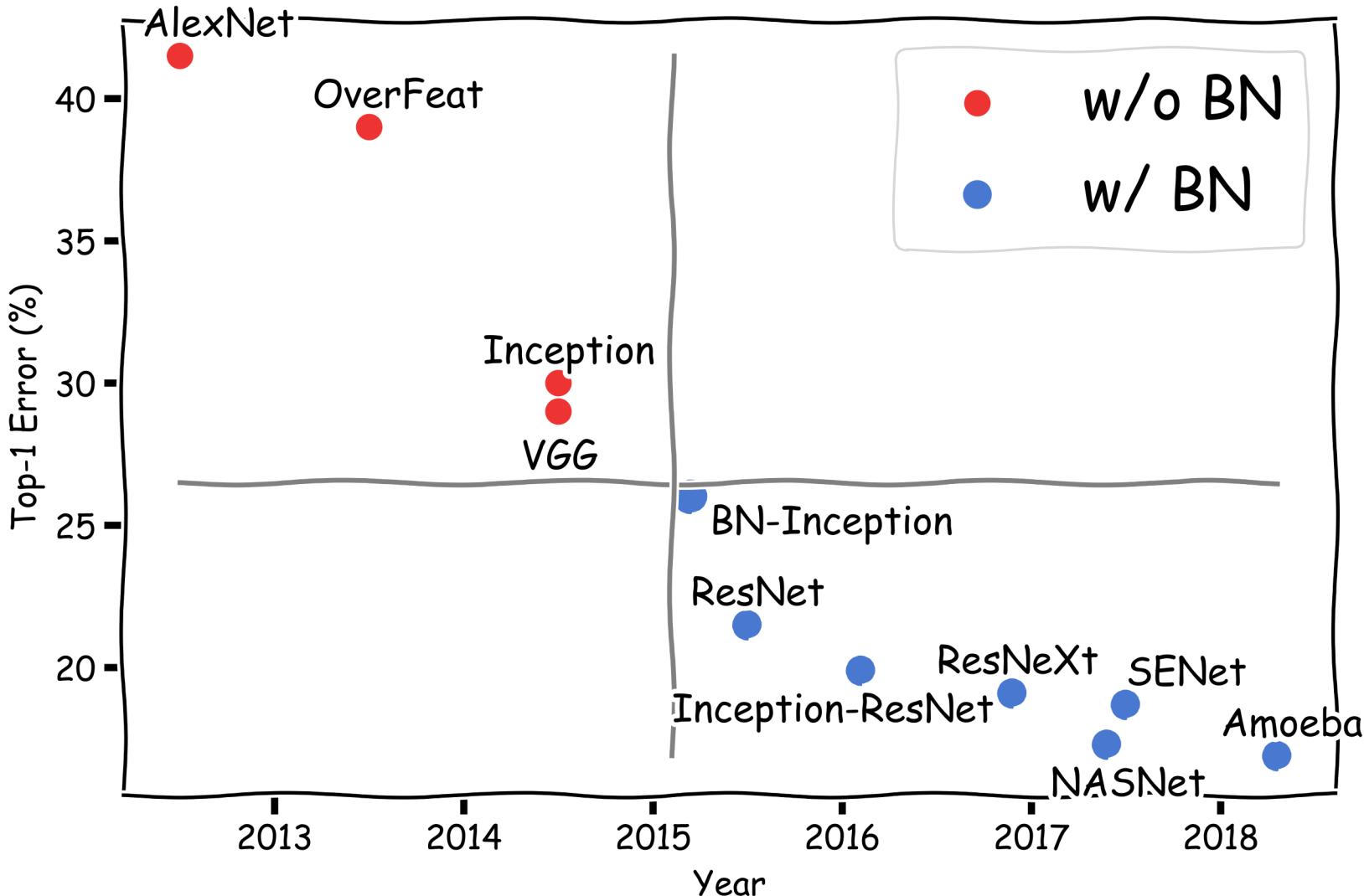


Group Normalization

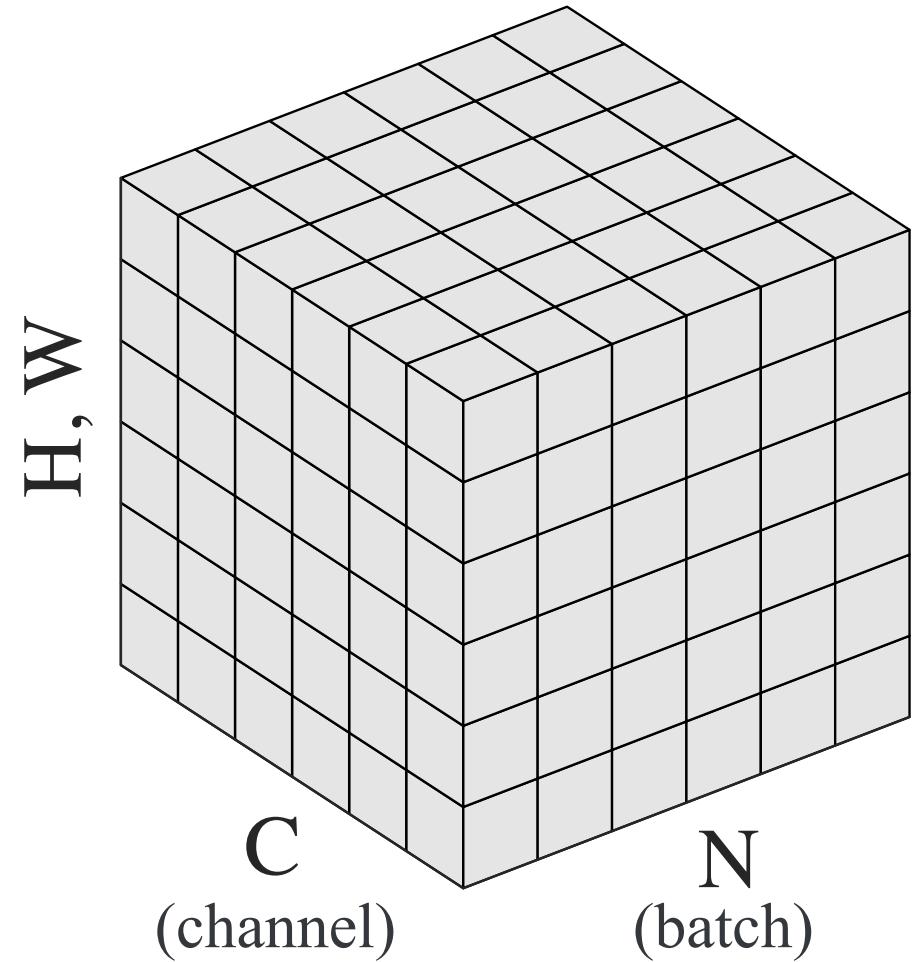
ECCV 2018, Munich

Yuxin Wu, Kaiming He
Facebook AI Research (FAIR)

Batch Normalization – a Milestone



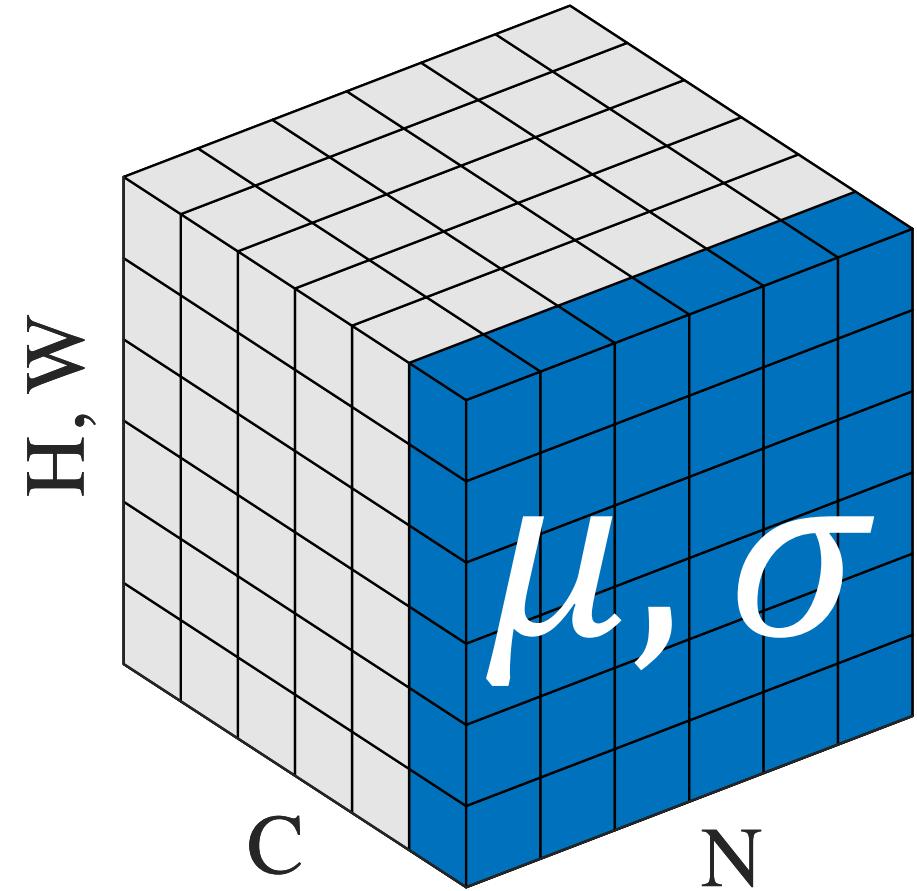
What's Batch Norm



What's Batch Norm

- Batch ...
- Normalization!

