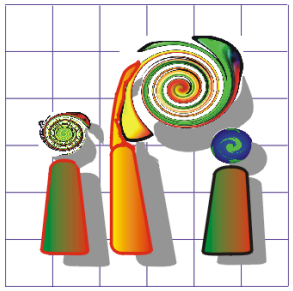
BUDAPEST GEOSUMMIT 2025

Advancing Geospatial Science and Engineering

MŰEGYETEM 1782

AI in Photogrammetry and Remote Sensing



Christian Heipke
IPI - Institute for Photogrammetry and GeoInformation
Leibniz Universität Hannover

Content

- **Introduction**
  - **what are we talking about?**
  - **photogrammetry or computer vision?**
- **Some AI and deep learning basics**
- **Examples of (more traditional) photogram-metry and remote sensing applications**
- **Conclusions**

Principles of photogrammetry and remote sensing

- determination of characteristics of the EMS radiation of a given wavelength, captured as a 2D image as reflected or emitted from some surface
  - energy, phase, polarisation, signal form, travel time
- derivation of object and surface characteristics from these measurements
  - geometry: position, size, shape
  - radiometry: rendered scene
  - semantics: object-ID, attributes

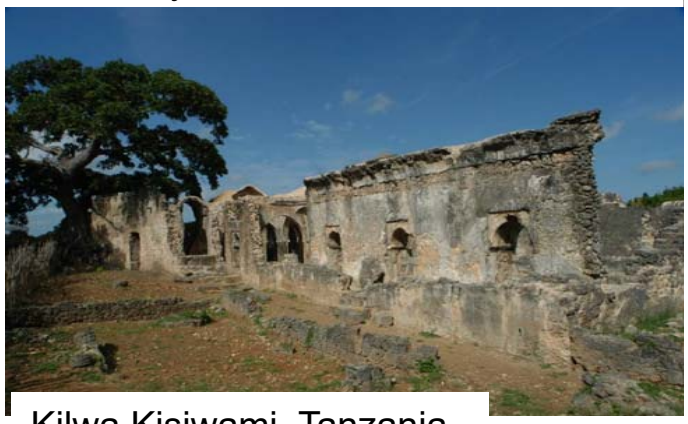
Documentation of world cultural heritage

Rationale for the work: „nothing is forever, time takes its toll”

Notre Dame de Paris, 2019

Bamiyan Buddha, before and after

Gate of All Nations, Persepolis



Kilwa Kisiwami, Tanzania

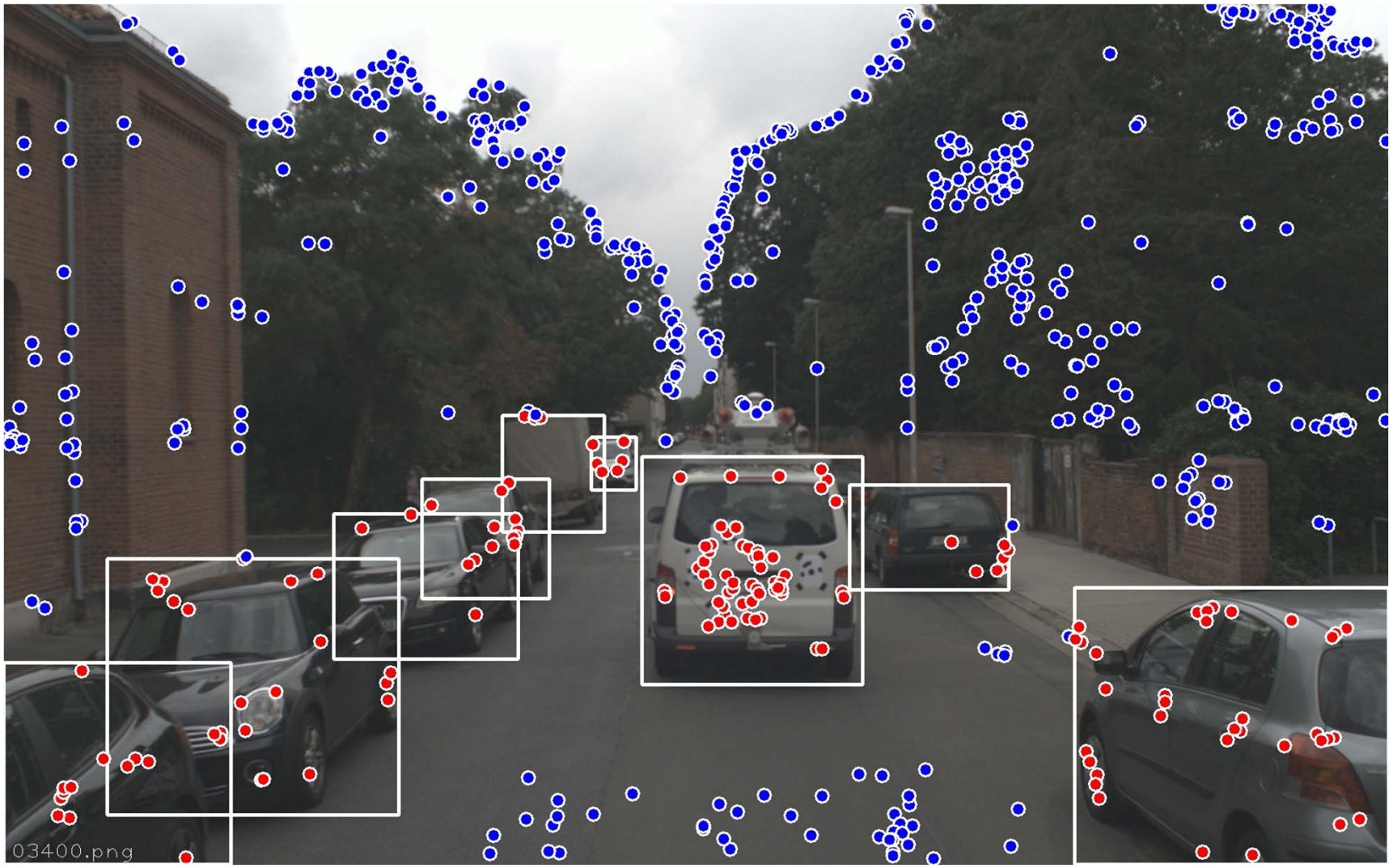
Sports

Real-time vision would have been useful here …

Referee Jorge Larrionda, Uruguay (June 27, 2010): no goal

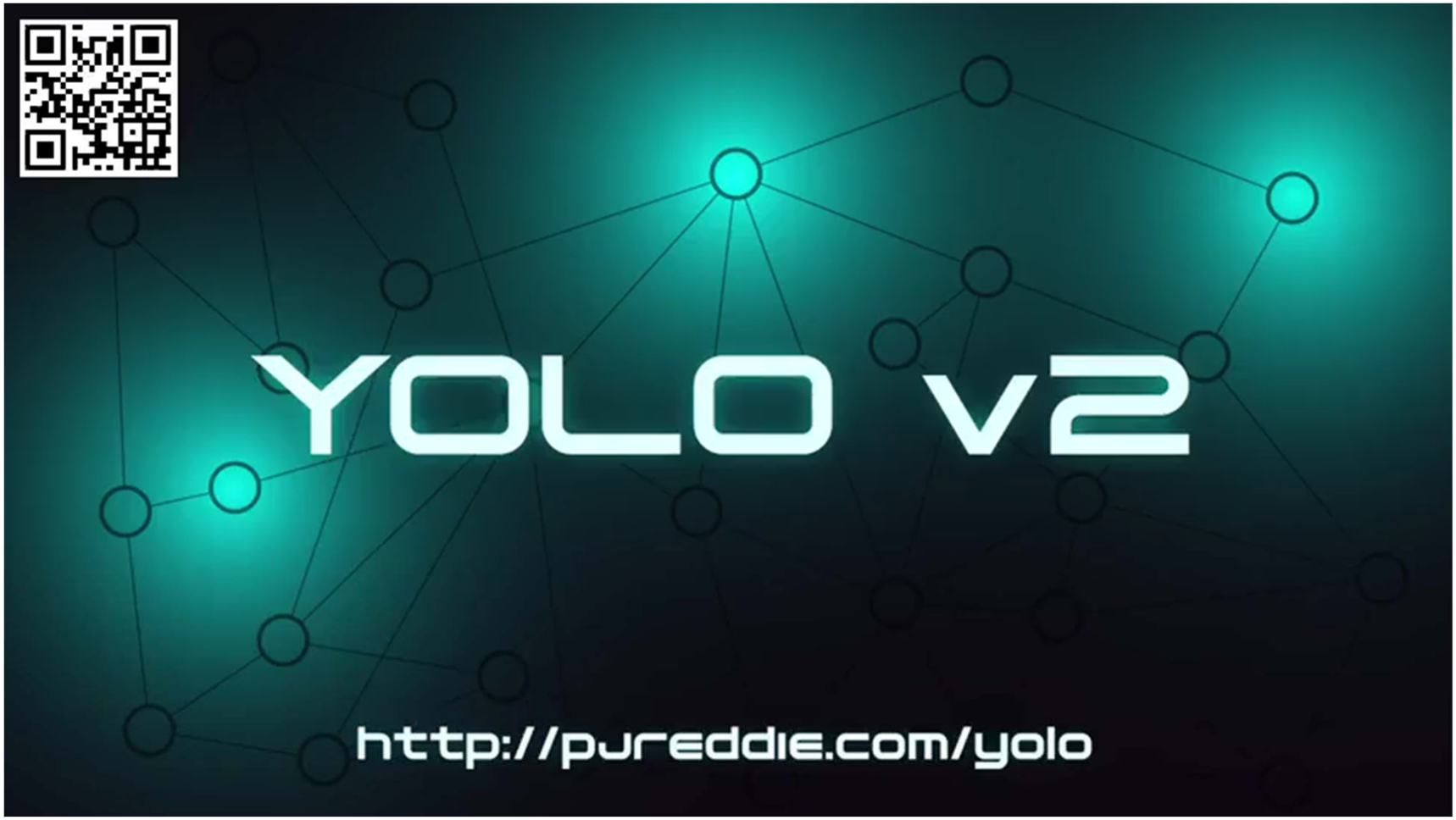
https://www.goal.com/en-gb/news/2890/world-cup-2010/2010/06/29/2001679/referee-jorge-larrionda-dropped-from-world-cup-2010

