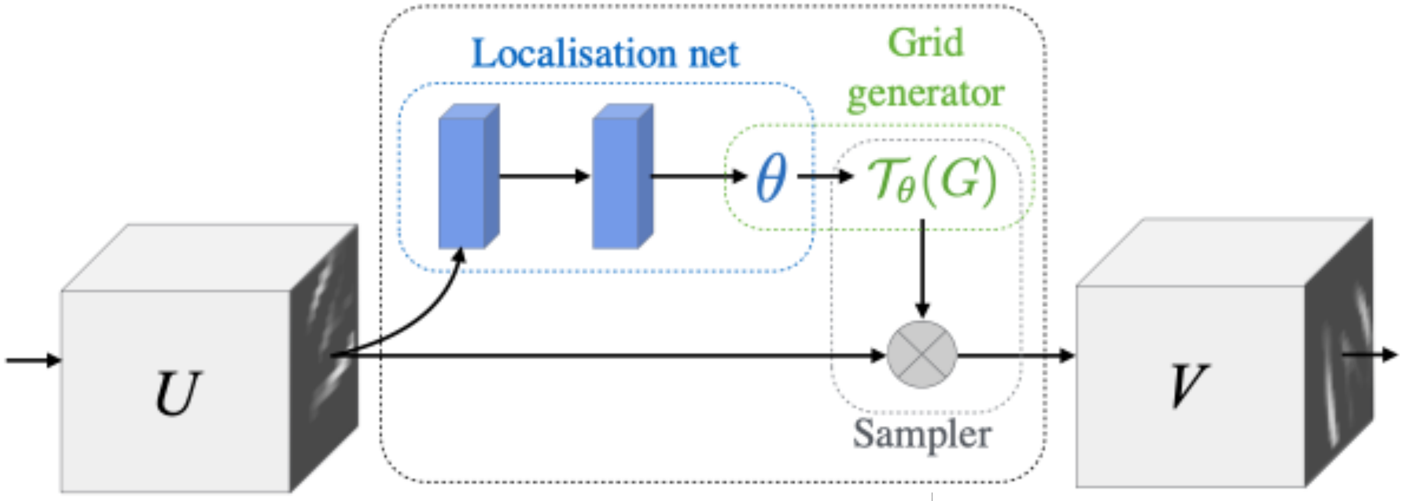
Understanding when Spatial Transformer Networks do not support invariance, and what to do about it

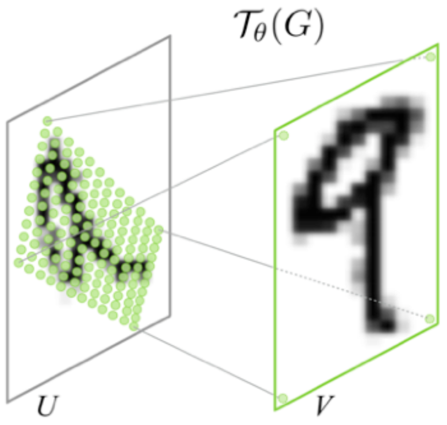
Lukas Finnveden, Ylva Jansson and Tony Lindeberg

Computational Brain Science Lab
Division of Computational Science and Technology
KTH Royal Institute of Technology
Stockholm, Sweden

