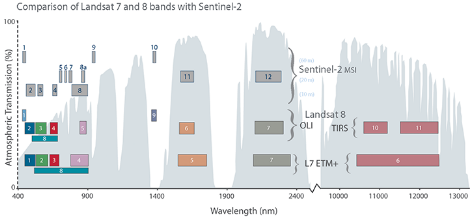
Transformers in Remote Sensing

Georg Zitzlsberger, IT4Innovations, 13-09-2023

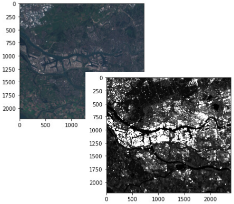
Remote Sensing Data - Types

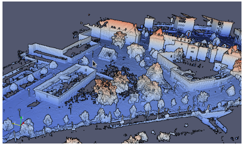
Most common remote sensing types are:

- Multi-/Hyper-spectral Optical:
  Passive scanning of reflected sunlight
- Synthetic Aperture Radar (SAR):
  Active emission of radar and detection of backscattered energy and polarizations
- Light Detection and Ranging (LiDAR):
  Active emission of laser and detection of backscattered energy



NASA/ESA

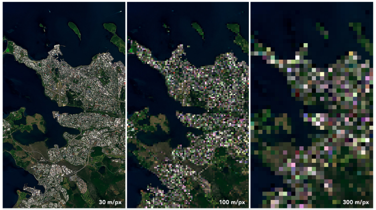

Optical (true color) and SAR observations of AoI Rotterdam


Bastian Steder, University of Freiburg, Dept. of Computer Science, Creative Commons Attribution License 3.0

Remote Sensing Data - Resolution

Resolutions can be categorized:

- Low Resolution (LR): > 30 m/pixel
- Medium Resolution (MR): 5-30 m/pixel
- High Resolution (HR): 1-5 m/pixel
- Very High Resolution (VHR): < 1 m/pixel



Landsat 8 image of Reykjavik, Iceland, acquired July 7, 2019, illustrating the difference in pixel resolution. Credit: NASA Earth Observatory.

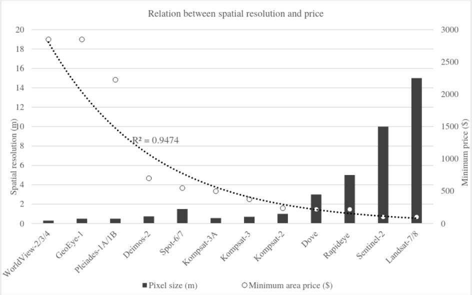


Figure 1 Comparison between price and spatial resolution

Based on the minimum price area and the estimate precision agriculture advantages (30$ ha^{-1}), a calculation of the break-even point for all the sensors was performed.

Image from Sozzi et al., 2018: Benchmark of Satellites Image Services for Precision Agricultural use

Change Detection

- First mentioned 1964 by Shepard: “A **change detector device** is needed which will automatically correlate the overlapping area and indicate all the changes.”
- Many menthods for Change Detection (CD) were created since then:
  - Difference (optical) or ratio (SAR)
  - Decision trees
  - Change Vector Analysis (CVA)
  - Slow Feature Analysis (SFA)
  - Principal Component Analysis (PCA) and clustering
  - ...
  - **Deep Neural Networks (DNNs)**

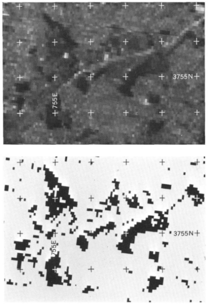


FIGURE 3.—Digital analysis of temporal overlay of Landsat band-5 data. Upper, ratioed data (band 5, 1972, divided by band 5, 1974). Lower, land-use-change theme extracted from ratioed data. Approximate location of UTM 1.0-km tick marks (zone 16) indicates scale and direction.

Image from Todd 1977: Urban and Regional Land Use Change Detected by using Landsat Data

What to Detect?

Detecting a specific type of change is hard:

- Buildings, Roads or infrastructure
- Mobile objects (cars, trucks, containers)
- Vegetation or phenology
- Trees and forests
- Water bodies

What about:

- Seasonal changes?
- Irradiance/lighting changes (shadows)?
- Atmospheric problems (clouds, haze, fog)?

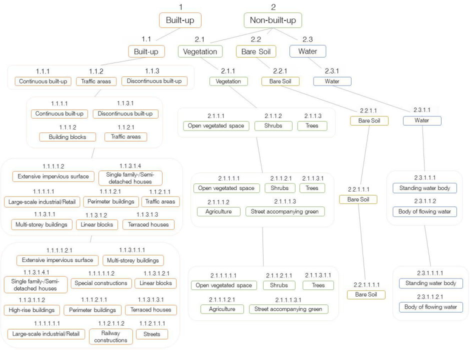
Figure 8. Examples of change detection in remote sensing images: left – input image A, middle – input image B, right – synthesized difference map

Image from Lebedev et al. 2018
(Season-varying CD Dataset)

Detecting Urban Changes

What is actually *urban*?