Separation between static and moving scene parts



Trusheim P., Mehltretter M., Rottensteiner F., Heipke C., 2024: Cooperative Image Orientation with Dynamic Objects. PFG (2024), 461-481, doi.org/10.1007/s41064-024-00296-w.

Multiple Object Extraction, real-time



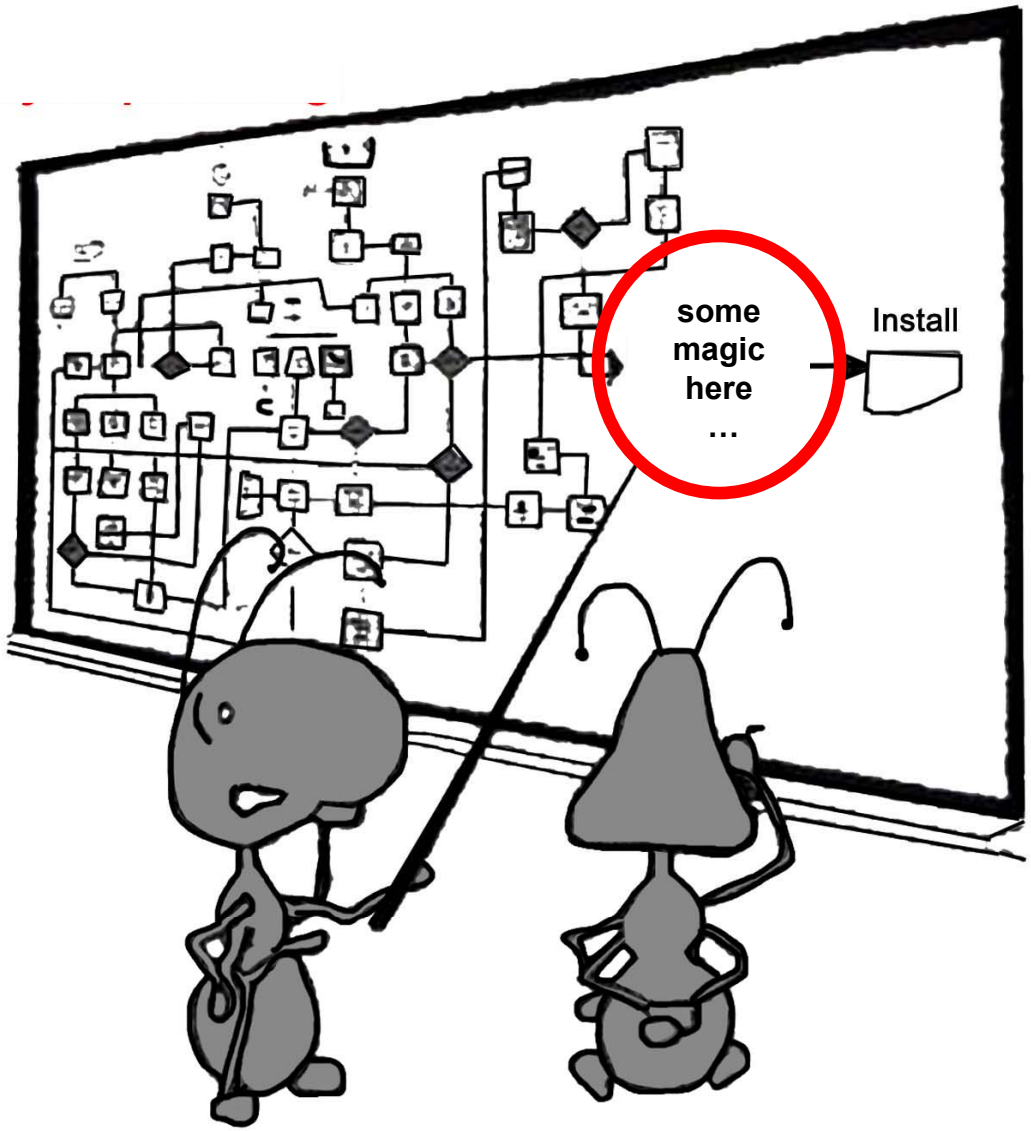
YOLO: You Only Look Once (version 2) – a Convolutional Neural Network (CNN) for object detection. J. Redmon, S. Divvalay, R. Girshick, A. Farhadiy, Uni. Washington, 2016

Multiple Object Tracking, articulated objects

https://github.com/CMU-Perceptual-Computing-Lab/openpose
https://medium.com/@rotemsha/how-does-openpose-actually-work-11a29872f30e

Autonomous driving (of course, real time)

Yi Zhu et al., 2019. Improving Semantic Sgmentation via Video Propagation and Label Relaxation (CVPR 2019). https://www.youtube.com/watch?v=ic0yCYro9H1kM

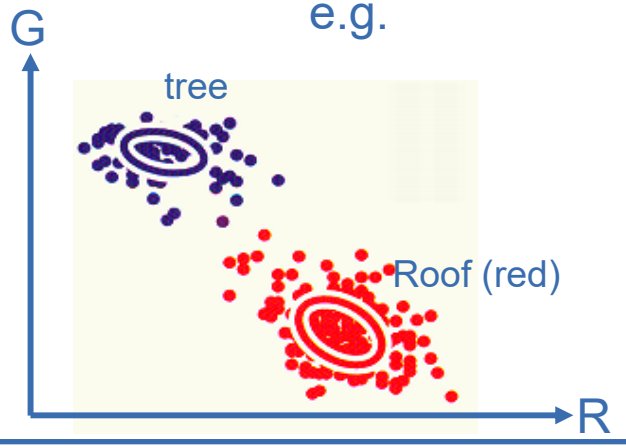


Very good work!
But don‘t you think we should have a few more details here?

Some AI and deep learning basics

Machine learning / data science / big data

- Object knowledge coded in examples → training data
- Analysis based on similarity of features in feature space

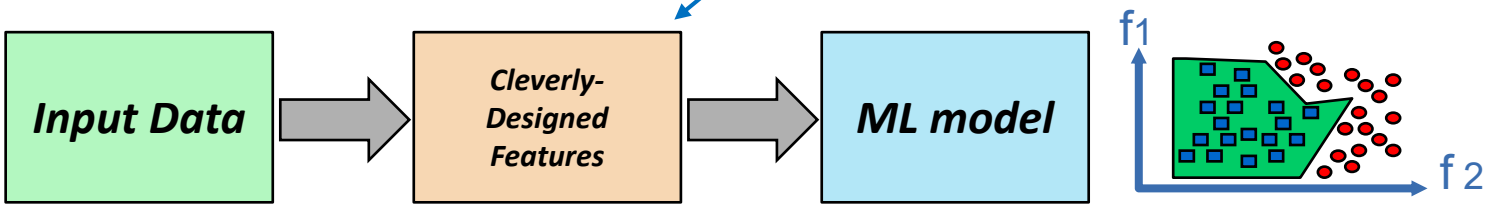


Deep learning - what is it, what is new?