Spatial transformer networks and invariance



Spatial transformer

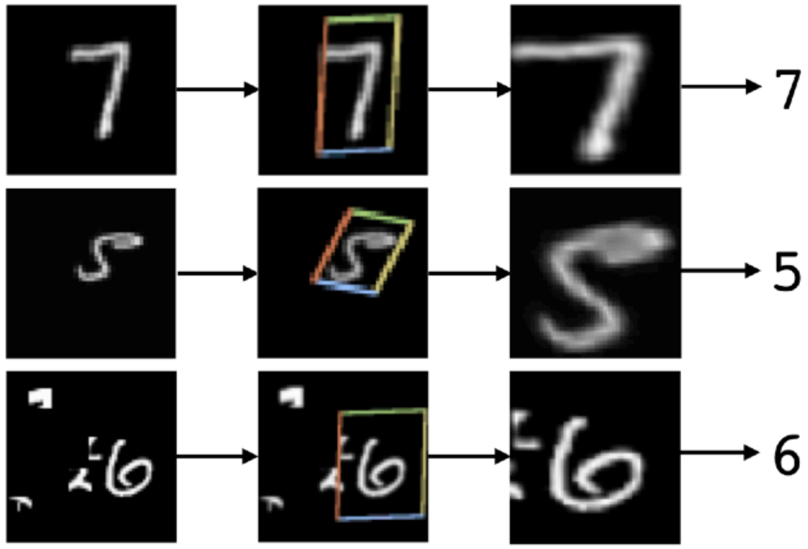


Resampling

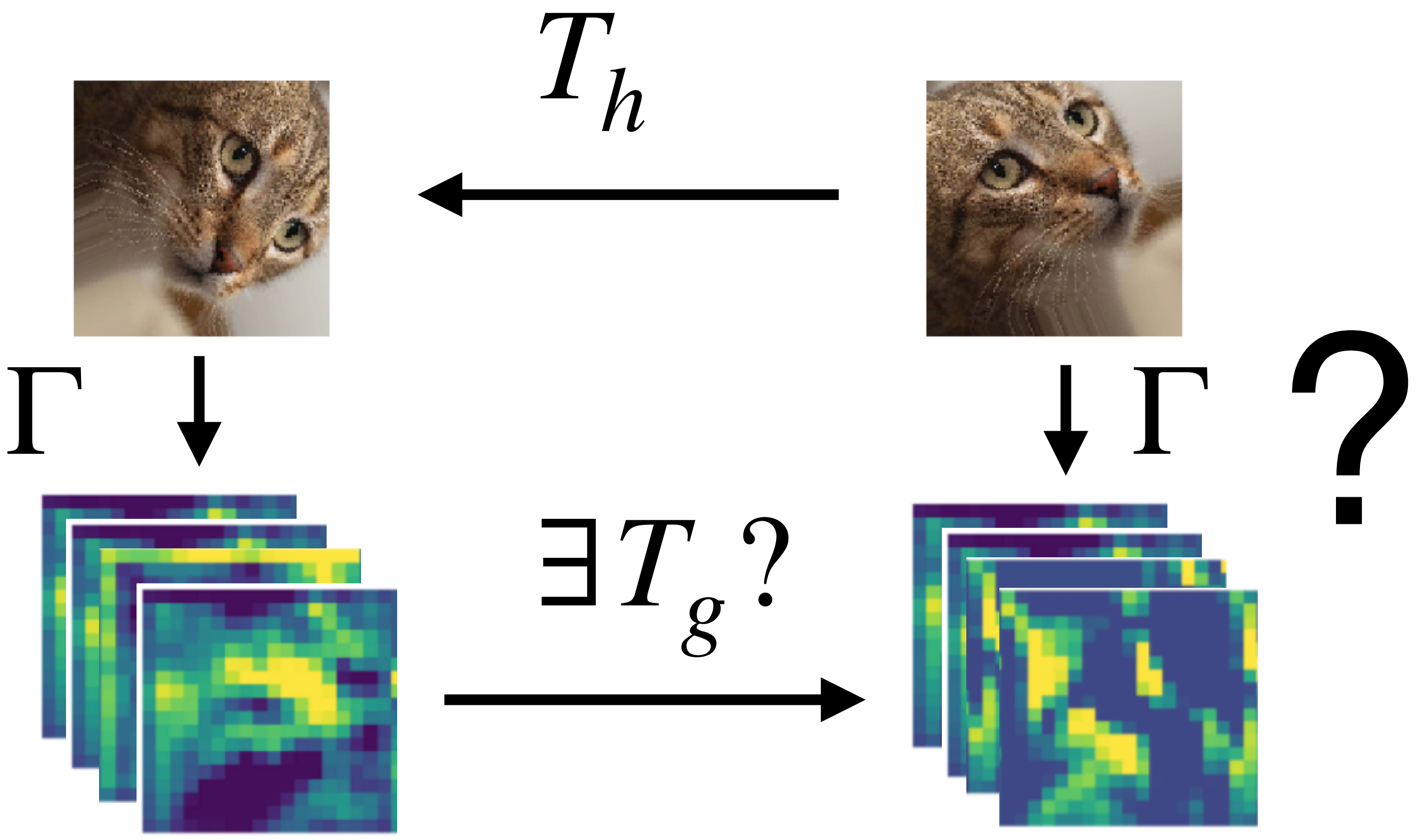
Jaderberg et al. (2015)