$$\hat{x} = \frac{x - \mu}{\sigma}$$

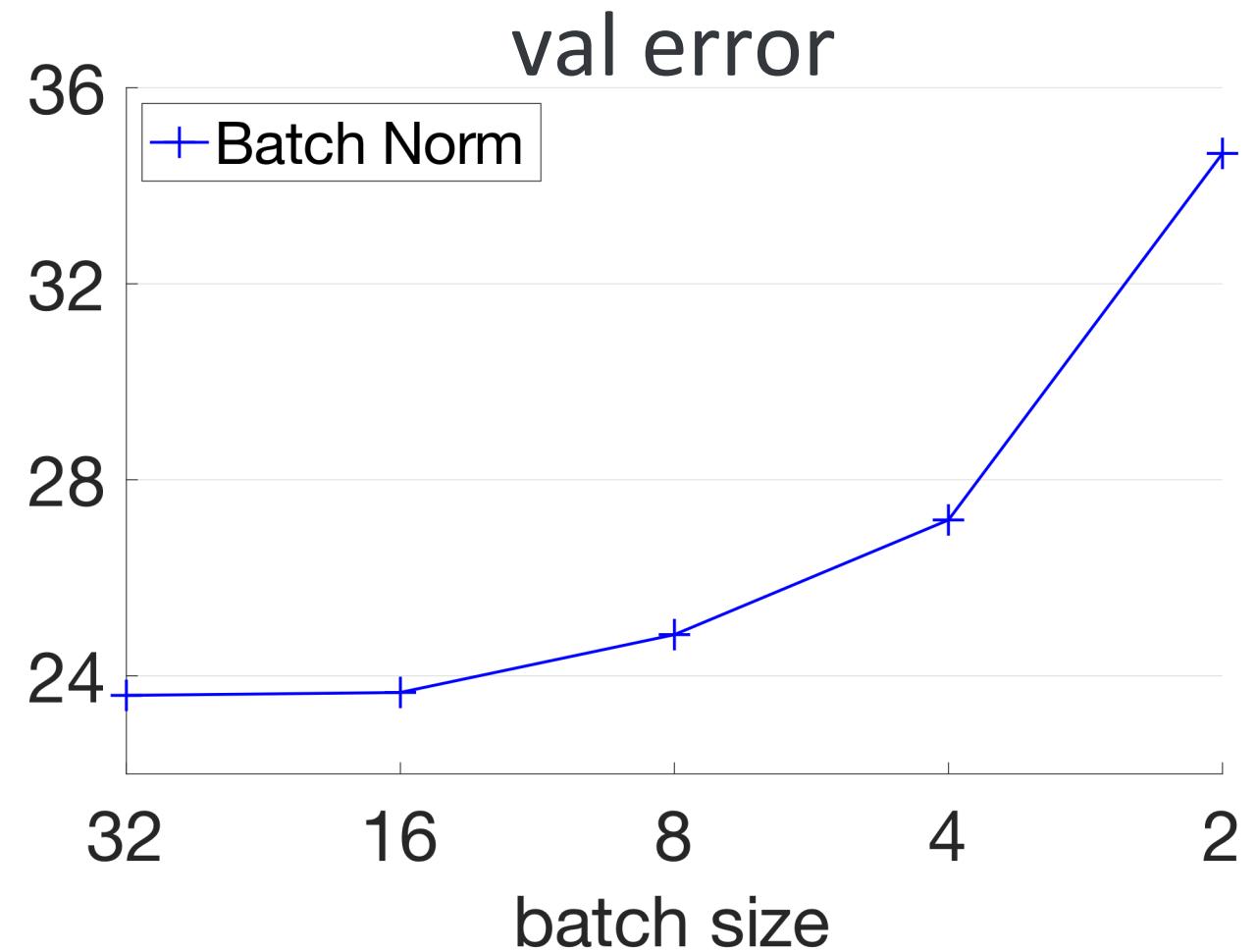


Batch: also source of drawbacks

- Small batch
 - large models
 - detection / segmentation / video classification ...