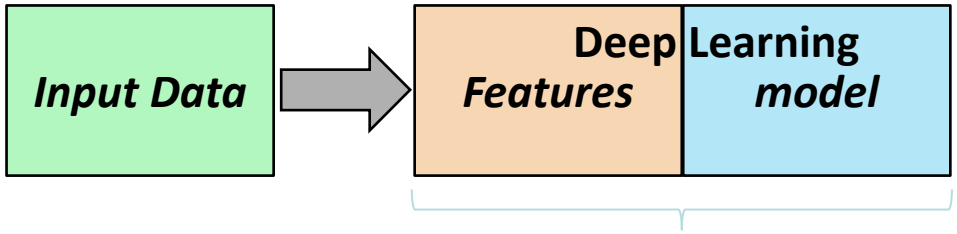
… "learning" an input-output mapping **from examples** by machines (classification, regression, …)

- traditional model:

most of the "heavy lifting" here, final performance only as good as feature set



- deep learning:

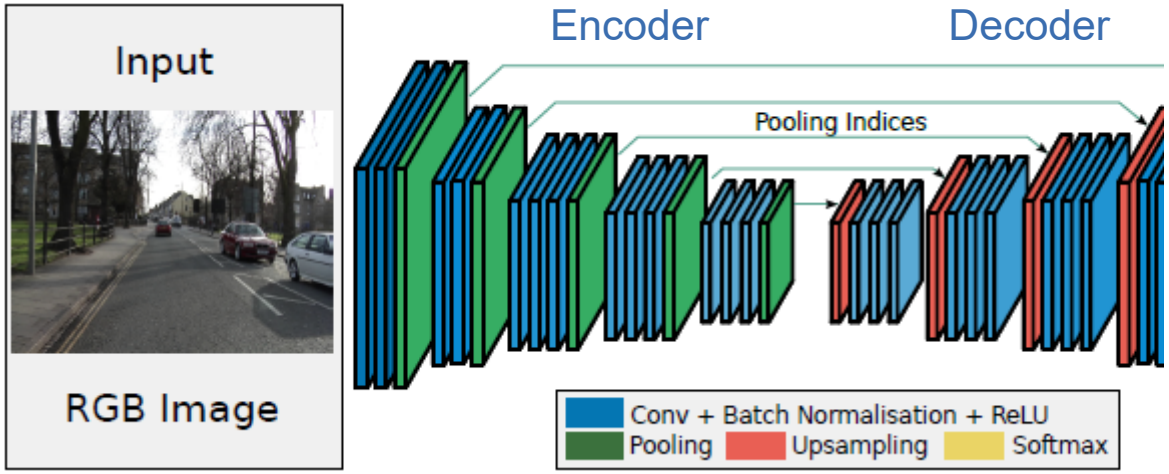


features and model learned together, mutually reinforcing each other

adapted from Michele Catasta

Different network architectures

- **Encoder-decoder network** [Ronneberger et al.,2015]:
  - field of view (“perceptive field”) increases with no. of layers
  - **skip connections** to preserve object boundaries
    - good delineation accuracy
  - standard for many RS applications (**pixel classification**)



© [Badrinarayanan et al., 2017]

Many more:

- 3D CNN
- GAN
- Residual Networks
- Trinet
- Graph NNs
- …

Ronneberger, O., Fischer, P., Brox, T., 2015: U-Net: Convolutional networks for biomedical image segmentation. *Proc. MICCAI*, Springer, LNCS, Vol. 9351, pp. 234-241.

Different network architectures

- **Attention is all you need** [Vaswani et al.,2017]:
  - based on “attention scores”, used extensively in NLP
  - sequential processing of so called “tokens” (e.g. words)

low att.

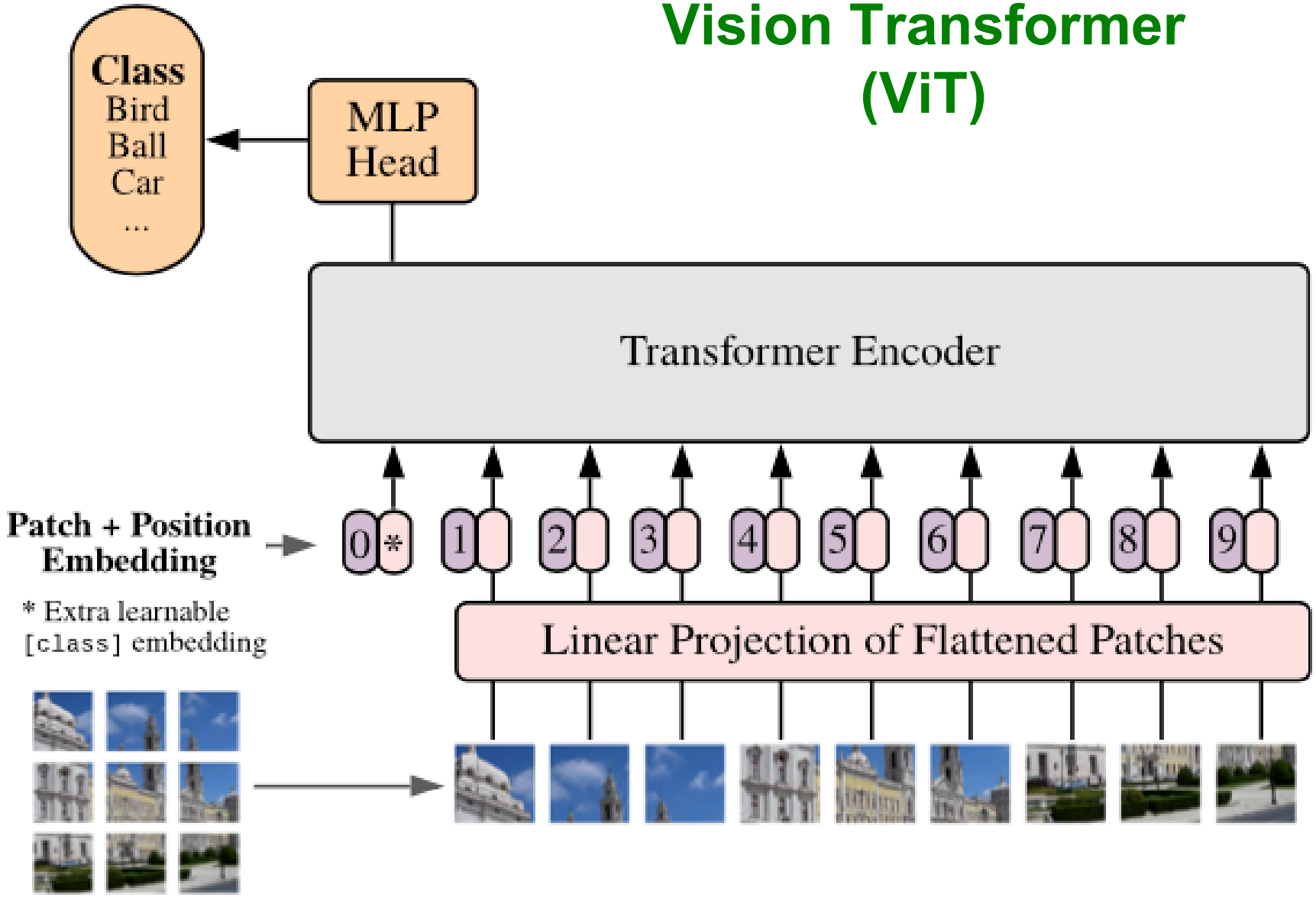
“I want a glass of beer”

high attention

  - attention is computed based on correlation of local features (derived from the tokens, ideally **ALL** of them – comp. burden!)
  - **more global context**, if considered useful (“high attention”)
  - for images: words -> pixels / patches

Vaswani, A. et al., 2017. Attention is all you need. I. Guyon et al. (eds.), Advances in Neural Information Processing Systems, 30, Curran Associates, Inc.

Vision Transformer (ViT)

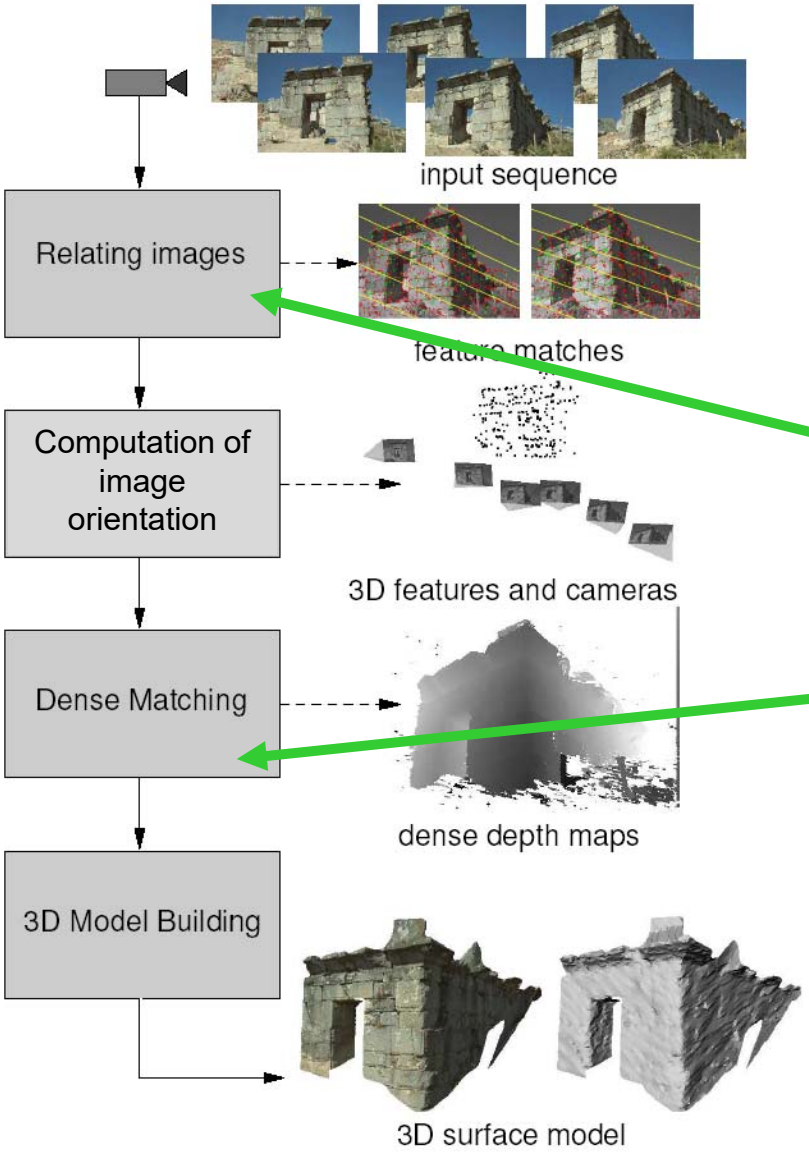


Dosovitskiy A, Beyer L, Kolesnikov A, et al., (2021): An image is worth 16x16 words: Transformers for image recognition at scale. Int. Conf. on Learning Representations (ICLR).