- Let's alias built-up as urban
- Many different built-up sub-types
- Any change involves a transition from/to any type to/from built-up
  Note: This also involves from and to build-up!



Urban Structure Type from Lehner et al. 2019

The Current DNN Methods

There are many methods but the current DNN-based ones are dominated by:

- Siamese networks: Two observations at two different times
- Curated VHR data, mostly RGB (only three spectral bands)

Fusion for Change Detection (CD):

- Image Level Fusion/Early Fusion
- Feature Level Fusion/Late Fusion
- Multi-scale Feature Fusion (Feature Pyramid, UNet, ...)

Attention:

- Channel attention
- Spatial attention
- Self-Attention

The Current DNN Methods - cont'd

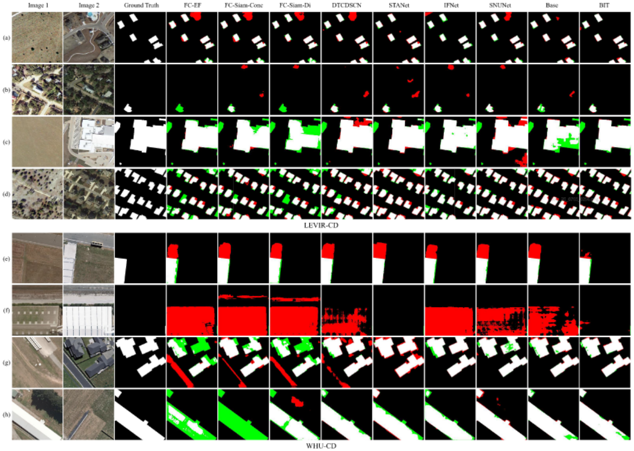


Image from Chen et al. 2022

Bitemporal Image Transformer (BIT)

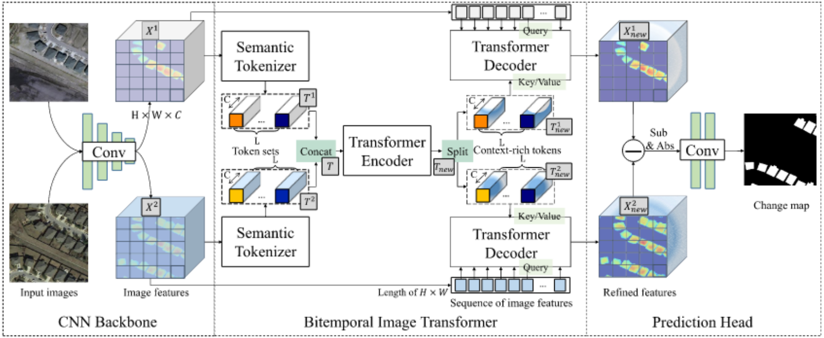


Fig. 2. Illustration of our BIT-based model. Our semantic tokenizer pools the image features extracted by a CNN backbone to a compact vocabulary set of tokens ($L \ll HW$). Then we feed the concatenated bitemporal tokens to the TE to relate concepts in token-based space-time. The resulting context-rich tokens for each temporal image are projected back to the pixel-space to refine the original features via the TD. Finally, our prediction head produces the pixel-level predictions by feeding the computed FDIs to a shallow CNN.

Bitemporal Image Transformer (BIT) from Chen et al. 2022

Bitemporal Image Transformer (BIT)

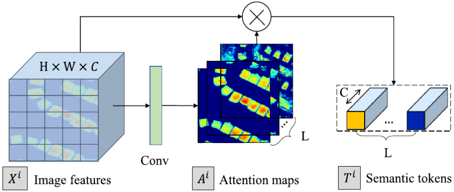


Fig. 3. Illustration of our semantic tokenizer.

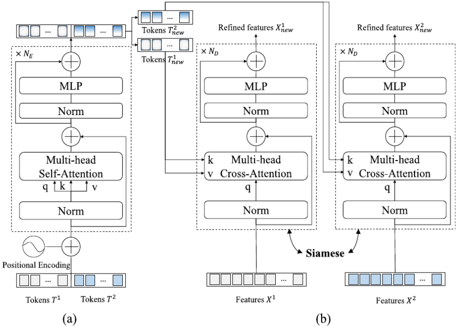


Fig. 4. Illustration of our (a) TE and (b) TD.

Bitemporal Image Transformer (BIT) from Chen et al. 2022

Bitemporal Image Transformer (BIT) - cont'd

- Code is available on ▸ GitHub
- The ▸ LEVIR-CD dataset is used:
  - Originally 637 image pairs with $1,024 \times 1,024$ pixels
  - The 637 pairs are tiled into 256×256 patches
  - Images are RGB
  - Temporal span of 5-14 years
  - Contains ground truth labels (binary)
  - Only building changes are covered

Let's look at the implementation

Thank you!

This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 101101903. The JU receives support from the Digital Europe Programme and Germany, Bulgaria, Austria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, Greece, Hungary, Ireland, Italy, Lithuania, Latvia, Poland, Portugal, Romania, Slovenia, Spain, Sweden, France, Netherlands, Belgium, Luxembourg, Slovakia, Norway, Türkiye, Republic of North Macedonia, Iceland, Montenegro, Serbia.