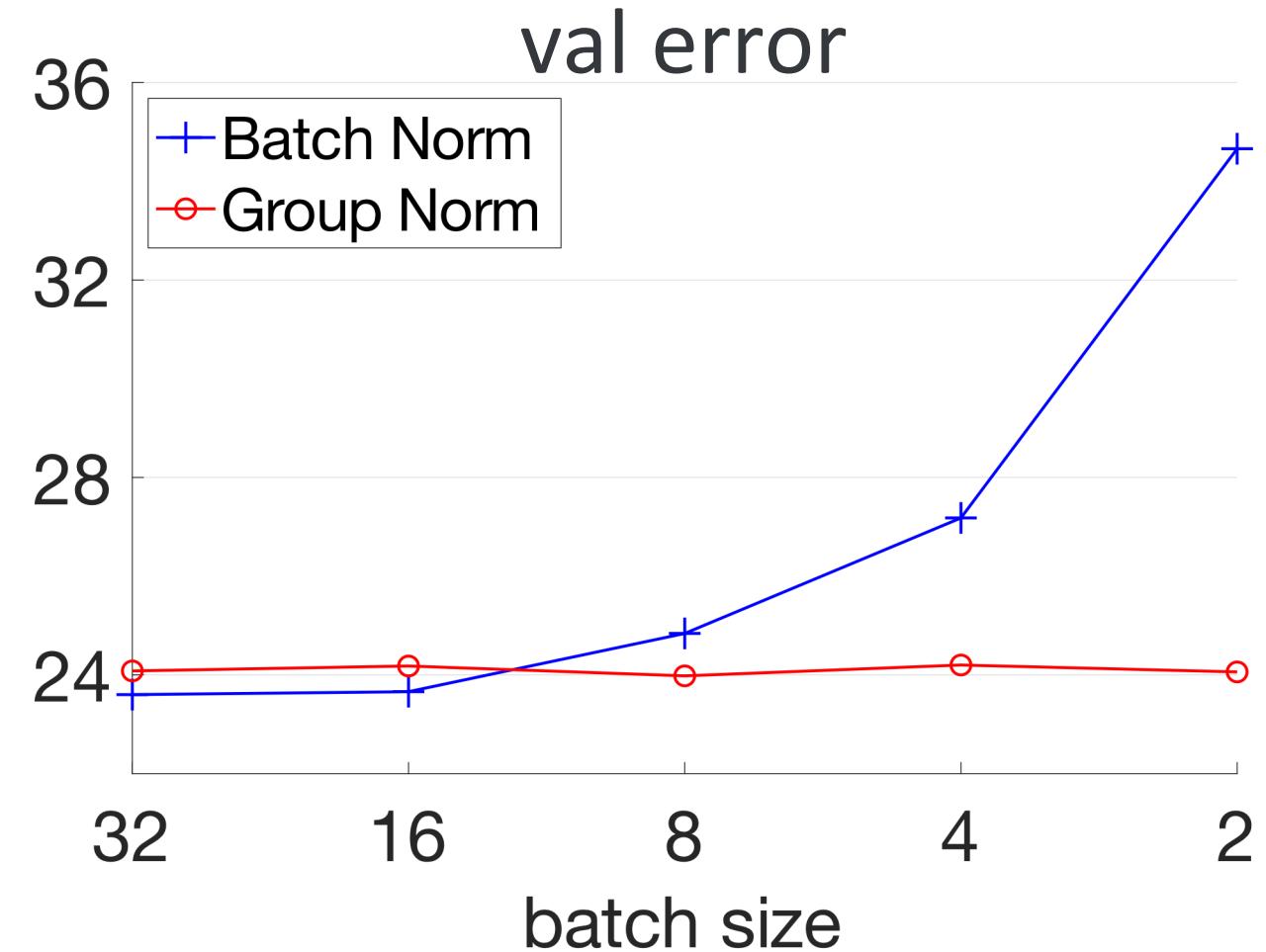
Batch: also source of drawbacks

- Small batch
- Varying batch

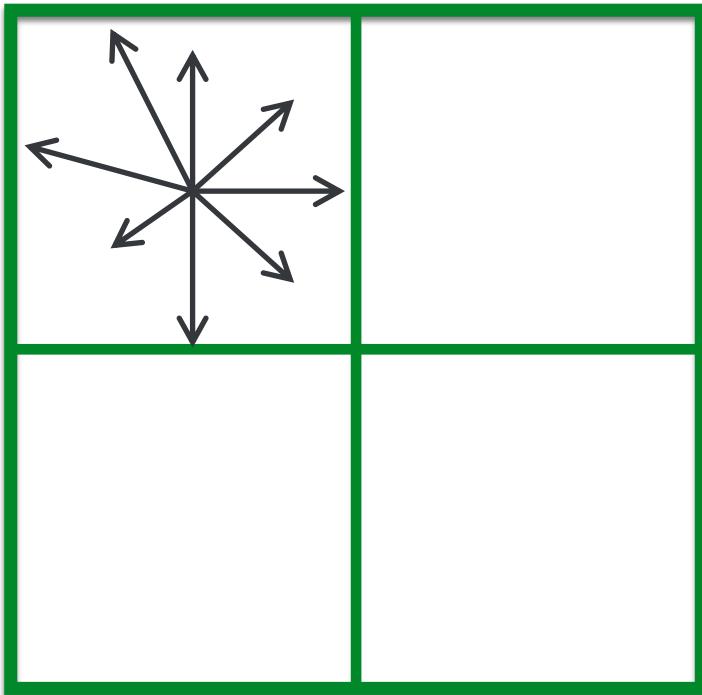


Our Method: Group Normalization

- GN is batch-independent
- ~~Small batch~~
- ~~Varying batch~~

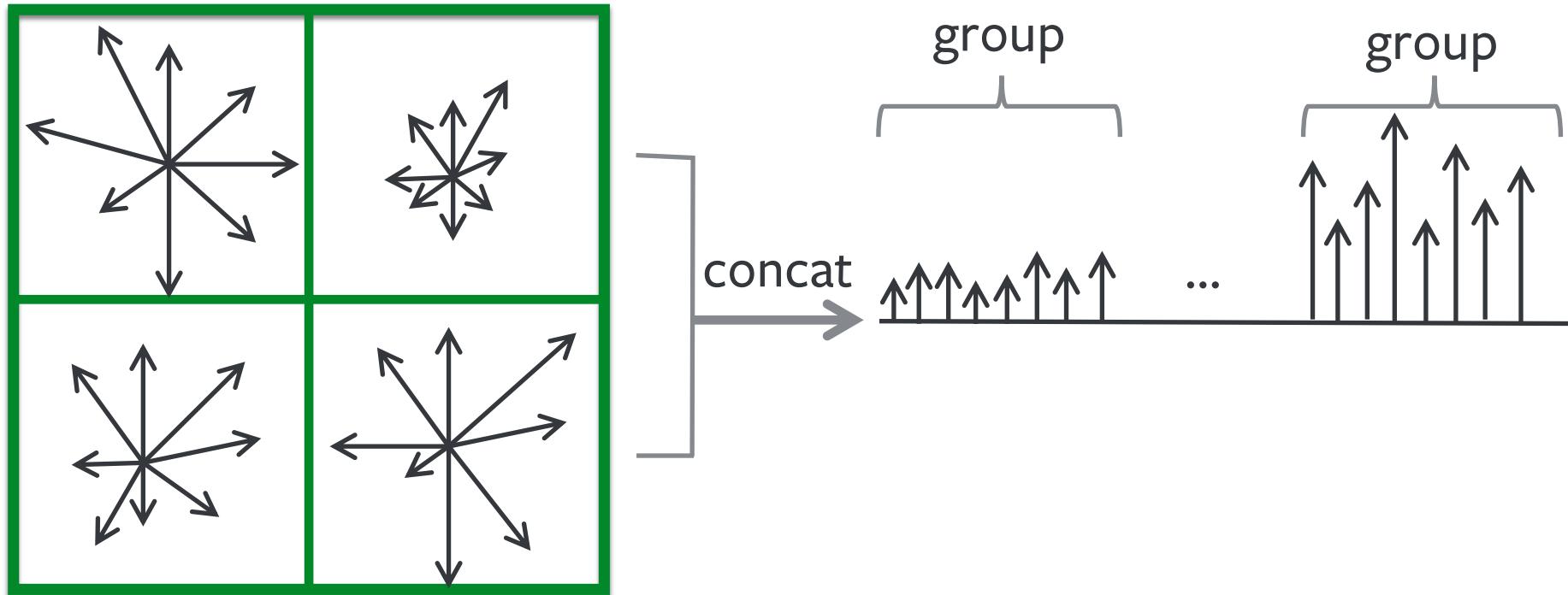


“Group Norm” in Retrospective: HOG/SIFT



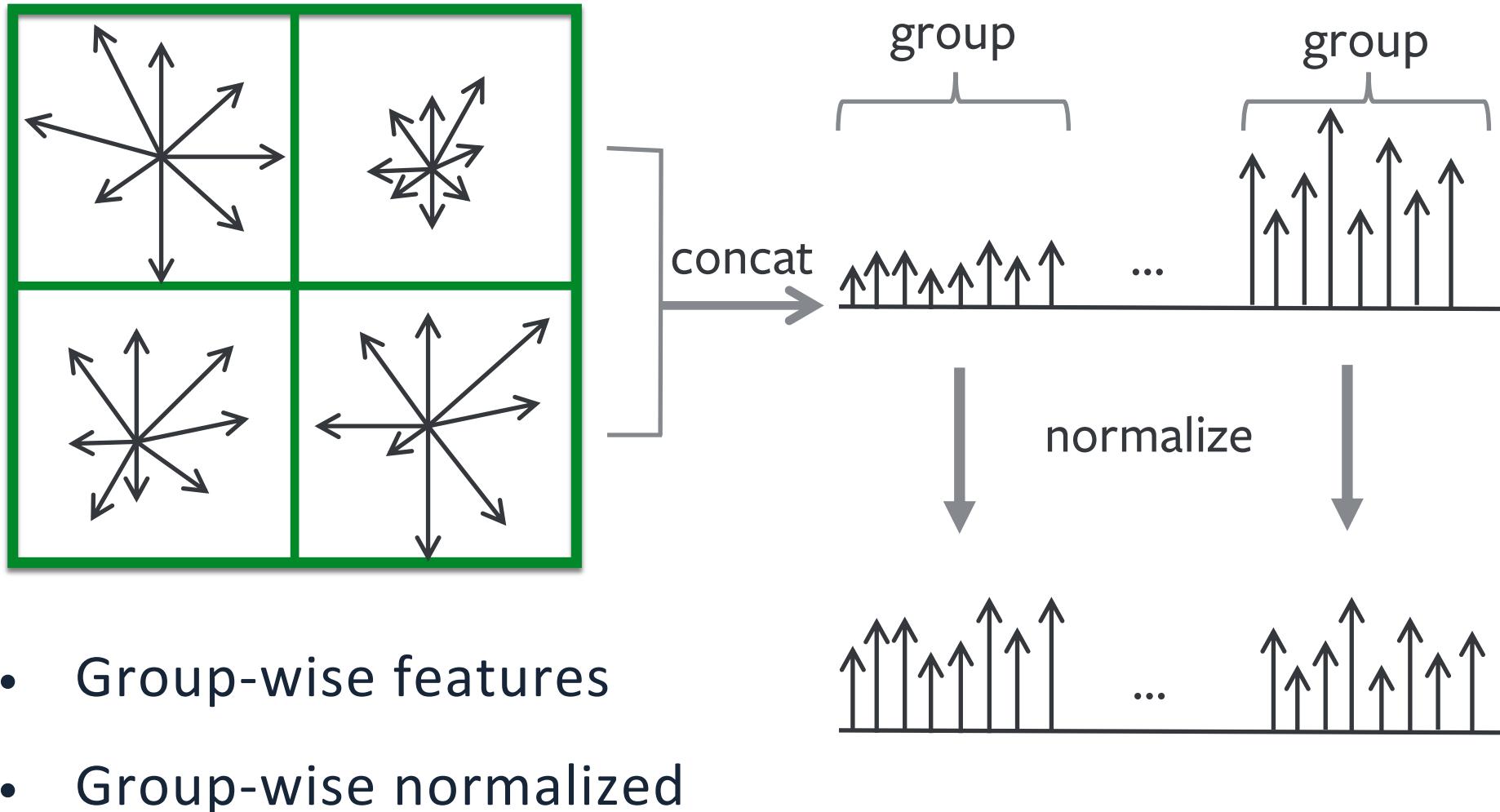
Dalal, Navneet, and Bill Triggs. "Histograms of oriented gradients for human detection." *Computer Vision and Pattern Recognition*, 2005
Lowe, David G. "Distinctive image features from scale-invariant keypoints." *International journal of computer vision* 60.2 (2004)