Examples of (more traditional) photogrammetry and remote sensing applications

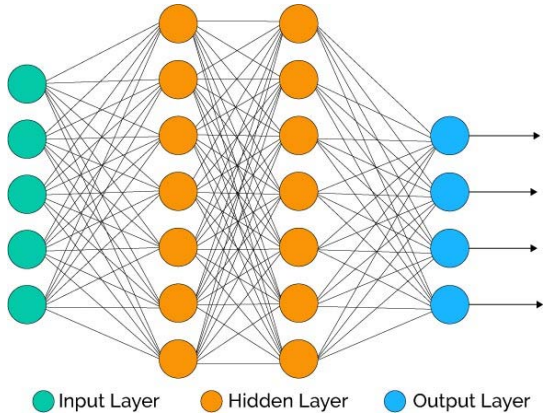
- image orientation
- dense stereo (and mono) 3D reconstruction
- land cover / land use classification

Photogrammetric workflow (geometry)



… a little extra

Deep learning

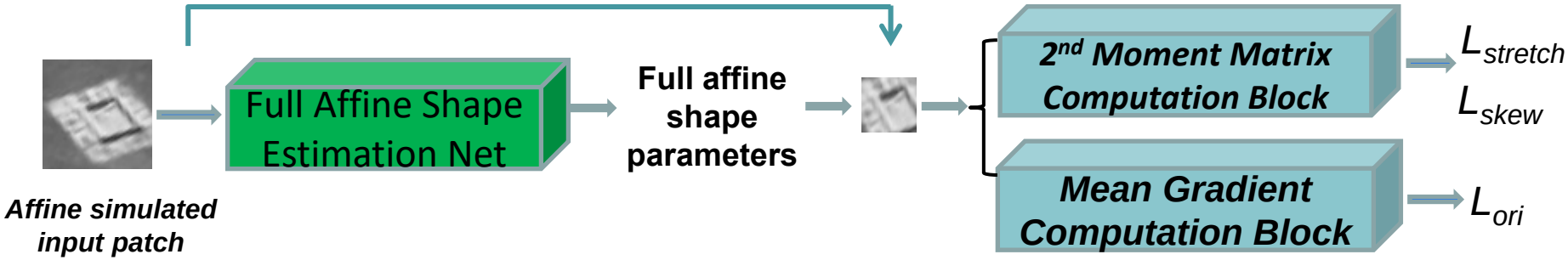


Example aerial image pair

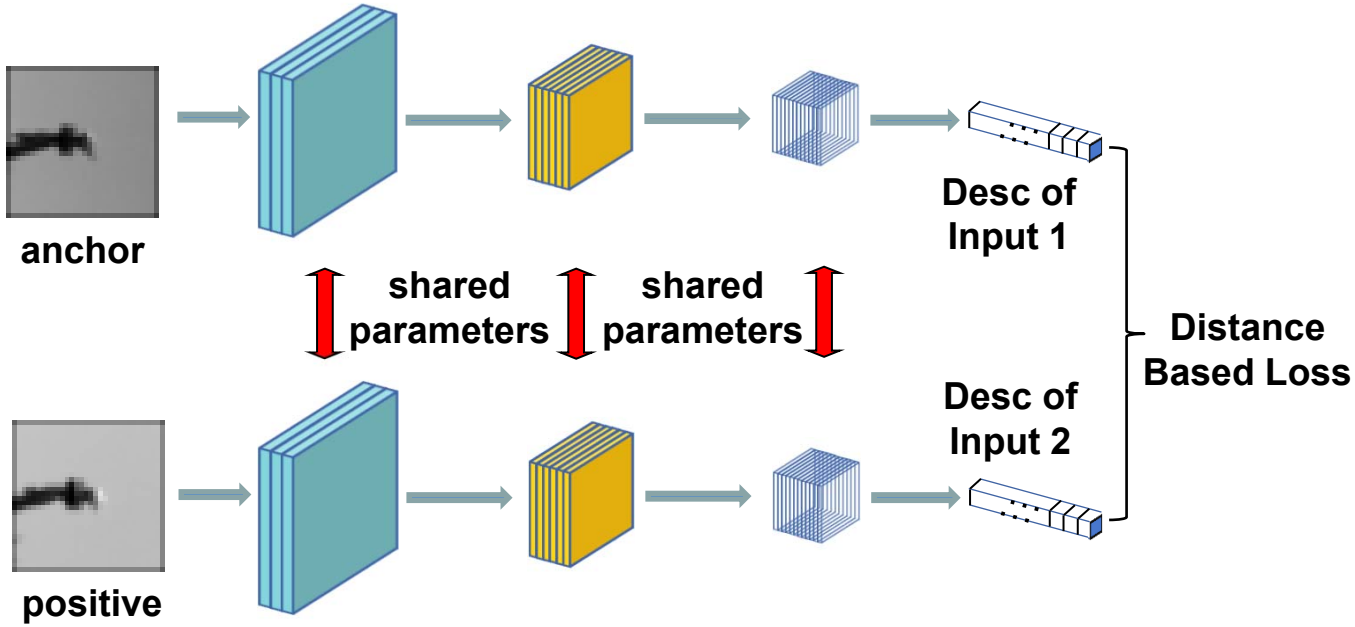
Chen L., 2021: Deep learning for feature based image matching. PhD thesis IPI, DGK C-867

Deep learning image orientation

- **Transformation into canonical representation**



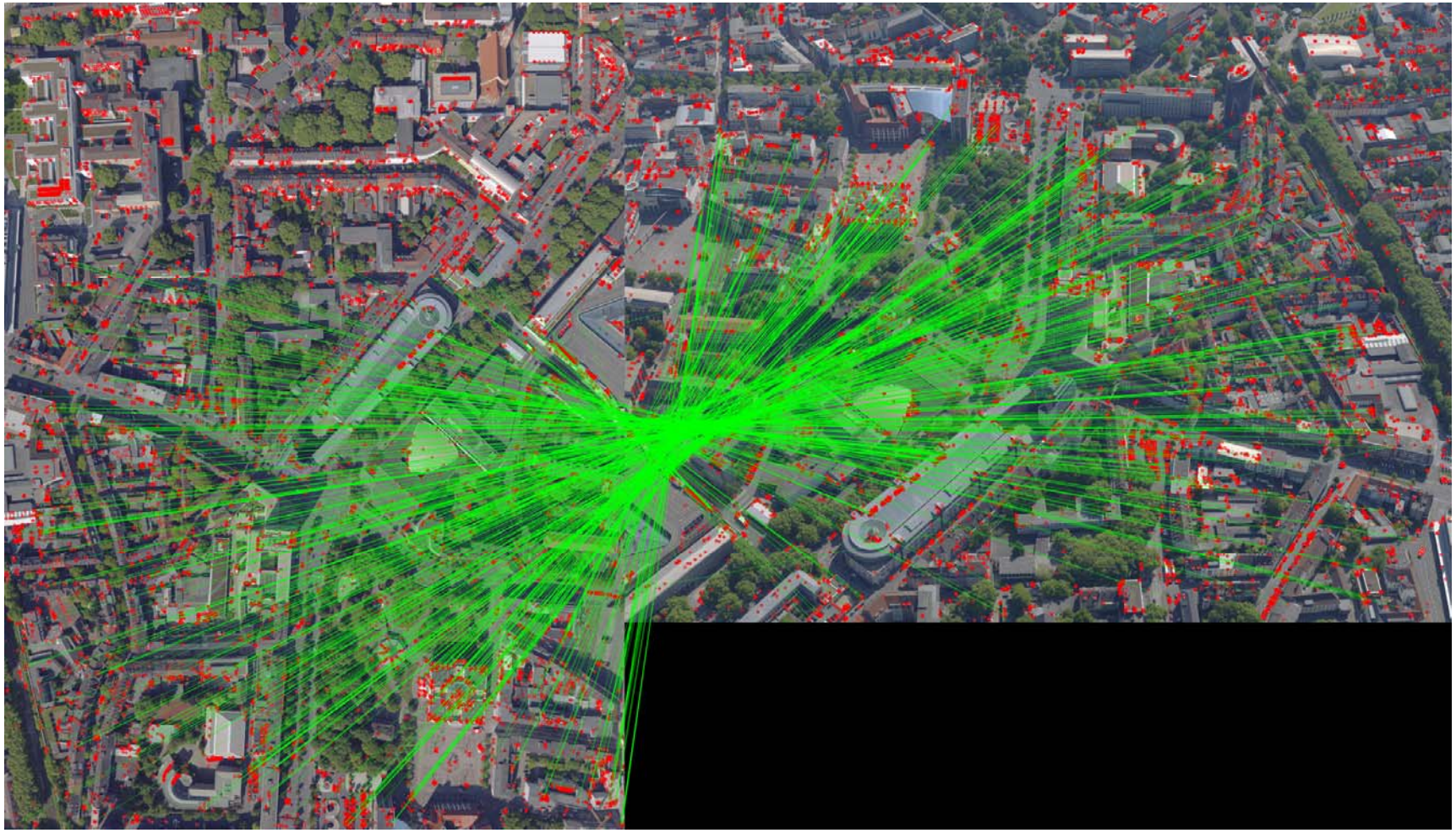
- **Feature matching and parameter estimation**



