Large-scale semantic classification:
Outcome of the first year of Inria aerial image labeling benchmark

B. Huang[1], K. Lu[2], N. Audebert[3,4], A. Khalel[5], Y. Tarabalka[6], J. Malof[1], A. Boulch[3], B. Le Saux[3], L. Collins[1], K. Bradbury[1], S. Lefèvre[4], M. El-Saban[5]

[1] Duke University; [2] NUS; [3] ONERA; [4] Univ. Bretagne-Sud, IRISA; [5] Raisa energy; [6] UCA, Inria.

⇒ `project.inria.fr/aerialimagelabeling/`

Inria Benchmark dataset and statistics

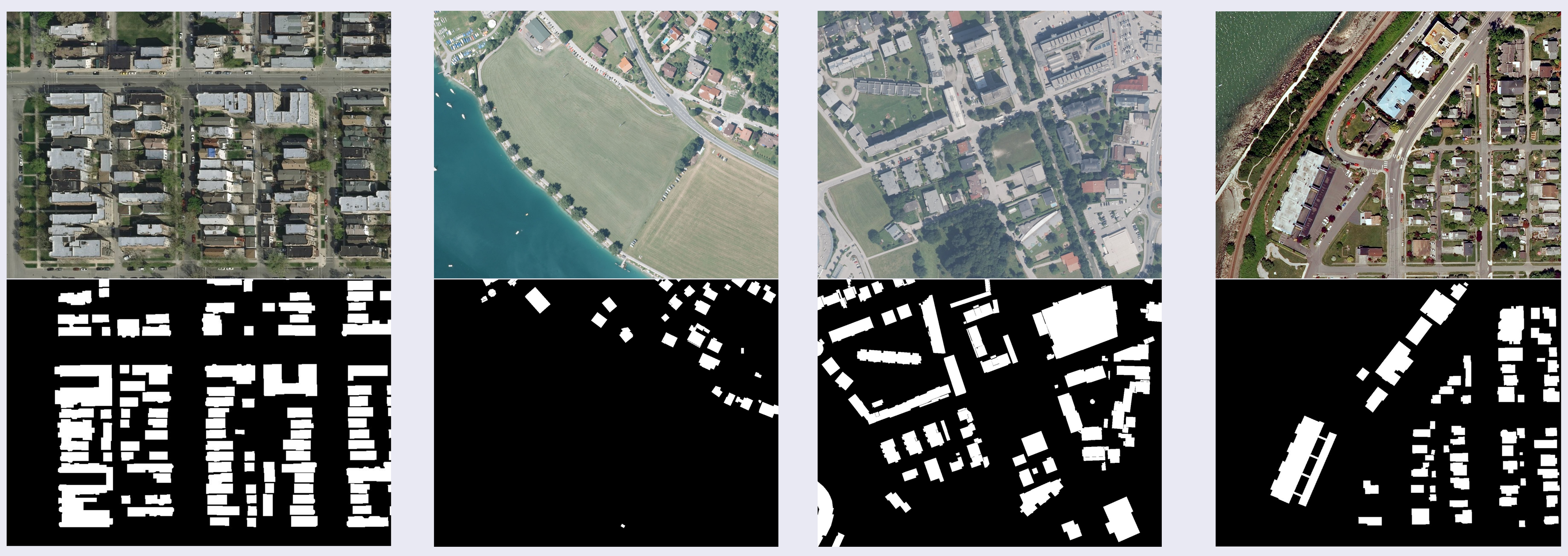
- Problem: Large-scale pixelwise semantic labeling of aerial images
- Two semantic classes: *building* and *not building* (ref. data by rasterizing building footprints)
- Different cities in train and test subsets
  ⇒ E.g., we should classify San Francisco without "seeing" it before
- European/American & high-/low-density urban landscapes in both subsets
- 0.3 m spatial resolution, 3 color bands, 360 tiles (1500^2 px each)

Statistics:

Train	Tiles	Total area	Test	Tiles	Total area
Austin, TX	36	81 km^2	Bellingham, WA	36	81 km^2
Chicago, IL	36	81 km^2	San Francisco, CA	36	81 km^2
Kitsap County, WA	36	81 km^2	Bloomington, IN	36	81 km^2
Vienna, Austria	36	81 km^2	Innsbruck, Austria	36	81 km^2
West Tyrol, Austria	36	81 km^2	East Tyrol, Austria	36	81 km^2
Total	180	**405 km^2**	Total	180	**405 km^2**

