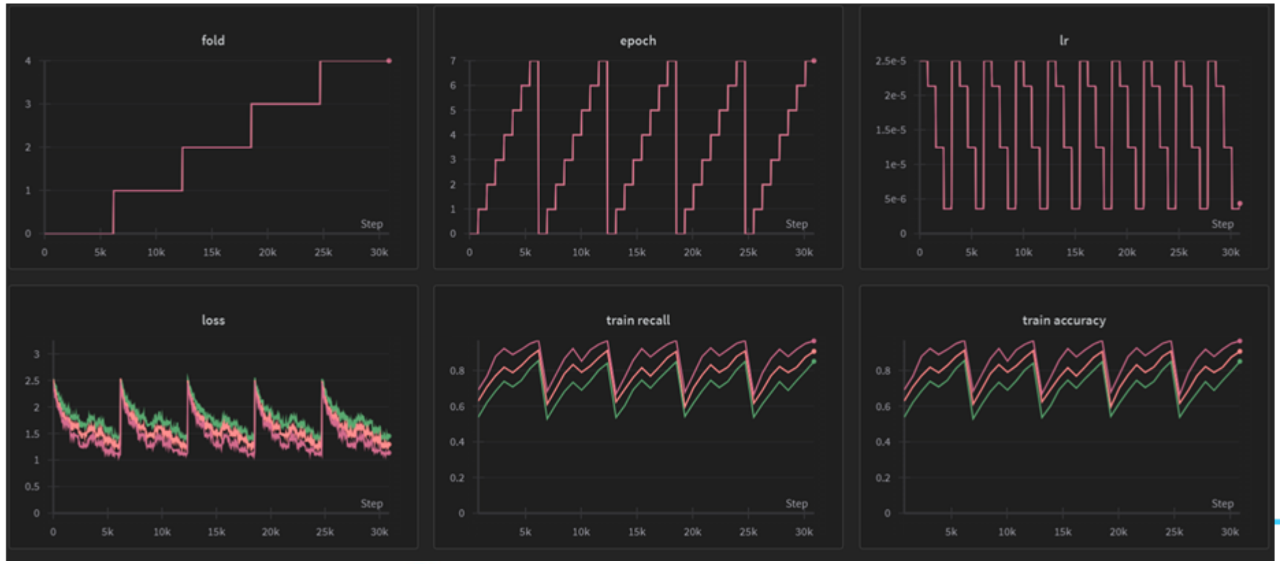
The approach for 1st Best Solution

Model Specification

MODEL	eva02_large_patch14_448.mim_m38m_ft_in22k_in1k
Task	Classification
Stratified K-Folds	5
Loss	Cross-Entropy
Label_Smoothing	0.3
Optimizer	AdamW
Scheduler	CosineAnnealingWarmRestarts
Learning Rate	2.5e-05
Image Resize	448 Max & 2:1 ratio(Street Veiw) 448(Top-view, Sentinel)
Image Augmentation	trivial
Image Interpolation	bicubic

- For image classification, I used EVA, which recently achieved SOTA (State Of The Art) on ImageNet.
- Since there are various types of labels, I used Stratified K-Folds to train five models (FOLDS) for each type of data (Street view, Top view, Sentinel).
- For the LOSS, I applied Label Smoothing (0.3) to Cross Entropy, effectively training it in a manner close to KL Divergence.
- I used AdamW as the optimizer and employed Cosine Annealing starting from a learning rate of 2.5e-05.
- For augmentation, I initially used griddropout, horizontalflip, Blur, and RandAugment, but to reduce overhead from parameter search and experiment more efficiently, I used PyTorch's trivial augment.

Experiment Logs



- Logs for individual models were recorded so that each encoder could achieve the best performance.
- The code was composed as argumnet-based and the experiment was conducted so that the variables in the experimental results were well stored and the experimental results could be reproduced.
- Combination and optimization were studied to achieve the best results while observing the learning curve and metrics.

Conclusion

Conclusion

- Street-view pictures convey many information about building, but they can hardly scale to regional/national levels
- Estimation of building age from space is feasible ! MPA_top-view ~60% over 7 classes
- When street view is available, MPA_all_modalities ~73%

- Perspectives:
  - Benefit of Sentinel-2 compared to VHR orthophoto?
  - Understanding of the importance of street-view images when training, even if this modality is not present during inference?
  - Generalization capabilities over unknown cities / countries?
  - How to improve performance? Add contextual information….

AI4EO

A bridge between EO and AI

esa

Φ-lab

NOVASPACE

MindEarth

Thank you !

More questions? Reach out:

Nicolas.Longepe@esa.int