“Group Norm” in Retrospective: HOG/SIFT

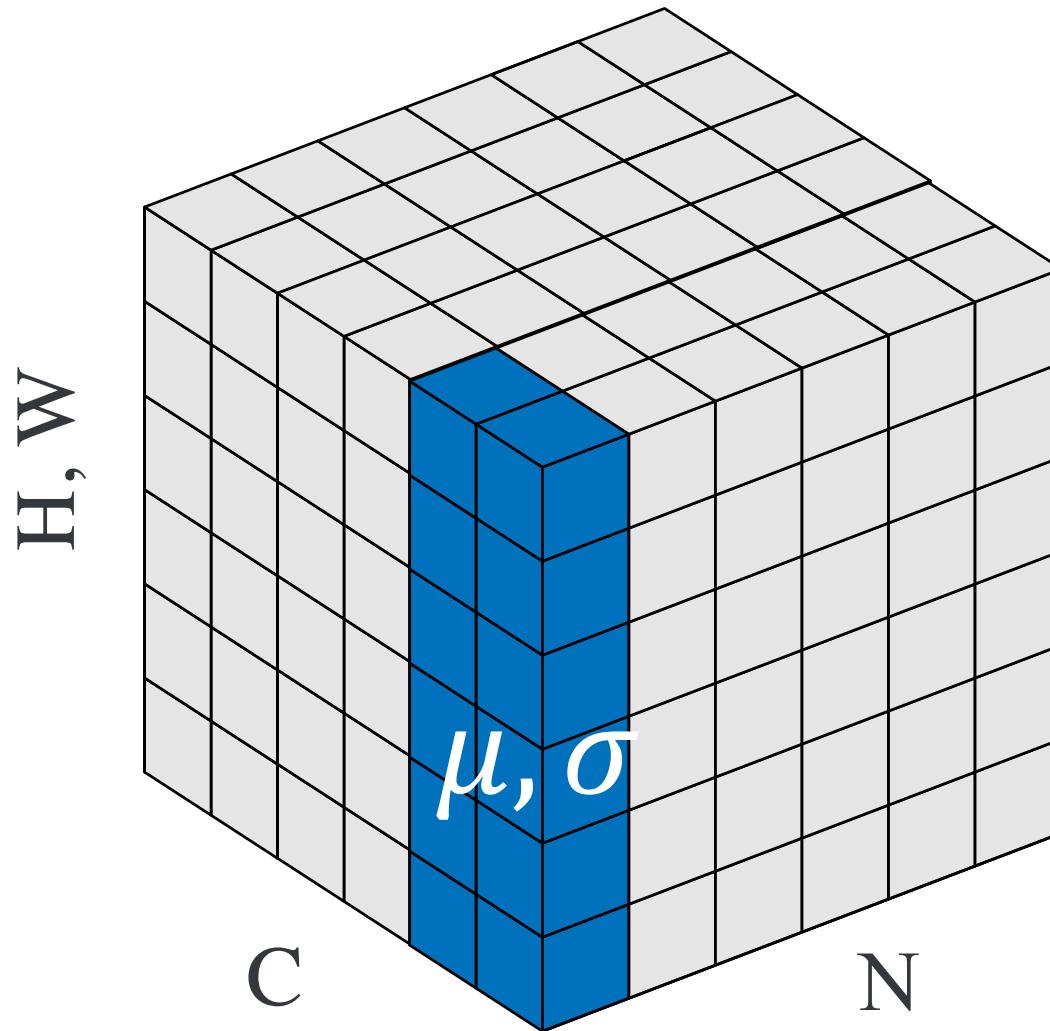


- Group-wise features

“Group Norm” in Retrospective: HOG/SIFT

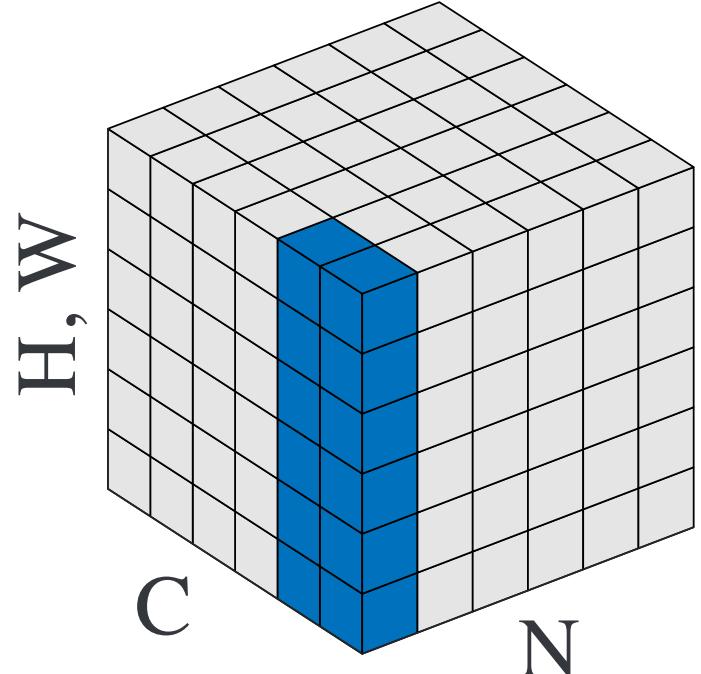


What's Group Norm

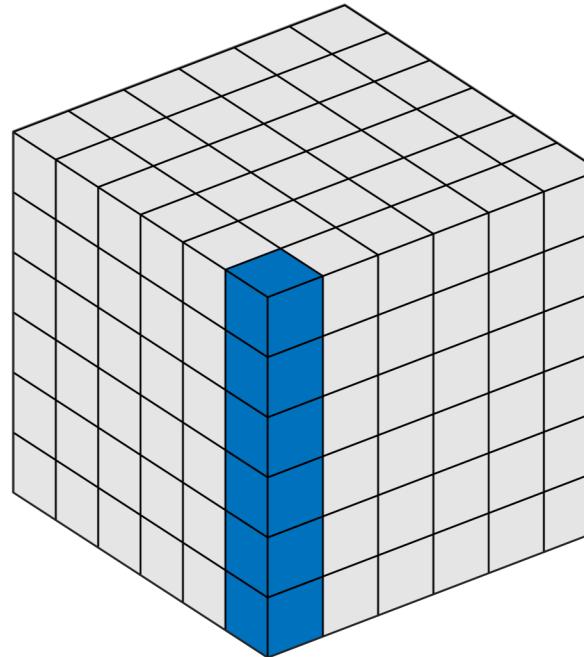


$$\hat{x} = \frac{x - \mu}{\sigma}$$