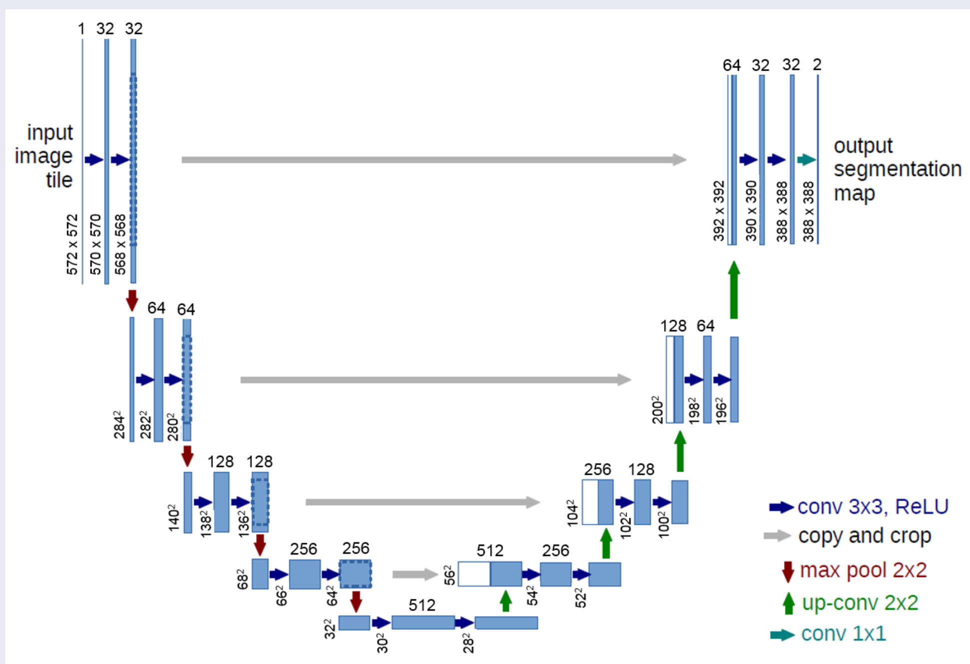
- During the first year after the benchmark release:
  - $>$ 800 downloads from all continents, from public & private institutes
  - 16 submissions with the results on the test set
  - **Which method is the best?** Four winning methods are detailed here

Close-ups of training and test sets



Chicago (train) West Tyrol (train) East Tyrol (test) Bellingham (test)

1st place: U-net with novel training/test strategy (AMLL, Duke Univ.)



- Original **U-net** architecture [1] with half as many filters at each layer
- **Training strategy:**
  - From training dataset: tiles 6-36 from each city for training, the rest for validation
  - Extract 572 × 572 input patches on a uniform grid, with 92 pixels of overlap between neighboring patches
  - Minibatch of 5 randomly selected patches
  - Data augmentation: vertical/horizontal flips and orthogonal rotations
  - Cross-entropy objective function
  - Adam optimizer: initial learning rate of $1e-3$, a momentum of 0.9
  - 100 epochs, each epoch processes 8000 minibatches
- **Label inference:**
  - U-net predicts poorly at the edge of its output
  - To mitigate this problem ⇒ use 2636 × 2636 input patches* during label inference

[1] Ronneberger et al., "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.

* Maximum size supported by 1080 Ti GPU.

2nd place: Dual-resolution U-net (NUS)

- **U-net** architecture with a pair of dual-resolution images as input
  - Crop high-resolution 384 × 384 patches
  - Crop 768 × 768 patches with the same center and downsample them to 384 × 384 patches
  - Features from high & low resolution patches are extracted by U-net
  - Result = weighted sum of dual-resolution score maps
- **Loss function** = combination of sigmoid cross-entropy (*sigmCE*) and a soft Jaccard loss [2]:

$$L_{NUS} = L_{sigmCE} - \log l_{soft-IOU}$$

- **Implementation details:**
  - Channels of the modified U-net are: 32, 64, 128, 128, 256, 128, 128, 64, 32
  - Data augmentation: vertical/horizontal flips
  - Adam optimizer: initial learning rate of $1e-3$, a momentum of 0.9, "poly" learning rate policy
  - 30 epochs

[2] Mattyus et al., "Deeproadmapper: Extracting road topology from aerial images," in *ICCV*, 2017.