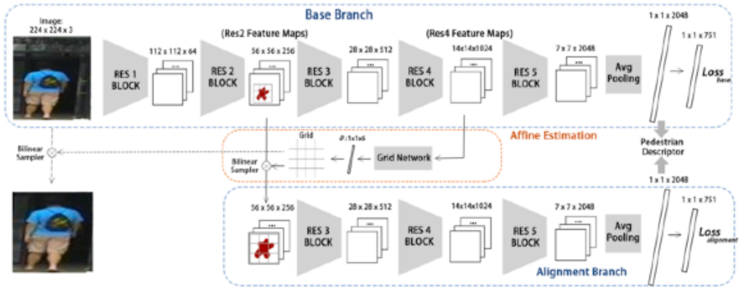
➡ A popular framework for learning invariance from data

➡ STNs can support invariant recognition by transforming all input images to a common pose

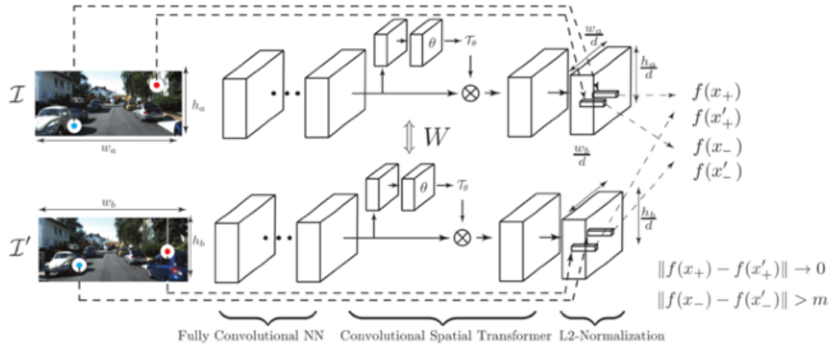


Transforming CNN feature maps?





Zheng et al. (2018)