Chen L., 2021: Deep learning for feature based image matching. PhD thesis IPI, DGK C-867

Example image pair

- nadir-oblique matching (> 800 matches, σ_o = 0.6 .. 0.7 pixel)



Chen L., 2021: Deep learning for feature based image matching. PhD thesis IPI, DGK C-867

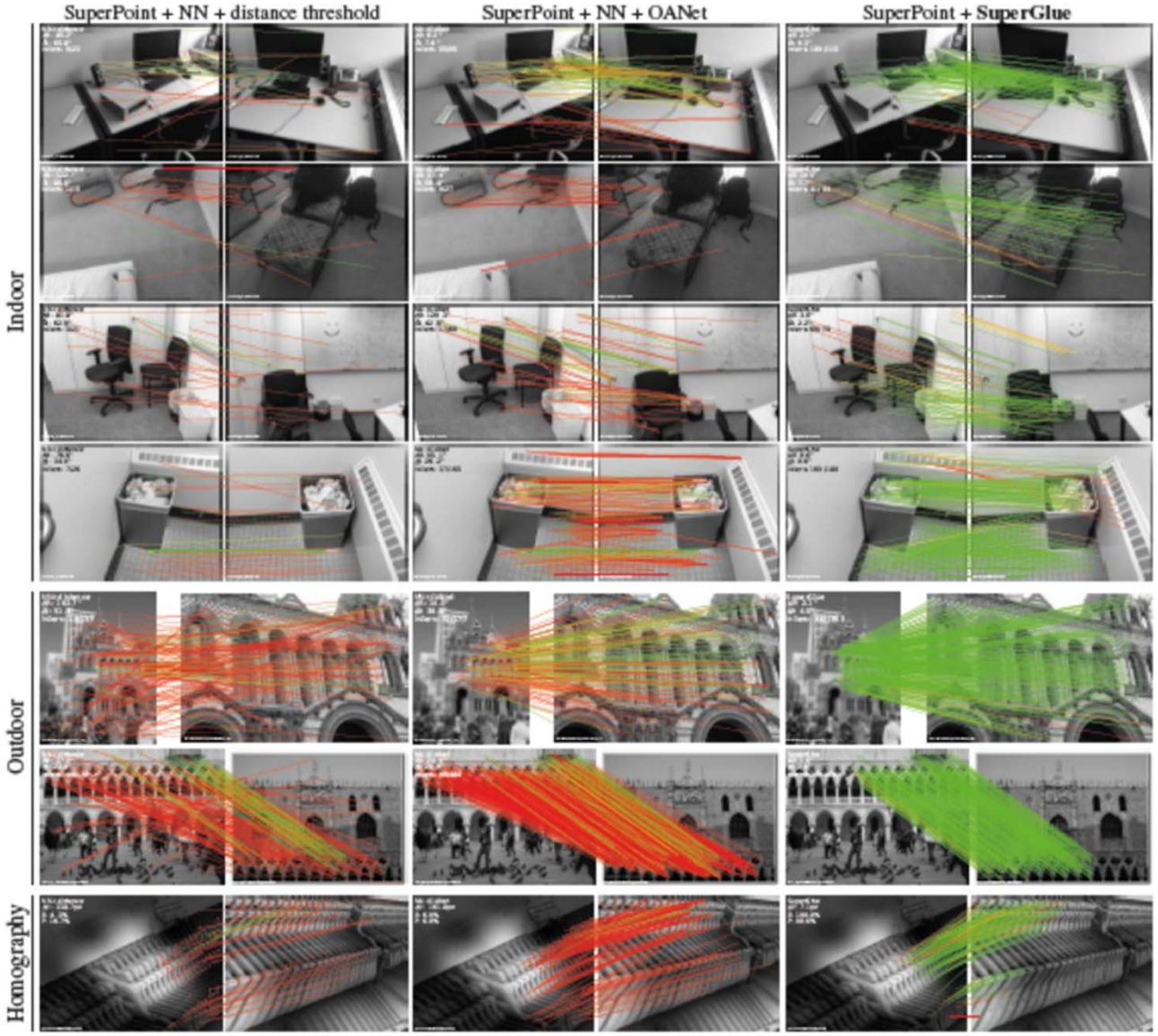


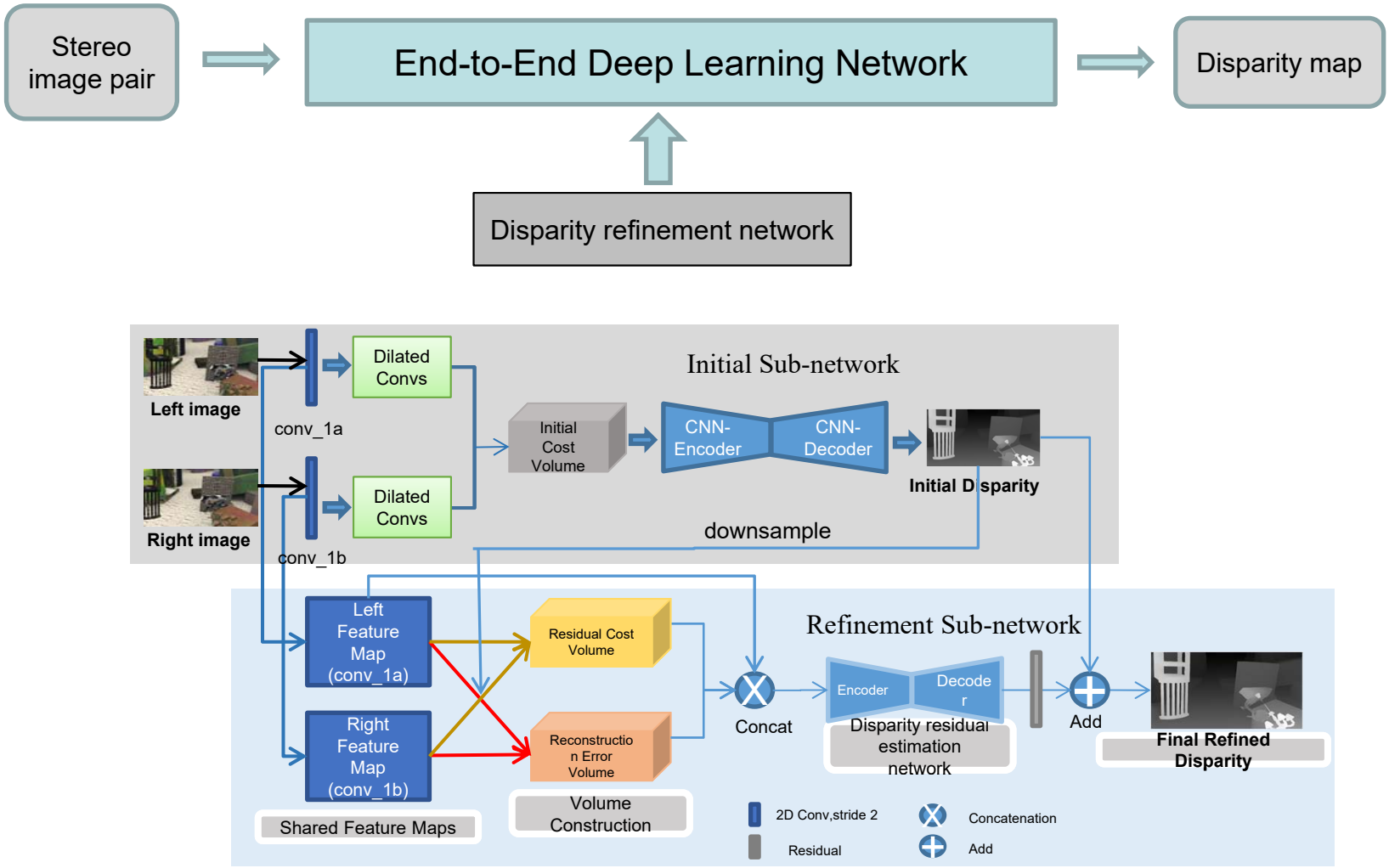
Figure 6: **Qualitative image matches.** We compare SuperGlue to the Nearest Neighbor (NN) matcher with two outlier rejectors, handcrafted and learned, in three environments. SuperGlue consistently estimates more correct matches (green lines) and fewer mismatches (red lines), successfully coping with repeated texture, large viewpoint, and illumination changes.