3rd place: Signed distance transform regression (ONERA)

- Standard **SegNet** architecture with pre-trained VGG-16 weights
  - 384 × 384 patches, stochastic gradient descent optimizer
- Include spatial context in optimization
  ⇒ Add a regularization loss computed on the Euclidean signed distance transform (SDT) [3]:

$$L_{ONERA} = NLLLoss(Z_{seg}, Y_{seg}) + \lambda L1(Z_{dist}, Y_{dist}),$$

where $NLLLoss$ = negative log-likelihood loss function, $L1$ = $L1$ penalty on SDT distances, λ = hyper-parameter

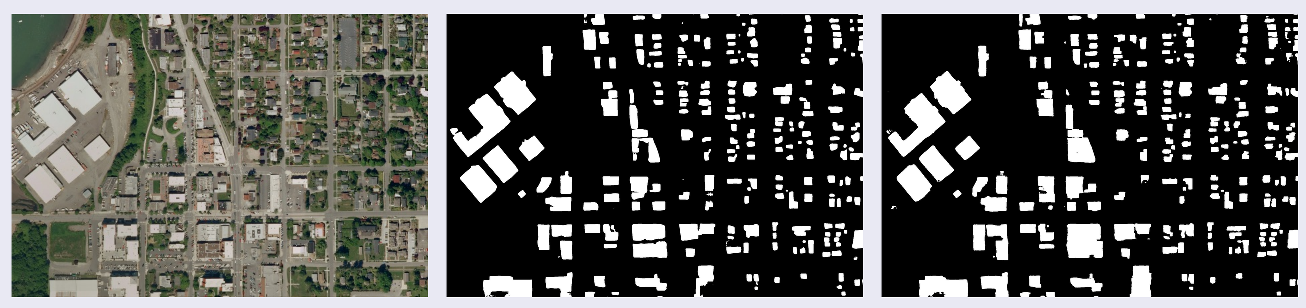
[3] Ye, "The signed Euclidean distance transform and its applciations," in *ICPR*, 1988.

4th place: Stacked U-nets (Raisa Energy)

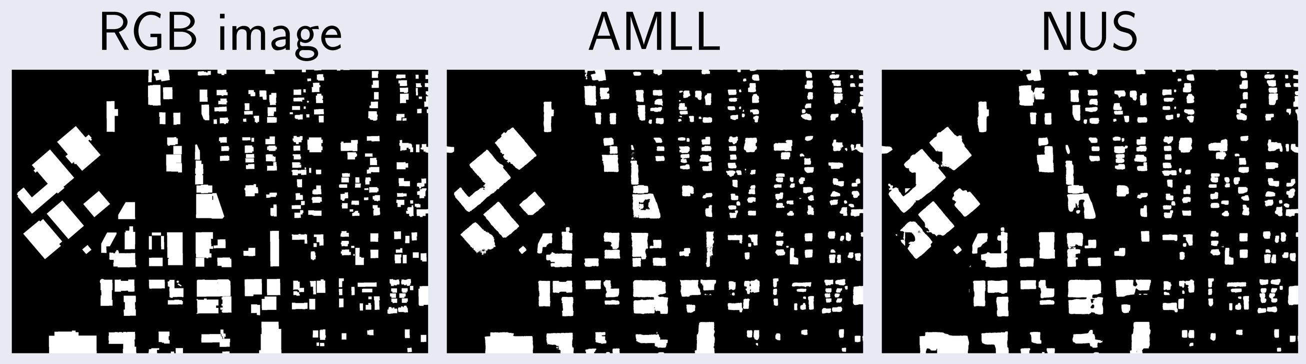
- Stack of two **U-nets** arranged end-to-end
  - Second net enhances predictions of the first net
- **Loss function** combines binary cross entropy and a differential form of Intersection over union (IoU) [2]

Experimental results

- `https://project.inria.fr/aerialimagelabeling/leaderboard/`



RGB image AMLL NUS



Ground truth Onera Raisa Energy

	Belling.	Bloom.	Inns.	S. Francisco	East Tyrol	Overall
AMLL	67.14	65.43	72.27	**75.72**	74.67	**72.55**
NUS	**70.74**	66.06	**73.17**	73.57	**76.06**	72.45
ONERA	68.92	**68.12**	71.87	71.17	74.75	71.02
RAISA	68.73	60.83	70.07	70.64	74.76	69.57

Numerical evaluation on test set (IoU scores)

Concluding remarks

- Active exploitation of the benchmark since its release
- U-net architecture has shown the highest performance
- Good choice of loss function & training strategy boosts results
- Published on Nov '16, $>$ 1500 downloads as of June '18, $>$ 50 submissions to contest
- **Contest still open** to submit results to benchmark!