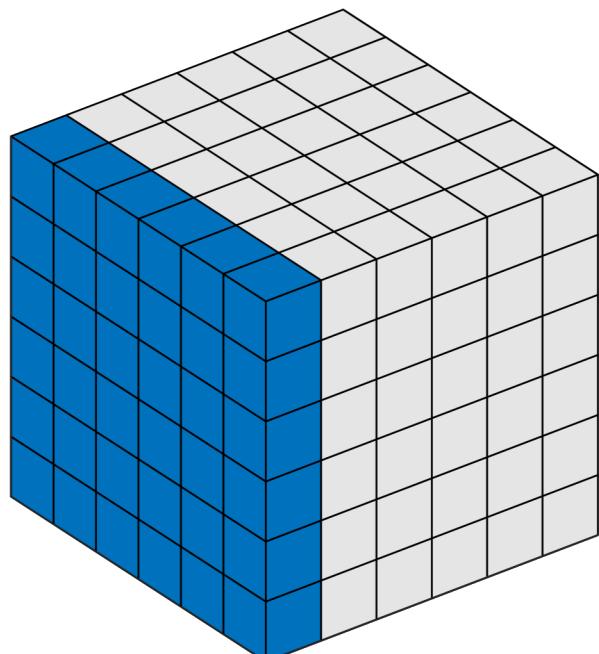
Group Norm



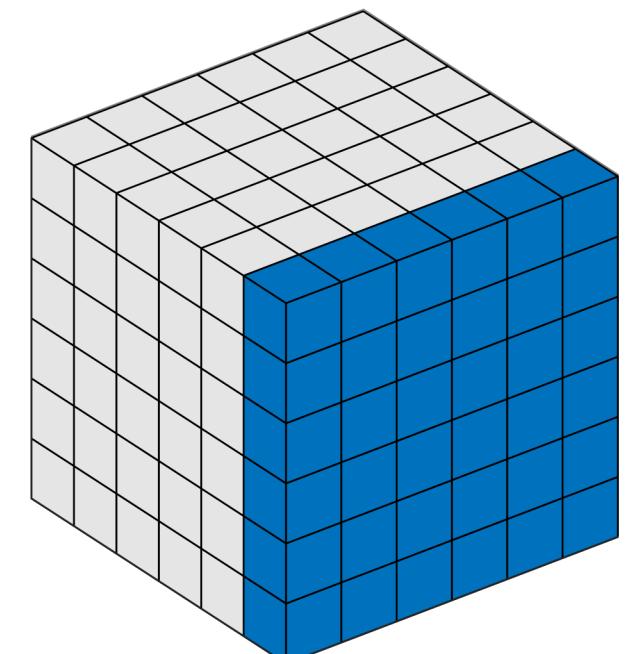
Instance Norm



Layer Norm



Batch Norm



all channels
in one group

one channel
per group

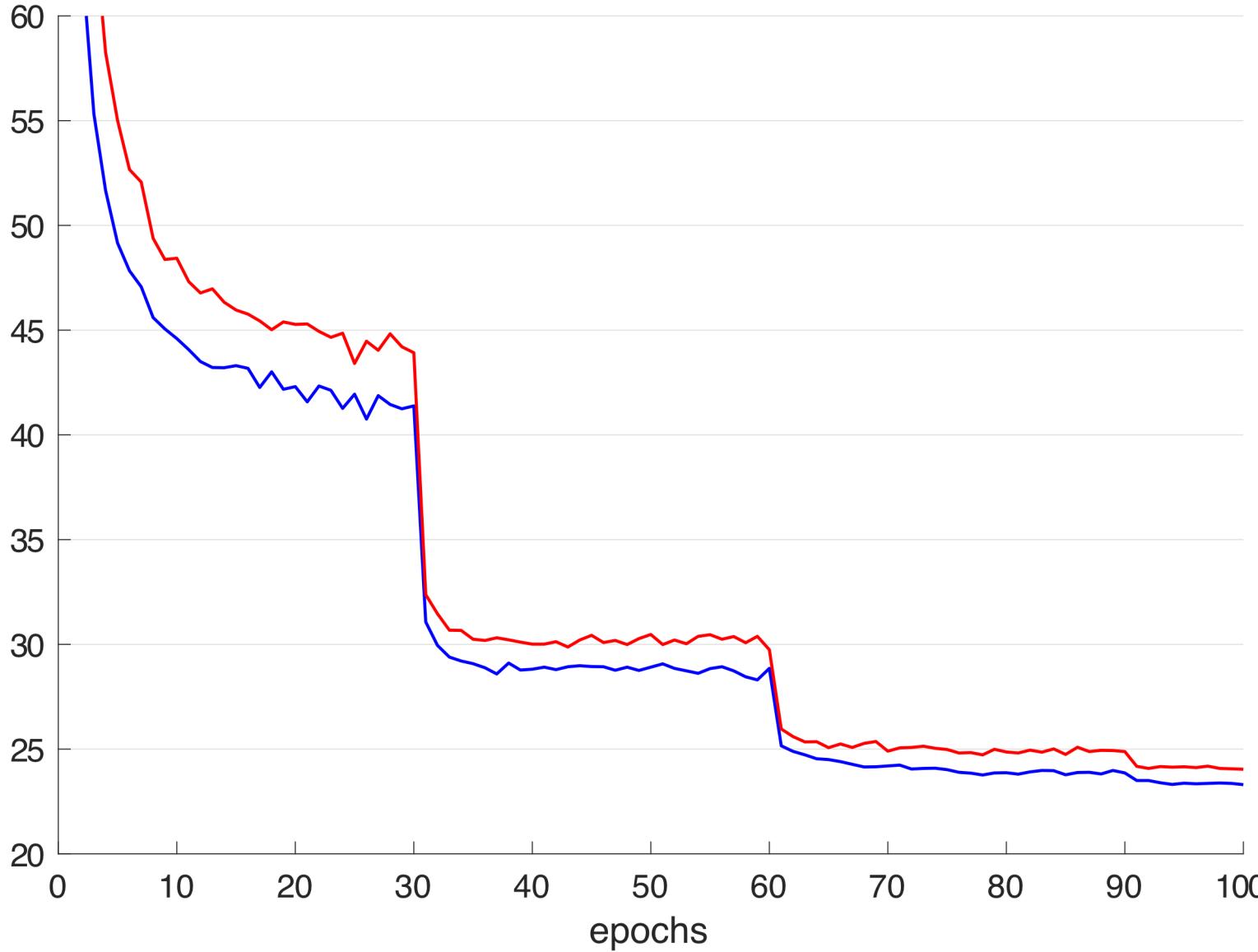
one channel

one image per batch