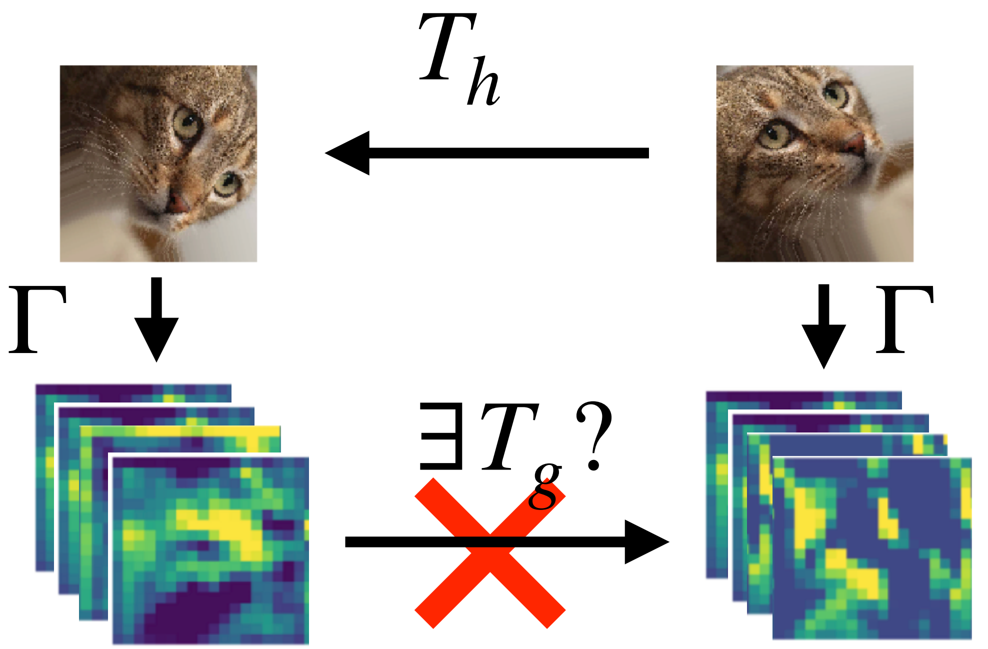


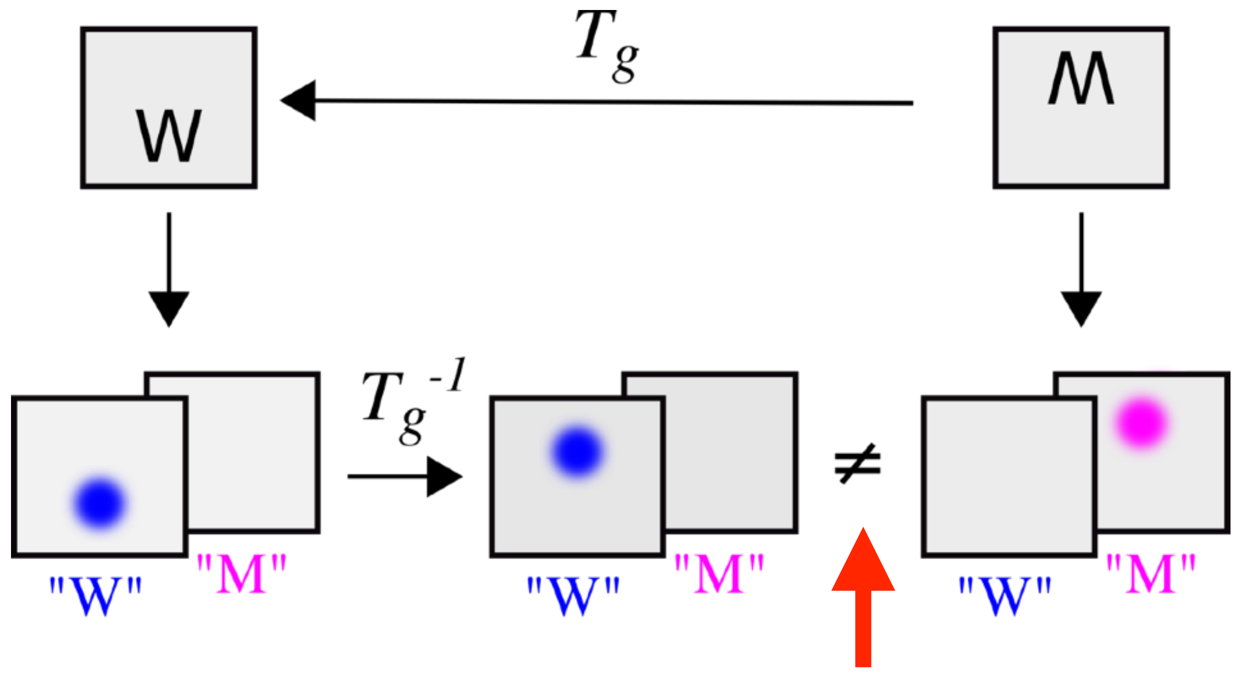
Choy et al. (2016)

- ➡ More complex features are useful but…
- ➡ … can invariance still be achieved if transforming CNN feature maps?

