GEOAI 2026 · GHENT

A good data model is worth a *complex* neural network

Evidence from building generalization using deep learning

Authors Zhou, **Winkler**, Fu, Weibel

Affiliation UZH · Dira GeoSystems AG

Dataset OSM-Stuttgart & swisstopo-Zürich

Funded by SNSF DeepGeneralization

BACKGROUND

Map generalization

When a map changes scale, features must be **simplified, merged, displaced or removed**: features are truly transformed, not just shrunk.

Elimination
small buildings removed

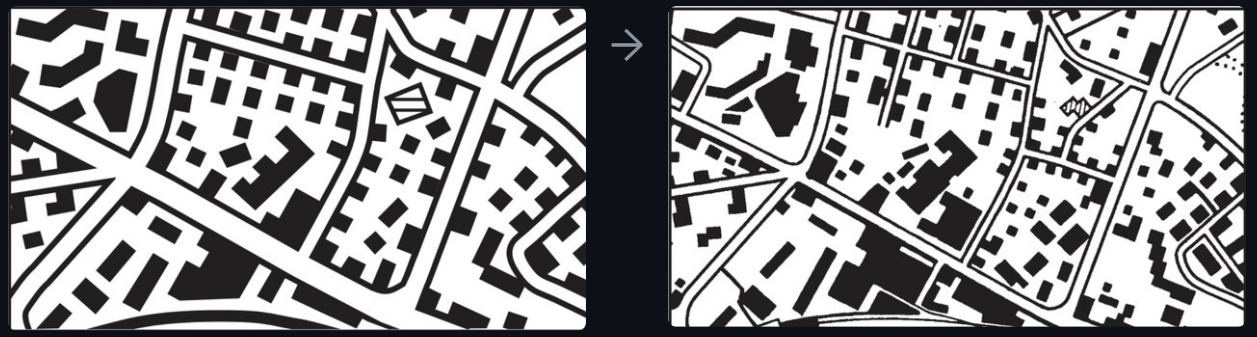
Simplification
walls straightened

Aggregation
touching buildings merged

Displacement
overlap resolved

Challenge: formalising *expert cartographic knowledge* into rules is a known bottleneck (Weibel et al. 1995). DL offers a data-driven alternative.

1 : 10 000 (SOURCE)

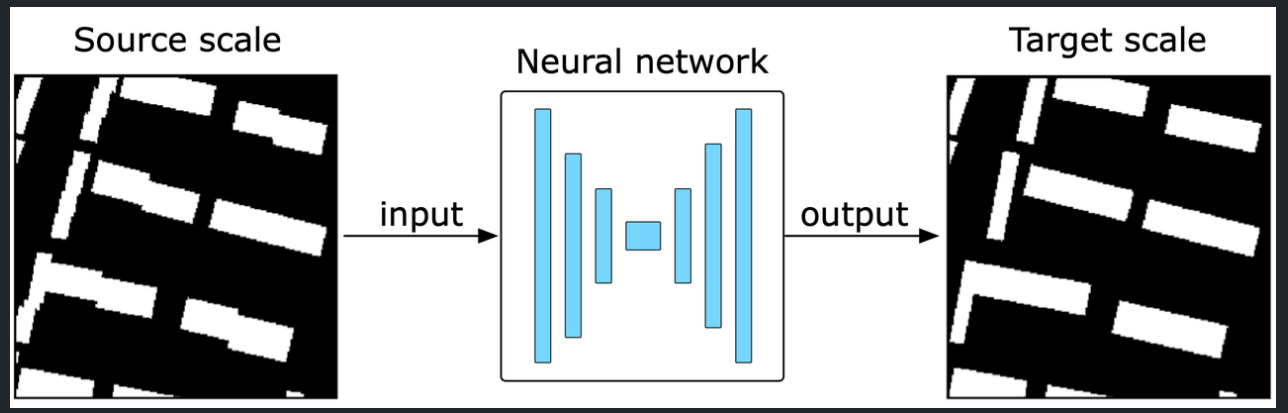


1 : 25 000 (TARGET)

swisstopo data, Beglinger 2022

PROBLEM FORMULATION

Generalization as image prediction



Building generalization as a raster image prediction task: 1:10 000 → 1:25 000 (224 × 224 px patches, binary maps)

PRIOR WORK (FENG ET AL. 2019)

U-Net, single-channel binary raster maps. Promising IoU results, but evaluated on one dataset (OSM-Stuttgart) with one architecture.

THIS STUDY

We vary **both axes independently**: data model (1-channel vs 2-channel) and architecture (U-Net vs Swin Transformer), across two datasets.

RESEARCH QUESTION

Does a better *data model* beat a more complex *neural network*?