Deep learning image orientation

SuperPoint/SuperGlue:
- combination of interest point extraction and matching based on attention and a graph neural network, real-time

Sarlin P.E., DeTone D., Malisiewicz, T. Rabinovich, A., 2020: SuperGlue: Learning Feature Matching with Graph Neural Networks, CVPR. https://doi.org/10.48550/arXiv.1911.11763.

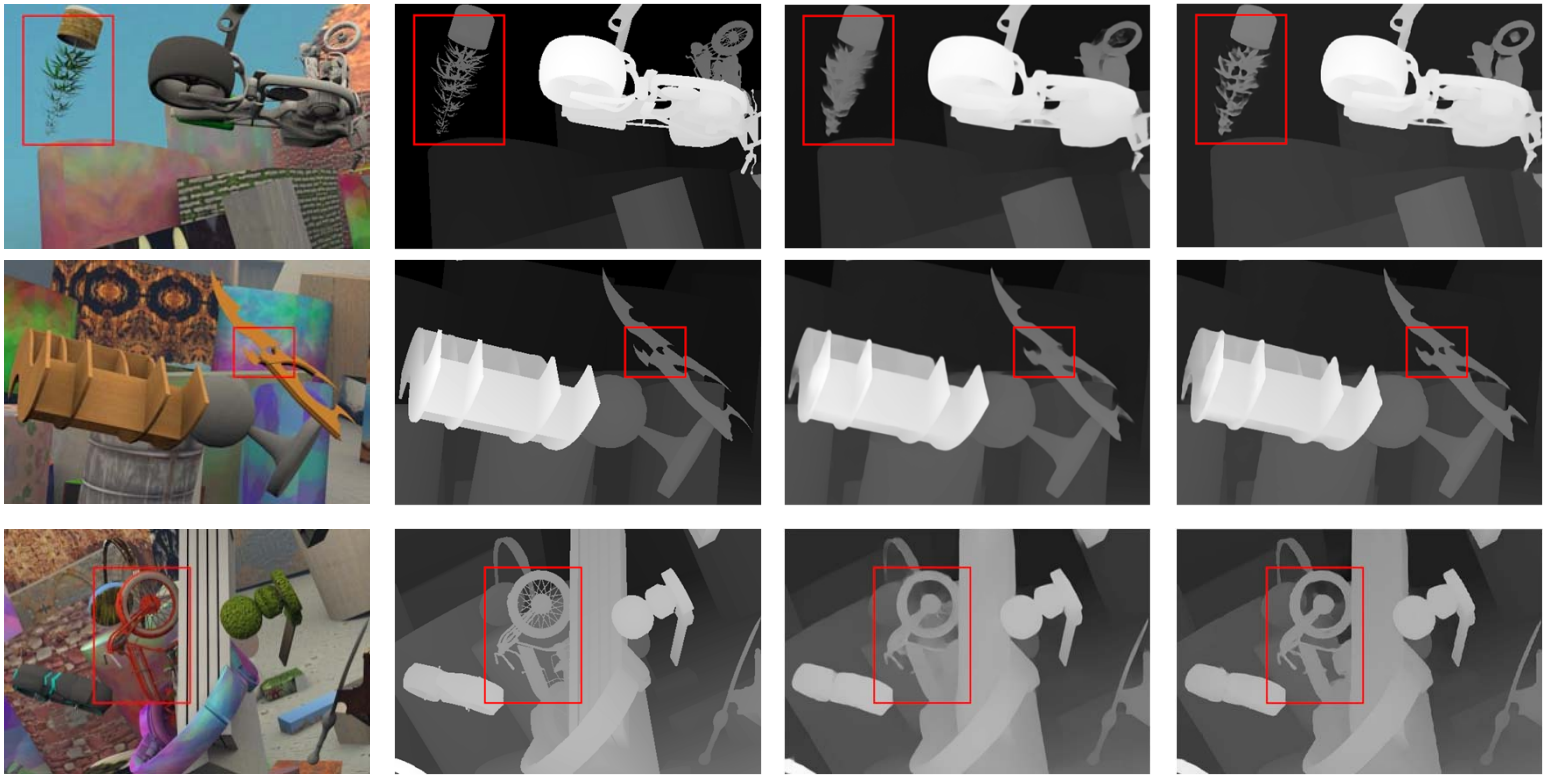
Dense stereo 3D reconstruction



Kang J., Chen L., Deng F., Heipke C., 2019: Context Pyramidal Network for Stereo Matching Regularized by Disparity Gradients. JPRS (157), 201-215.

Kang J., 2020: Deep learning for feature based image matching. PhD thesis, Wuhan University.

Dense stereo 3D reconstruction



left image | ground truth | initial disparity | refined disparity

left image | right image | baseline | ours

Kang J., Chen L., Deng F., Heipke C., 2019: Context Pyramidal Networks for Stereo Matching Regularized by Disparity Gradients. JPRS (157), 201-215.

Dense stereo 3D reconstruction

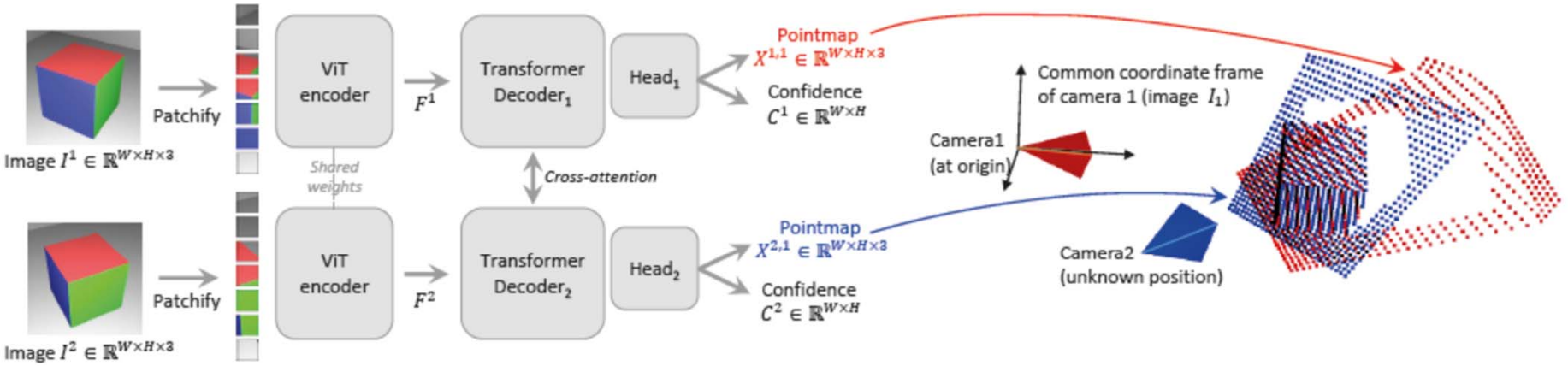


Figure 2. **Architecture of the network.** Two views of a scene (I^1, I^2) are first encoded in a Siamese manner with a shared ViT encoder. The resulting token representations F^1 and F^2 are then passed to two transformer decoders that constantly exchange information via cross-attention. Finally, two regression heads output the two corresponding pointmaps and associated confidence maps. Importantly, the two pointmaps are expressed in the same coordinate frame of the first image I^1. The network is trained using a simple regression loss (Eq. (4))

DUSt3R (Dense and Unconstrained Stereo 3D Reconstruction)

- integration of image orientation and dense 3D surface reconstruction
- vision transformers, also computes confidence map

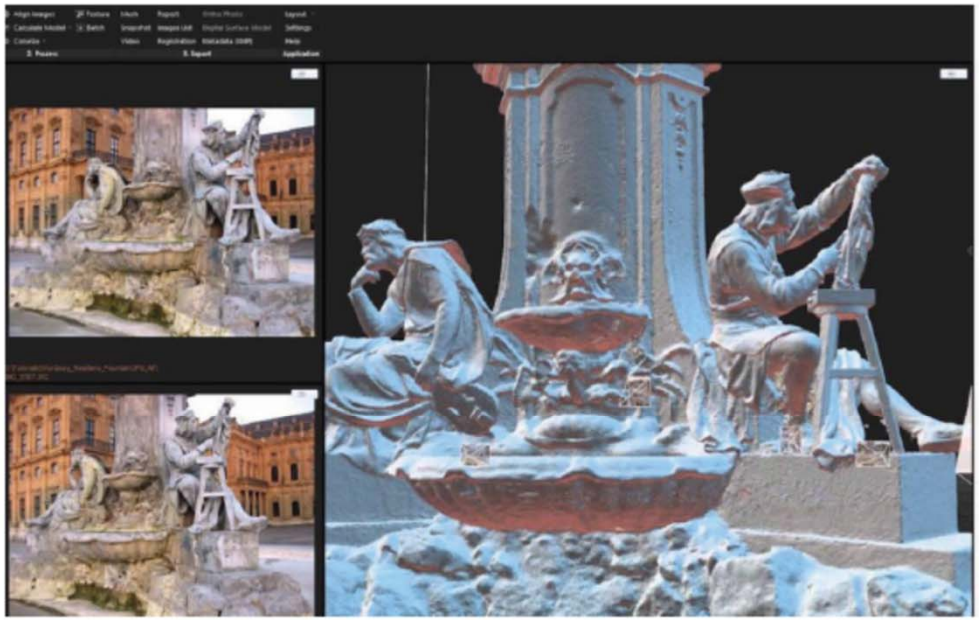
Wang, S., Leroy, V, Cabon, Y., Chidlovskii, B., Revaud, J., 2024: DUSt3R: Geometric 3D Vision Made Easy, CVPR. https://doi.org/10.48550/arXiv.2312.14132.