Experiments: ImageNet Classification

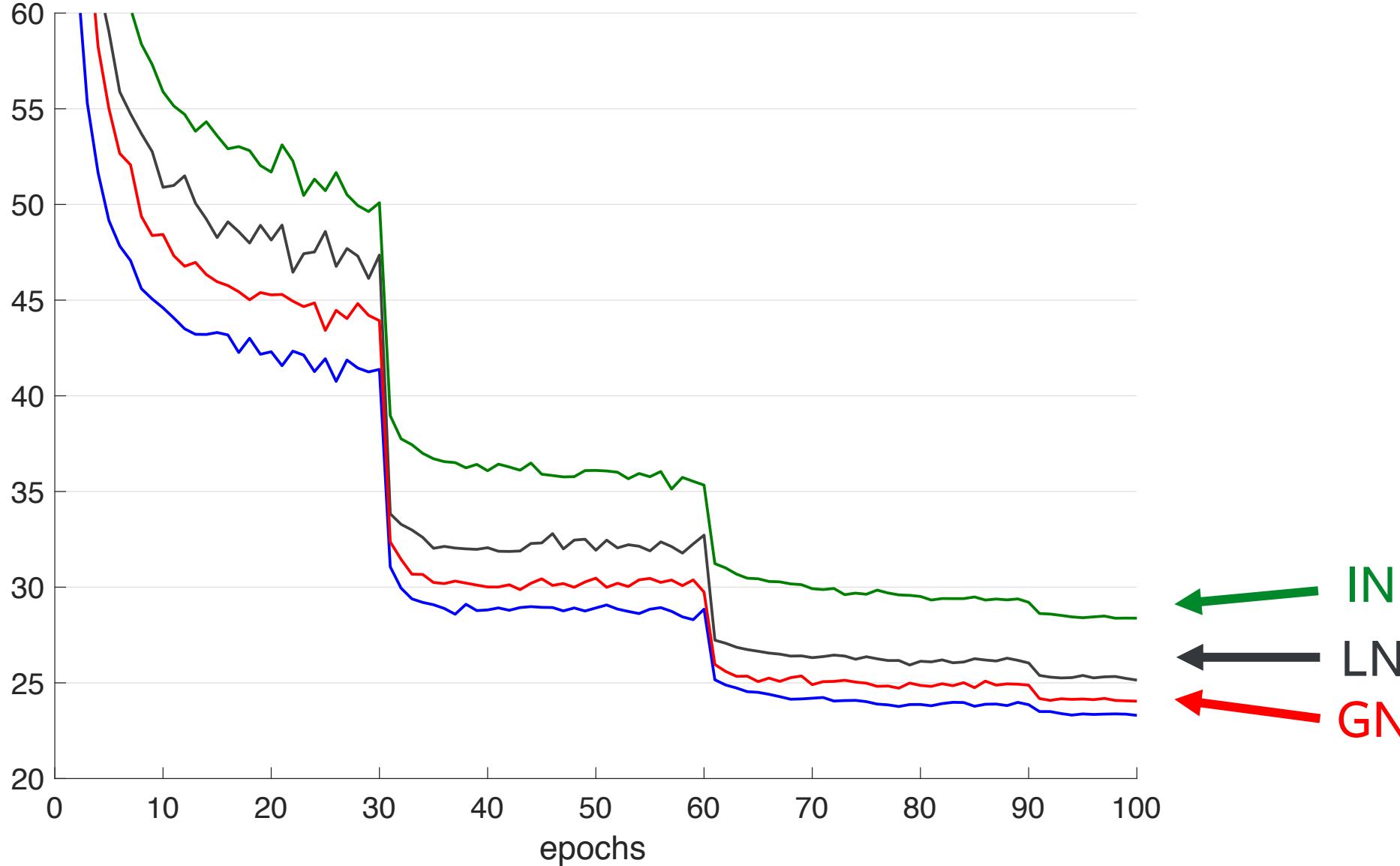
Standard Batch Size

val error (%), batch=32



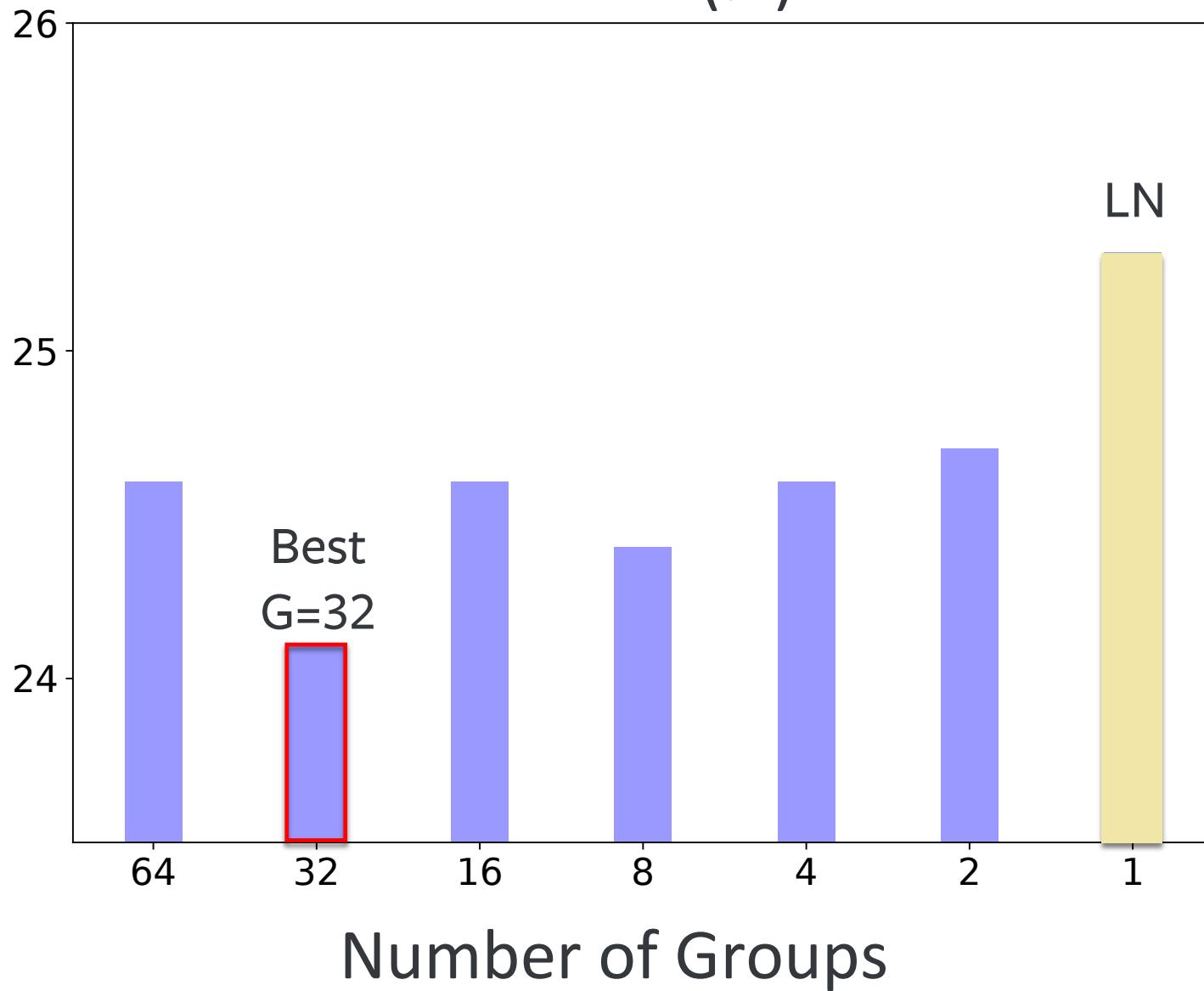
GN: 24.1
BN: 23.6

val error (%), batch=32



IN
LN
GN

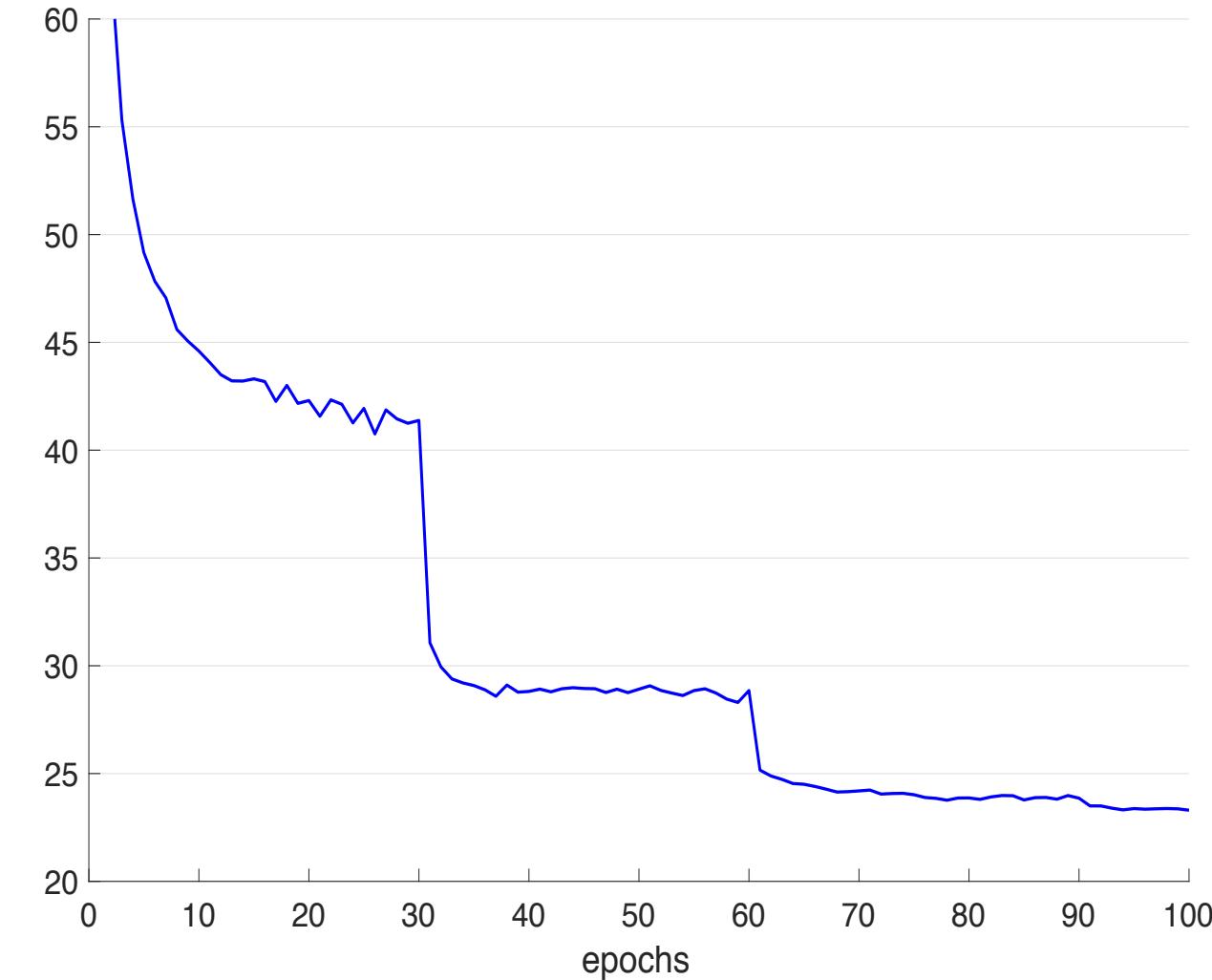
val error (%)