AXIS 1: DATA MODEL

Single-channel (baseline)
vs. two-channel (focal + context)

AXIS 2: ARCHITECTURE

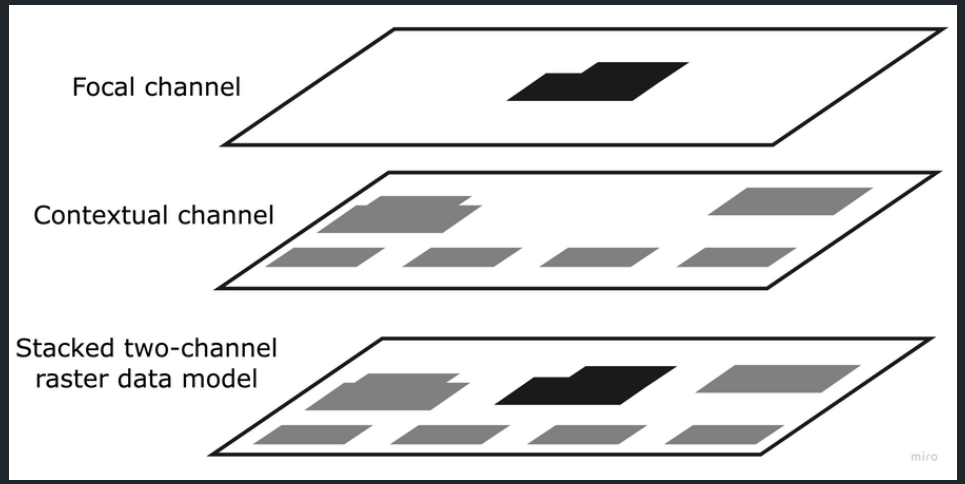
U-Net (30M params, CNN)
vs. Swin Transformer (85M params)

VALIDATED ON

OSM-Stuttgart (simple)
swisstopo-Zürich (complex)

METHOD: DATA MODELS

Single-channel vs. two-channel



Focal channel (one building) + contextual channel (neighbours) stacked as a 2-channel input

SINGLE-CHANNEL (1-CH)

All buildings in one binary map. Network must decide which to keep, merge, or remove, without explicit focus.

TWO-CHANNEL: CENTERED (2-CH-CEN)

Focal building always centred. Larger dataset by design (one patch per building). Proposed by Fu et al. 2024.

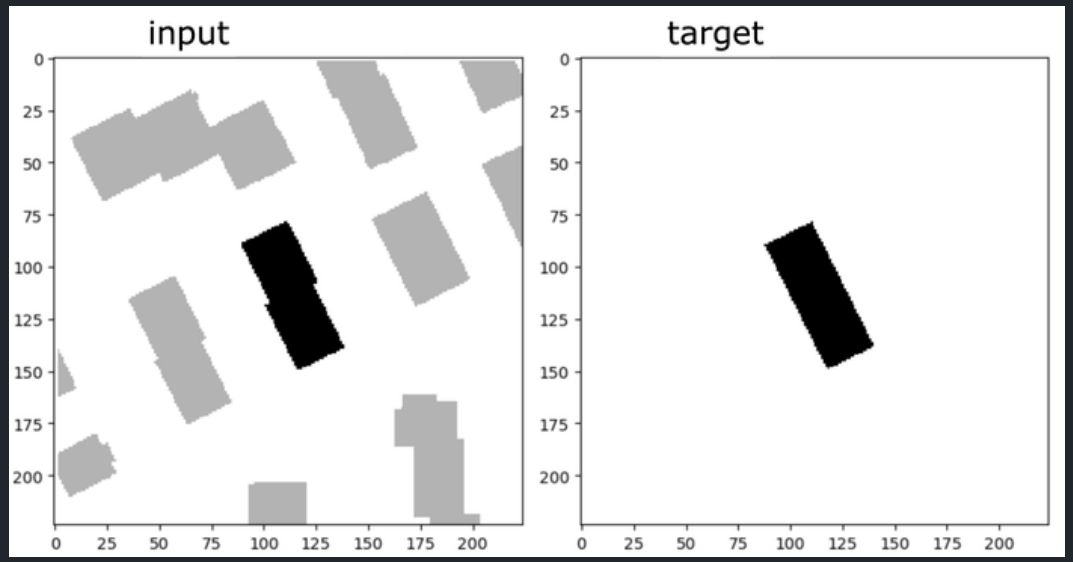
TWO-CHANNEL: RANDOM (2-CH-RND) NEW

Focal building placed randomly. Same dataset size as 1-Ch. Controls for training-set-size confound.