Dense stereo 3D reconstruction

Stereo Vision

ETH zürich PRS Photogrammetry Remote Sensing

A success story of computer vision and photogrammetry research: mature, “solved”, well-understood (including limitations)

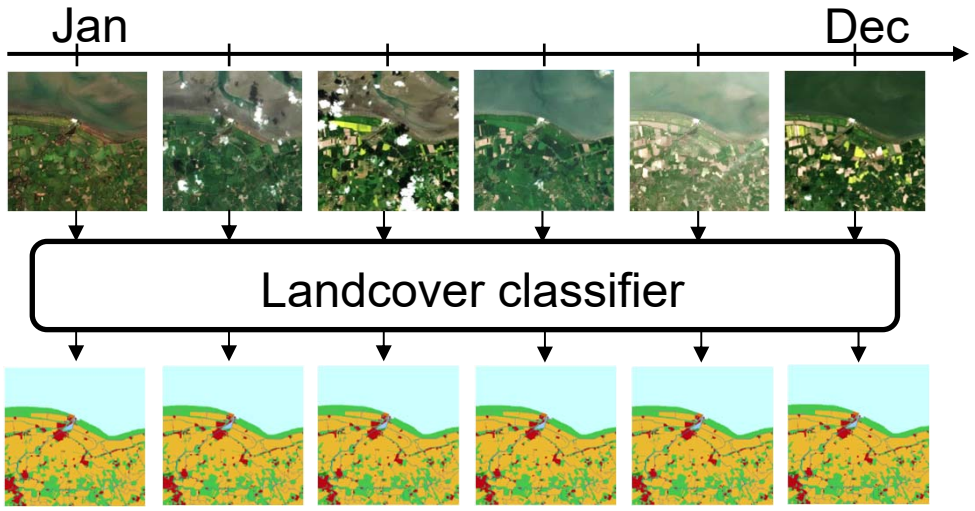


We have universal, off-the-shelf deployable methods and tools that can handle most reasonable images

Dense monocular 3D reconstruction

Multi-temp. LC class. using existing map data

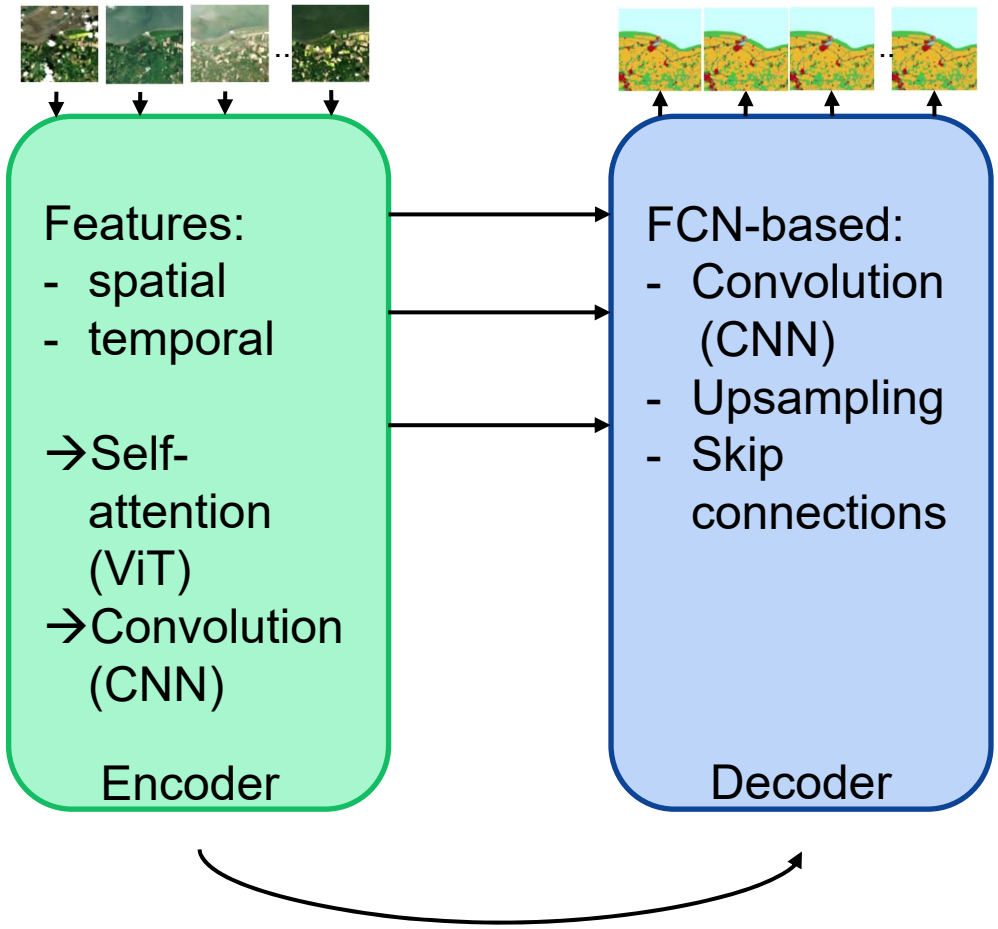
- Satellite image time series:
  - High **temporal** resolution, however irregular intervals
  - Medium **spatial** resolution
  - Seasonal variations for some classes, e.g. vegetation



- Multi-temporal in- & output, allows to analyse and monitor current state and development on the Earth surface (**"change"**)

- Goal: **multi-temporal** land cover classification
  - using existing DB for training

Multi-temp. LC class. using existing map data



Encoder - decoder neural network

Encoder:

- CNN for spatial features
- ViT for temporal features

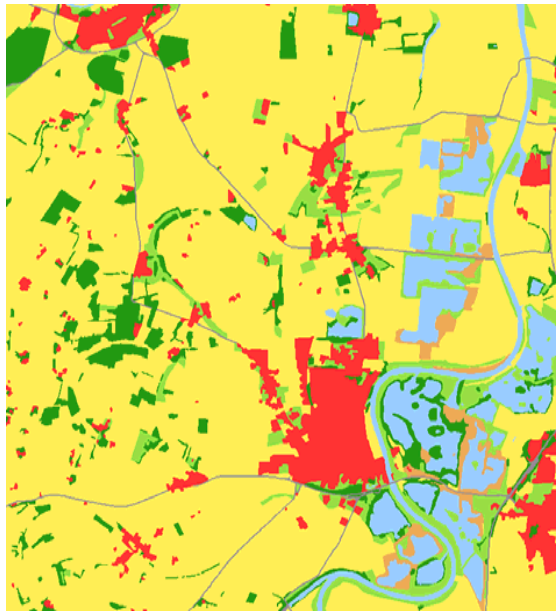
Decoder:

- Fully convolutional network (FCN)

Voelsen M., Rottensteiner F., Heipke, C., 2024: Transformer models for Land Cover Classification with Satellite Image Time Series. PFG (2024), 547-568. doi.org/10.1007/s41064-024-00299-7.