Experiments: ImageNet Classification

Small Batch Size

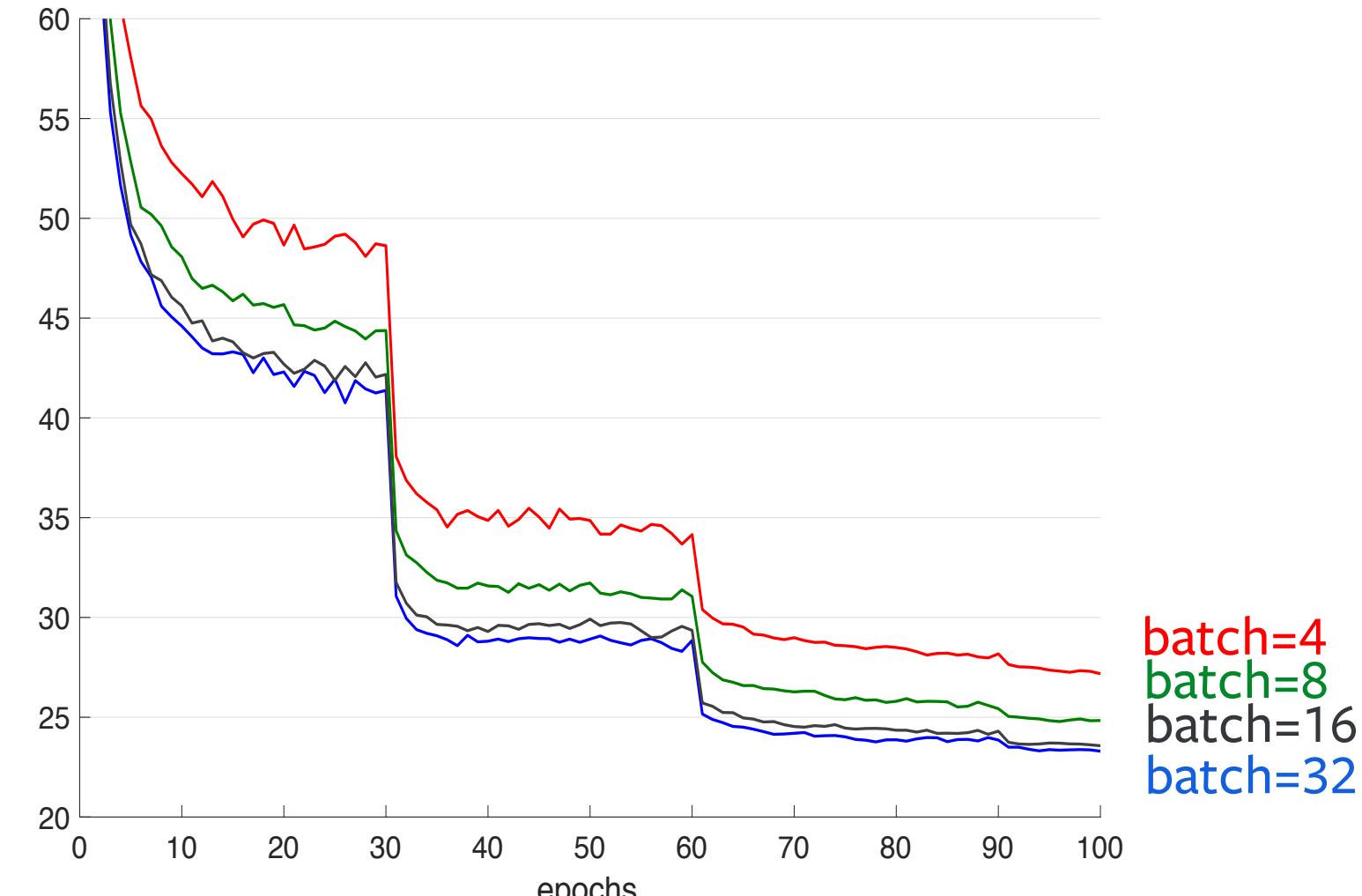
val error



batch=32

Batch Norm 😞

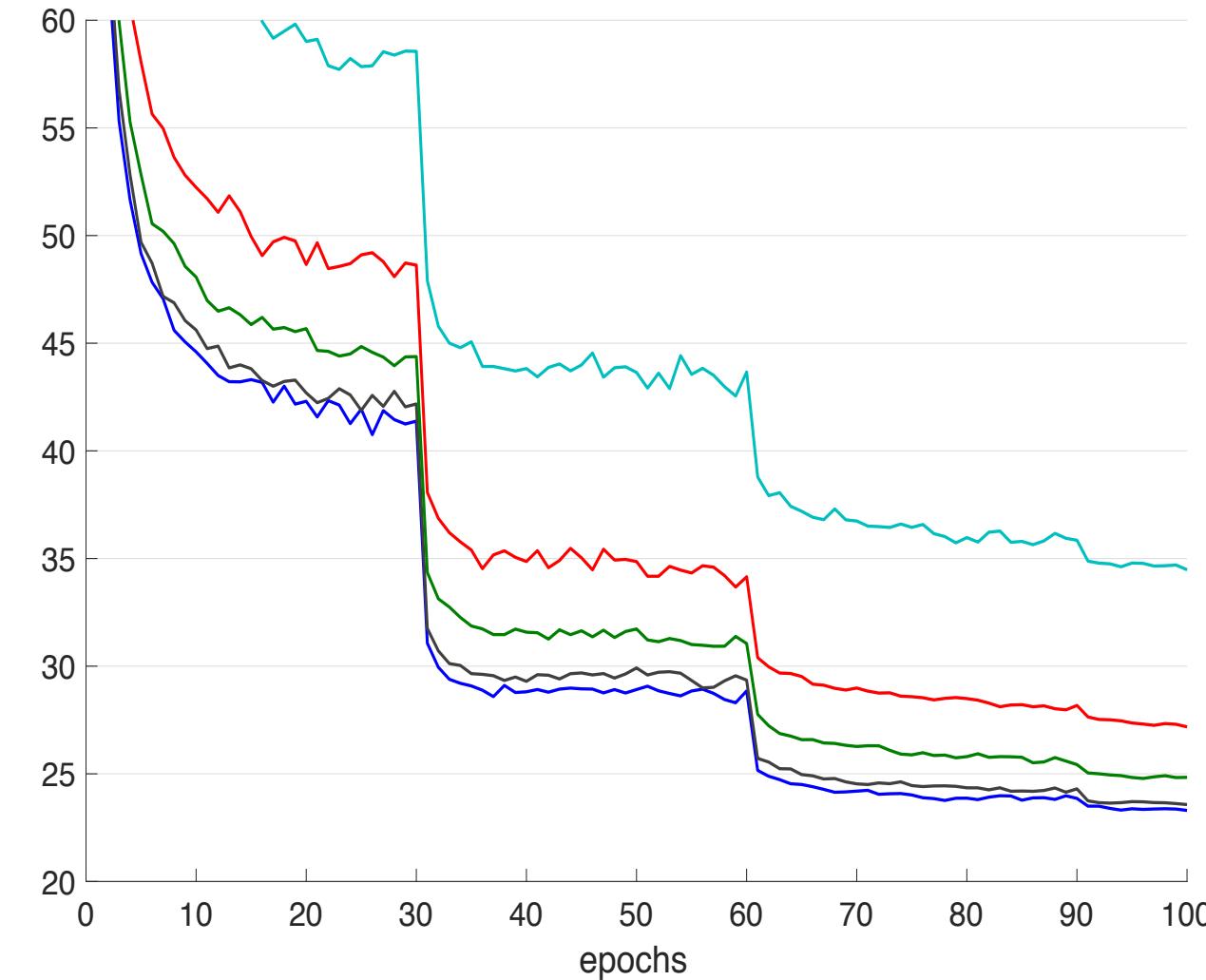
val error



batch=4
batch=8
batch=16
batch=32