STNs that transform feature maps do not support invariant recognition

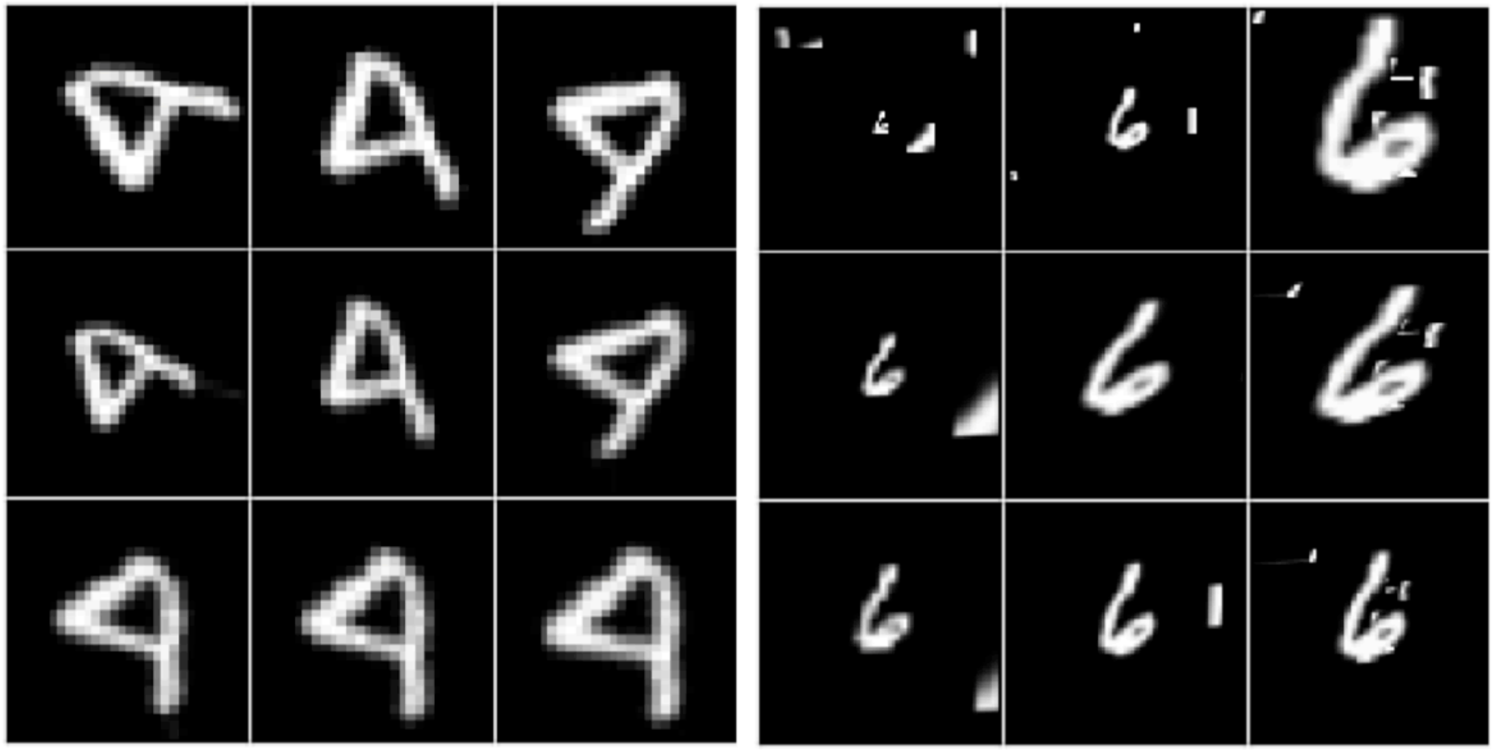
Full proof: ***arXiv:2004.14716***





- ➡ We prove that a spatial transformation is, **not enough to align the CNN feature maps** of an original and transformed image
- ➡ E.g. a rotation of an image typically results in a shift in **which feature channels** respond the strongest, which cannot be corrected by a spatial transformation

STNs that transform feature maps do not work as intended



Rotation

Scale

INPUT

Transform CNN feature maps →

Transform the input →

→ (!) Does not find a canonical orientation/scale

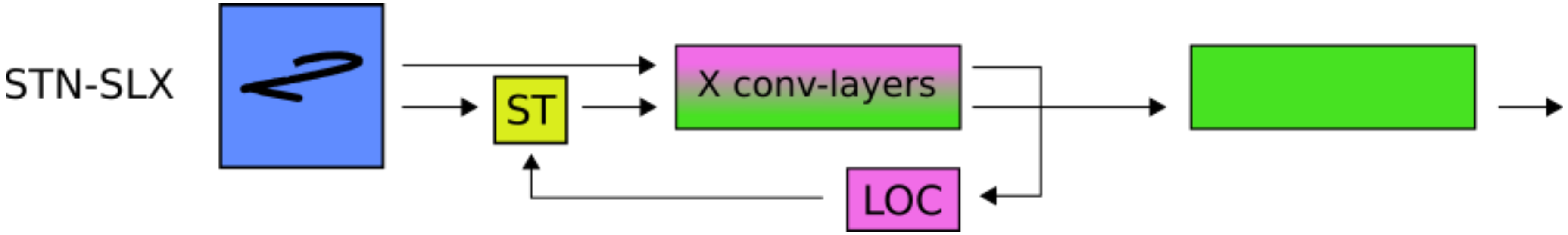
→ Finds a canonical orientation and scale

➡ An STN that transforms feature maps does not learn to transform objects to a **canonical orientation/scale**