METHOD: DATA MODELS

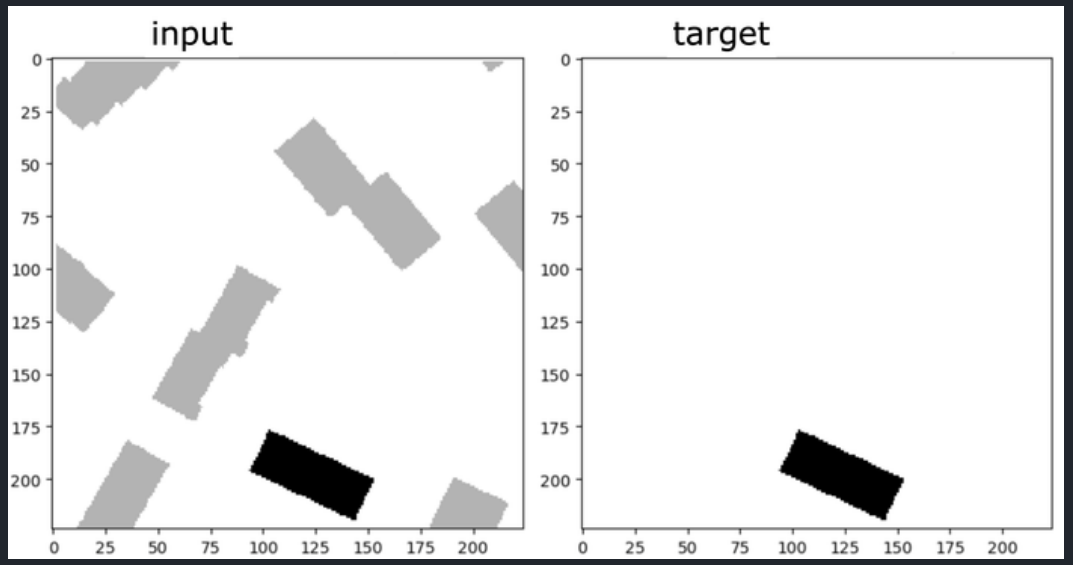
What the network actually sees

CENTERED-FOCAL (2-CH-CEN)



Focal building (black) centred; contextual neighbours (grey). Network predicts only the focal building.

RANDOM-FOCAL (2-CH-RND)



Same building count as 1-Ch. Isolates the benefit of the focal/context split from dataset-size effects.

Key insight: the 2-Ch-Rnd model has the *same number of patches* as 1-Ch, so any improvement is purely from the layered representation, not from more data.

METHOD: NETWORKS & DATASETS

Two networks, two datasets

U-Net

~30M

PARAMETERS

Classical CNN encoder-decoder with skip connections. Prior SOTA for this task.

Swin Transformer

~85M

PARAMETERS

Hierarchical vision transformer with shifted-window self-attention. 3× more complex. Proven in semantic segmentation (Liu et al. 2021).

OSM – STUTTGART

~30k patches · Simple generalization (fewer operators) · established benchmark from Feng et al.

1:10K → 1:25K

SWISSTOPO – ZÜRICH

~37k patches · Complex generalization (all operators active) · Swiss topographic standard

1:10K → 1:25K

NEW DATASET

RESULTS: FINDING 1

Two-channel consistently beats single-channel

OSM – STUTTGART (SWIN TRANSFORMER)

DATA MODEL	MIOU	Δ VS 1-CH
1-Ch (baseline)	0.961	ref
2-Ch-Cen	**0.974**	**+0.013**
2-Ch-Rnd	**0.965**	**+0.004**

SWISSTOPO – ZÜRICH (SWIN TRANSFORMER)

DATA MODEL	MIOU	Δ VS 1-CH
1-Ch (baseline)	0.915	ref
2-Ch-Cen	**0.941**	**+0.026**
2-Ch-Rnd	**0.951**	**+0.036**

> **2-Ch-Rnd beats 1-Ch with the *same number of patches*: the improvement comes from the representation, not the data volume.**

The effect is larger on the harder Zürich dataset (+3.6 pp): the focal/context split helps most when generalization is complex.

RESULTS: FINDING 2

U-Net outperforms Swin Transformer

despite being 3× smaller

SAME DATA MODEL (2-CH-CEN): BOTH DATASETS

DATASET	NETWORK	MIOU
OSM – Stuttgart	**U-Net (~30M)**	**0.999**
OSM – Stuttgart	Swin T. (~85M)	0.978
swisstopo – Zürich	**U-Net (~30M)**	**0.958**
swisstopo – Zürich	Swin T. (~85M)	0.942

> **More parameters ≠ better performance. Inductive bias of CNNs aligns with the local spatial structure of buildings.**

The Swin Transformer's weaker inductive bias means it needs far more training data to catch up, a known limitation on small datasets.

CONCLUSIONS

What this means for practitioners

> **A good data model grounded in *domain knowledge* can outperform a 3× more complex neural network.**

BEFORE REACHING FOR A BIGGER MODEL…

- Encode domain knowledge into the data representation
- Align it with expert attention: focal building + context
- Try a simpler architecture first: U-Net beat Swin here

OPEN QUESTIONS

- Will Transformers close the gap with larger training sets?
- Does the two-channel approach generalise to other feature types (roads, land use)?
- Cross-national transfer: does Swin's weaker inductive bias help?

Zhou, Winkler, Fu, Weibel (2025) · SNSF DeepGeneralization 200021_204081 · Data: swisstopo open data