Multi-temp. LC class. using existing map data

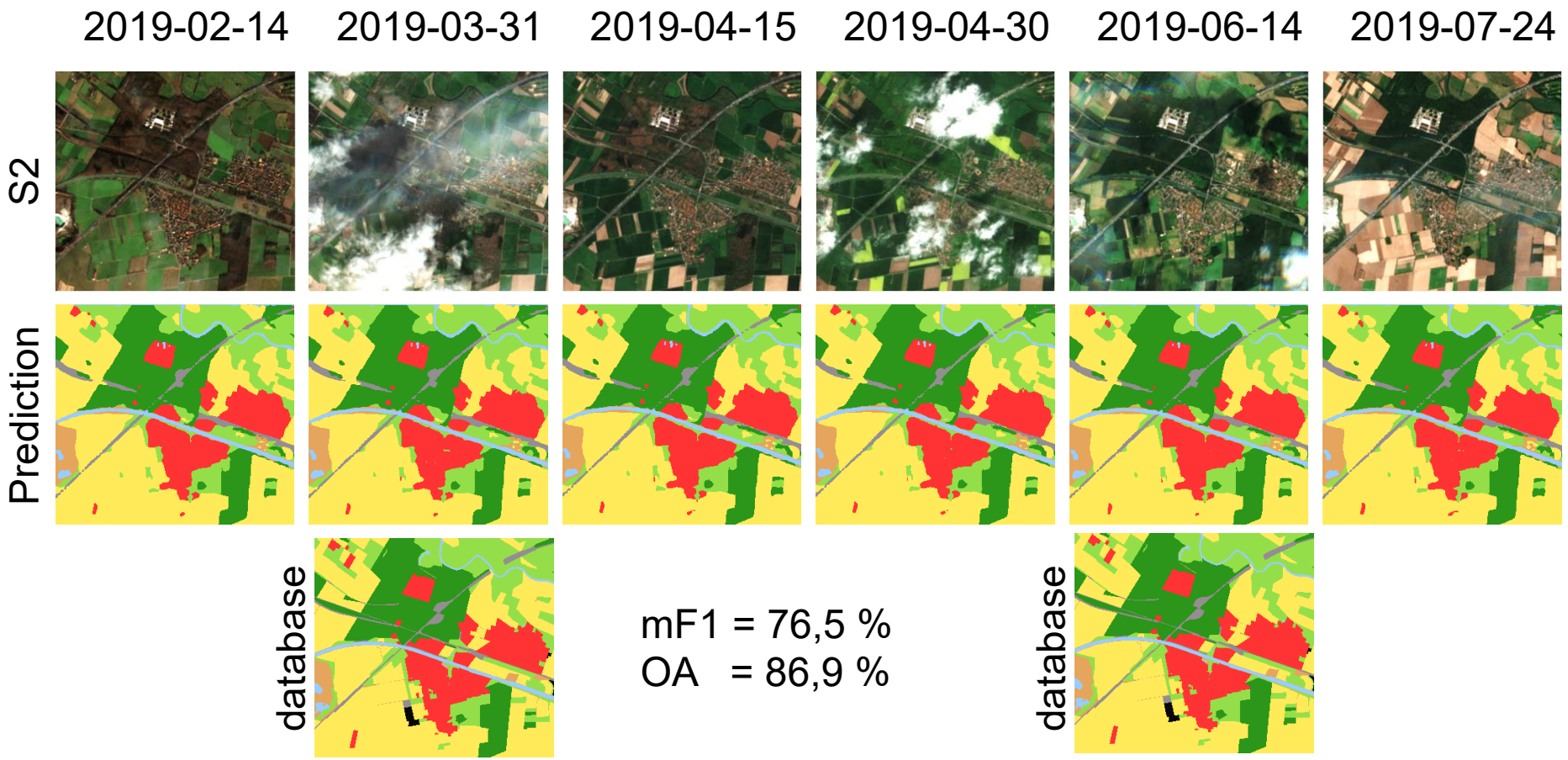
- Experimental data set: Sentinel-2 RGB and NIR channels, 10 m GSD
  - Lower Saxony (47.600 km^2)
  - all images captured between 2019 and 2022; < 5% cloud cover
- 7 land cover classes
  - highly imbalanced
- Top. DB for training
  - LGLN ATKIS[1]
  - no need for manual labelling
- Ground truth also from ATKIS DB



[1] LGLN: Land Survey Office of Lower Saxony
ATKIS: Authoritative Topographic-Cartographic Information System

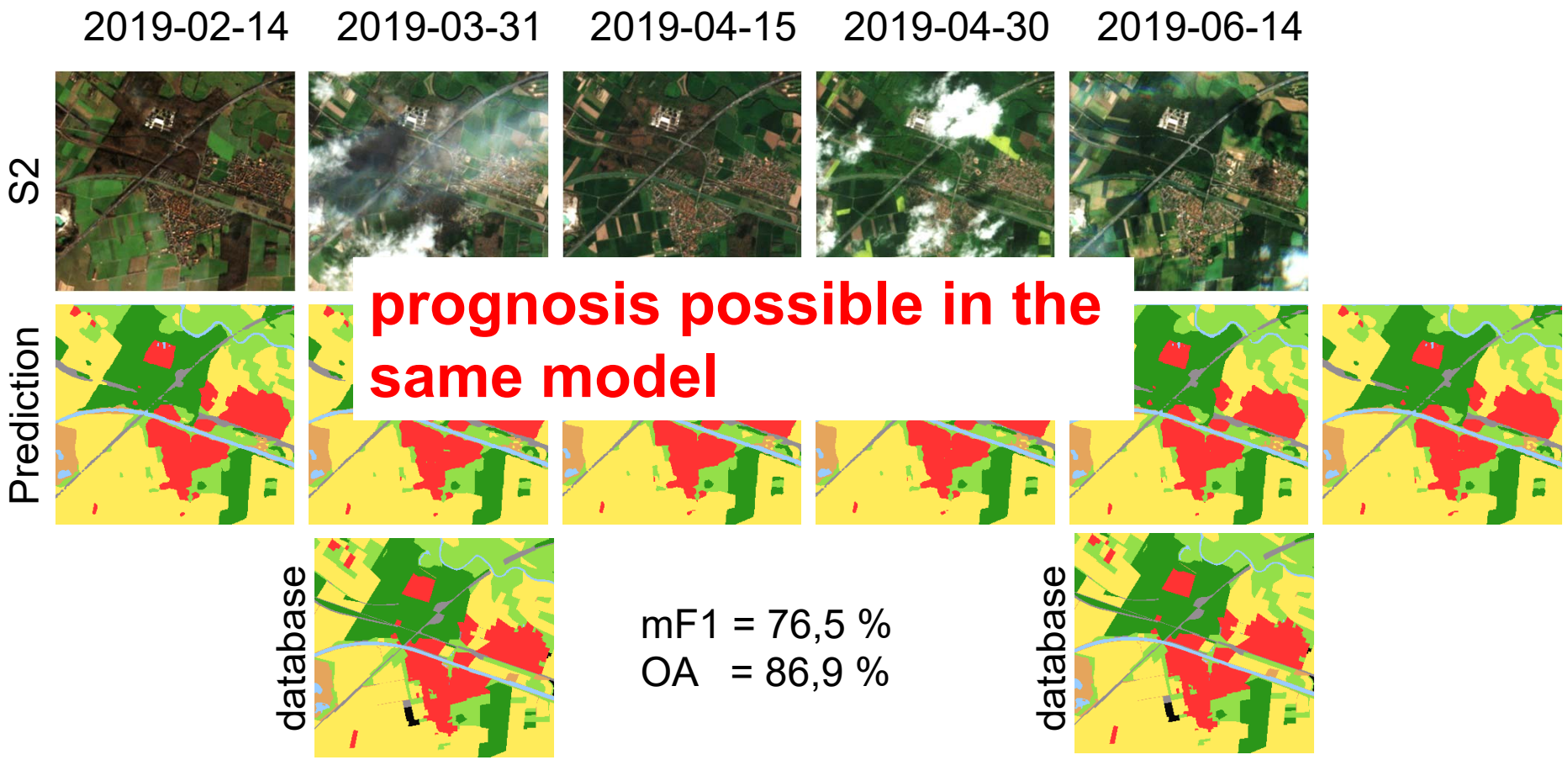
Voelsen M., Rottensteiner F., Heipke, C., 2024: Transformer models for Land Cover Classification with Satellite Image Time Series. PFG (2024), 547-568. doi.org/10.1007/s41064-024-00299-7.

Multi-temp. LC class. using existing map data



Voelsen M., Rottensteiner F., Heipke, C., 2024: Transformer models for Land Cover Classification with Satellite Image Time Series. PFG (2024), 547-568. doi.org/10.1007/s41064-024-00299-7.

Multi-temp. LC class. using existing map data



Voelsen M., Rottensteiner F., Heipke, C., 2024: Transformer models for Land Cover Classification with Satellite Image Time Series. PFG (2024), 547-568. doi.org/10.1007/s41064-024-00299-7.

Conclusions

Conclusions – where do we stand?

- **Deep learning** is here to stay, for semantics & geometry
  - implicit models have outperformed explicit ones everywhere
    - but remember: **Deep learning is statistics**, not magic …
    - details do matter (adapted architecture, hyper-parameters, …)
- Open issues require further research efforts
  - unbalanced data sets in training: **bias towards majority classes**
  - **lack of training data** in many applications
  - no integration of domain knowledge (“don’t learn what you already know”)
  - …
- Deep learning is a black box
  - quality of results for unseen data unclear, in particular for large end-to-end systems (**how well can DL methods generalize**?)

Conclusions – technical trends

- Sensor / data **fusion** for more challenging scenarios
  - images + X (laser scanner, radar, GNSS, IMU, …)
- Integration of **geometry and semantics**
  - synergy: geometry and semantics can support each other
  - using, e.g., deep implicit semantic occupancy fields for surface representation
- **NeRF** (Neural Radiance Fields) and **Gaussian splatting**
  - new view synthesis, needs “good enough” 3D surface model
- **Time series** processing, e.g., in the context of sustainability
  - combined evaluation of images, maps etc. to monitor environment and predict landscape changes and environmental parameters
- …

Conclusions – some challenges

- CNN: local, rigid neighbourhood only (digital filters)
  - Visual Transformers, wider context
- DL models: extremely high model complexity
  - **vast amounts of unknowns**, in particular for **foundation models** (these need a lot of computing power for training)
  - only few data centres have **required computational resources**
- Need for a significant amount of labelled training data
  - general pre-training, followed by fine-tuning (**transfer learning**)
  - **self-supervised learning** (pre-training without labels, e.g. MAE)
  - **label noise**: quality of training data not always sufficient
  - simulated training data, e.g., using generative AI
- Fake news
  - how do we know we can **trust the results**?

IPI 20205

… thank you very much for your attention!