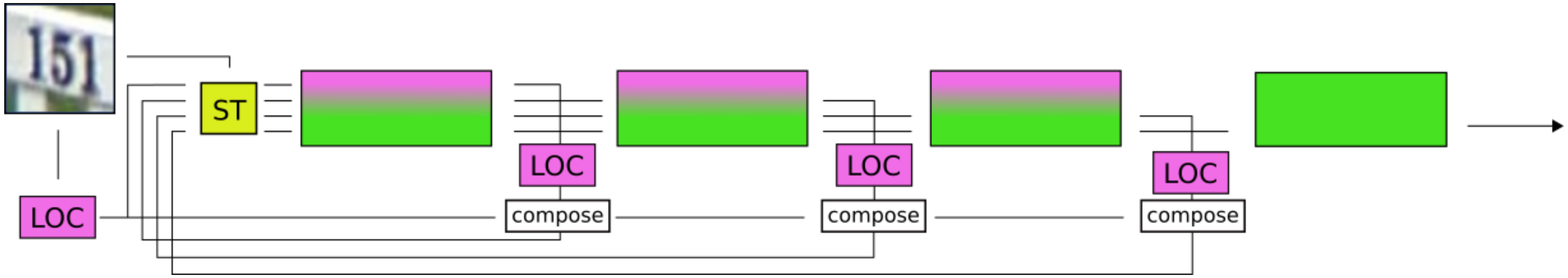
➡ This is because a rotation/scaling of CNN **feature maps** is **not enough** to align the feature maps of a translated image to those of the original

What are alternative options for using deeper features?

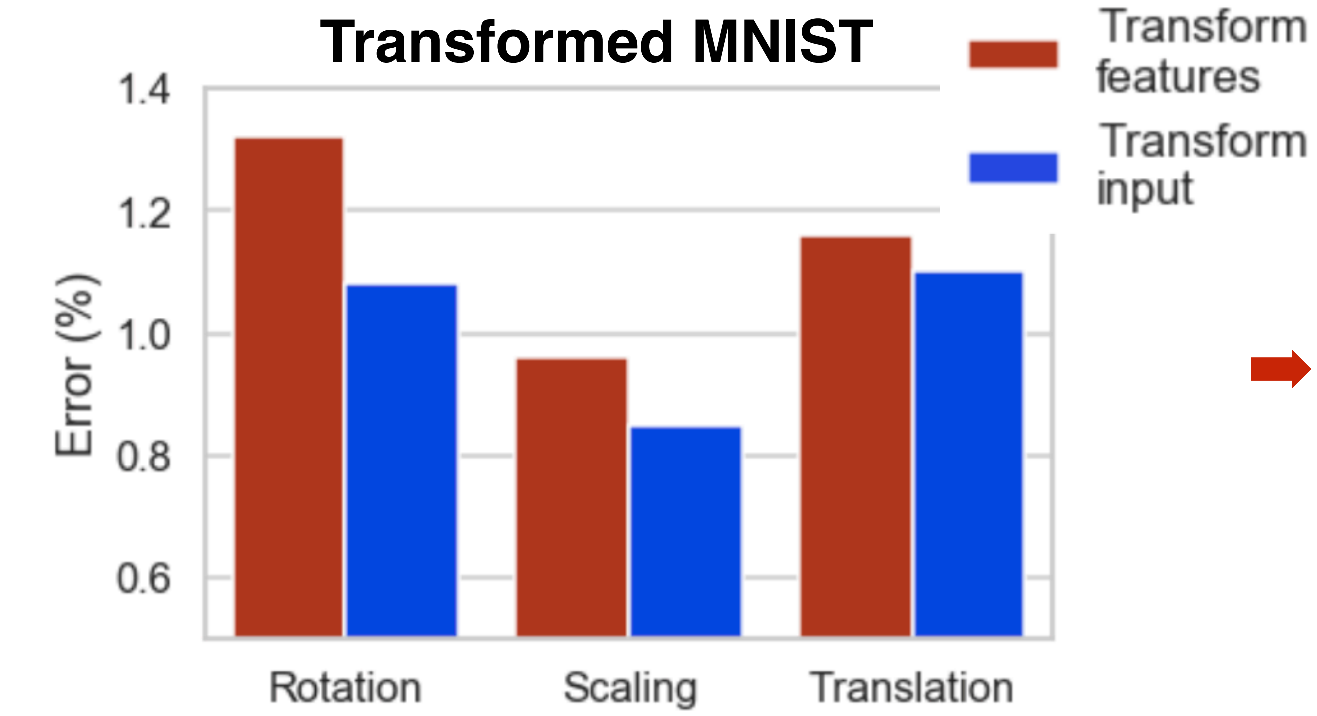
➡ Can **parameter sharing** between the localization and classification networks enable more stable training?



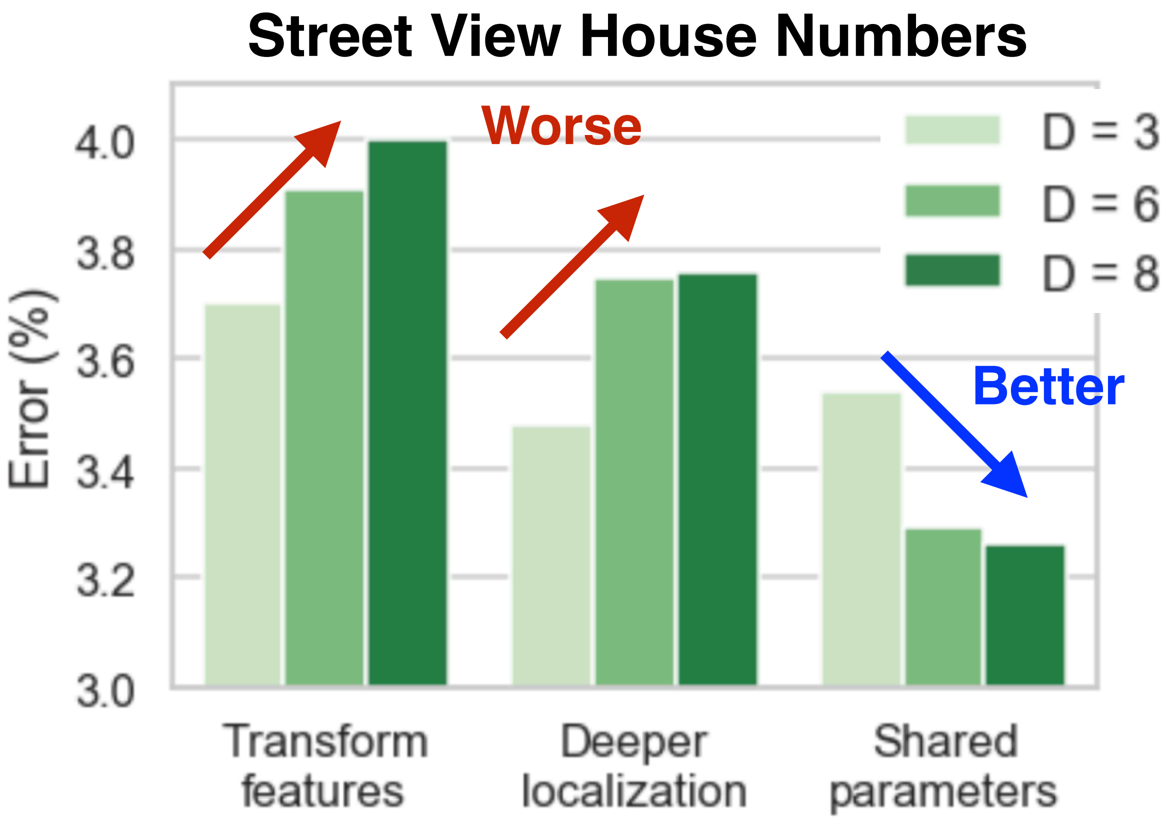
➡ Is **iterative image alignment** complementary to using deeper features?



Does it matter for performance? Yes!



➡ Transforming CNN feature maps **negatively impacts classification performance**



➡ Parameter **sharing enables** training **deeper localisation networks**

Summary

Full proof: ***arXiv:2004.14716***

- STNs that transform feature maps **do not enable invariant recognition**
- We present **a simple proof** and **an experimental evaluation** of the consequences of this result.
- Our experiments demonstrate the advantage of always **transforming the input**.
- We instead suggest **sharing parameters** between the classification and the localisation networks to enable training of deeper localization networks.
- Our results have implications also for **other approaches that spatially transform CNN feature maps**.