Batch Norm 😞

val error



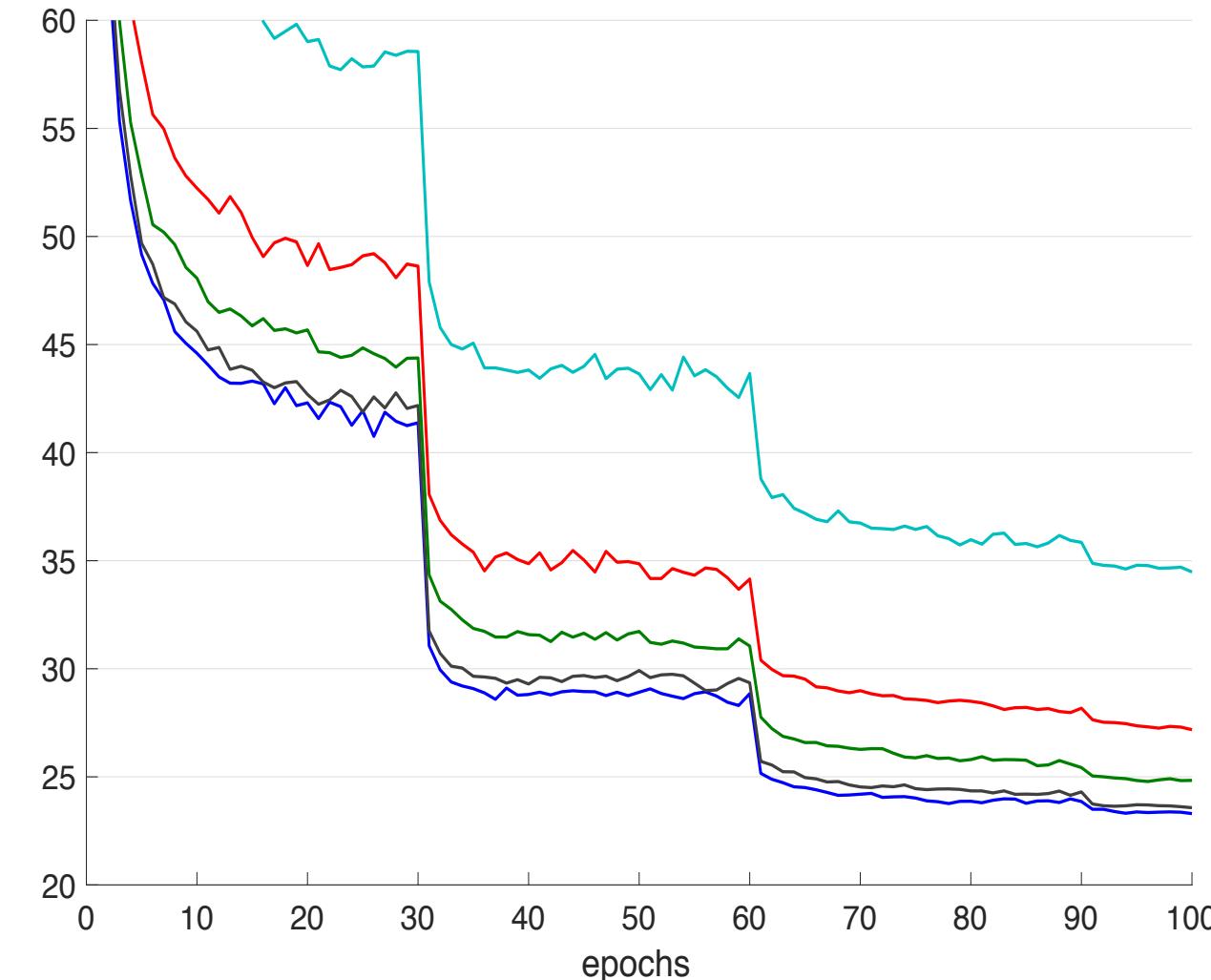
batch=2

11%

batch=32

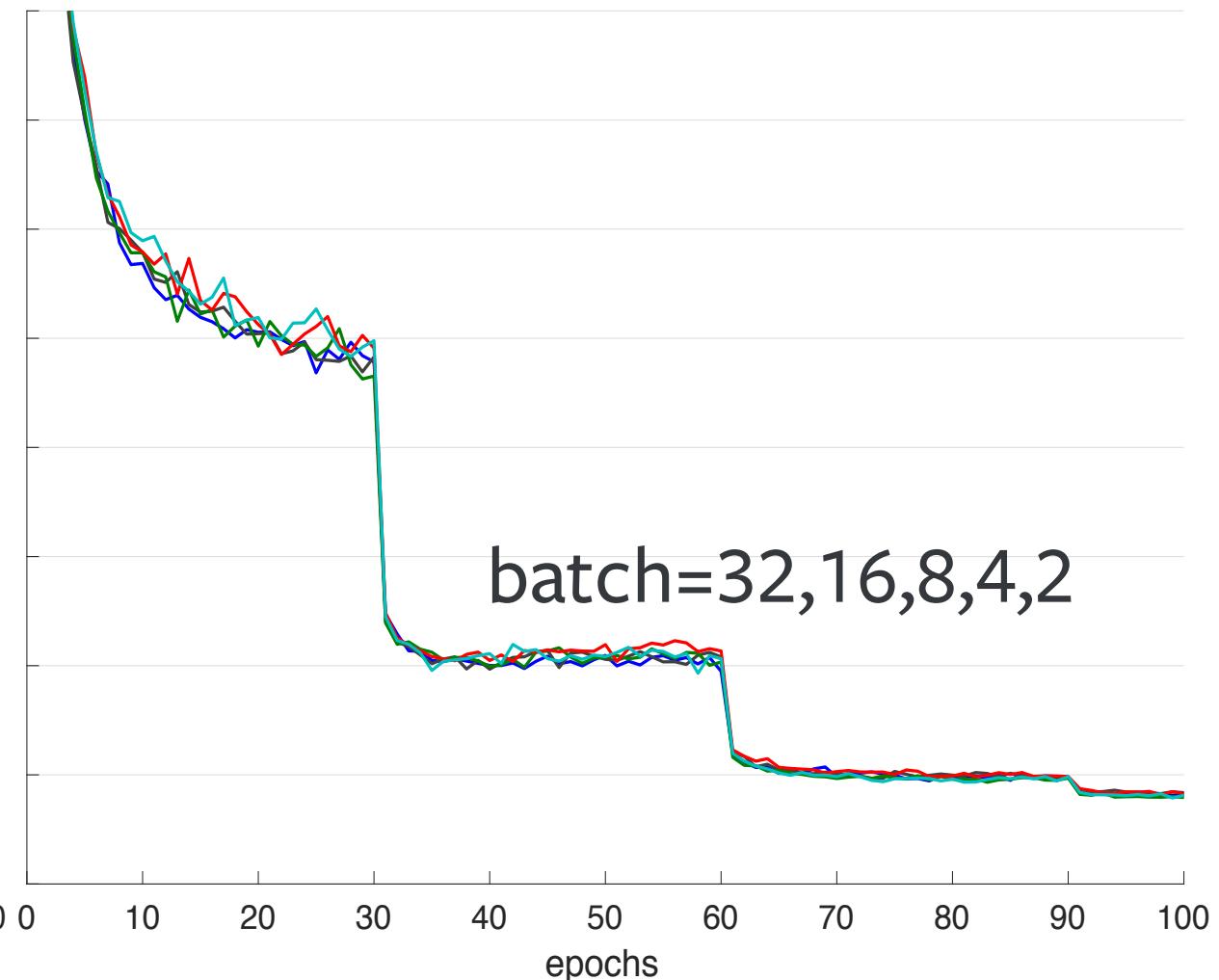
Batch Norm 😞

val error



Batch Norm 😞

val error



batch=32,16,8,4,2

Group Norm 😊
curves match

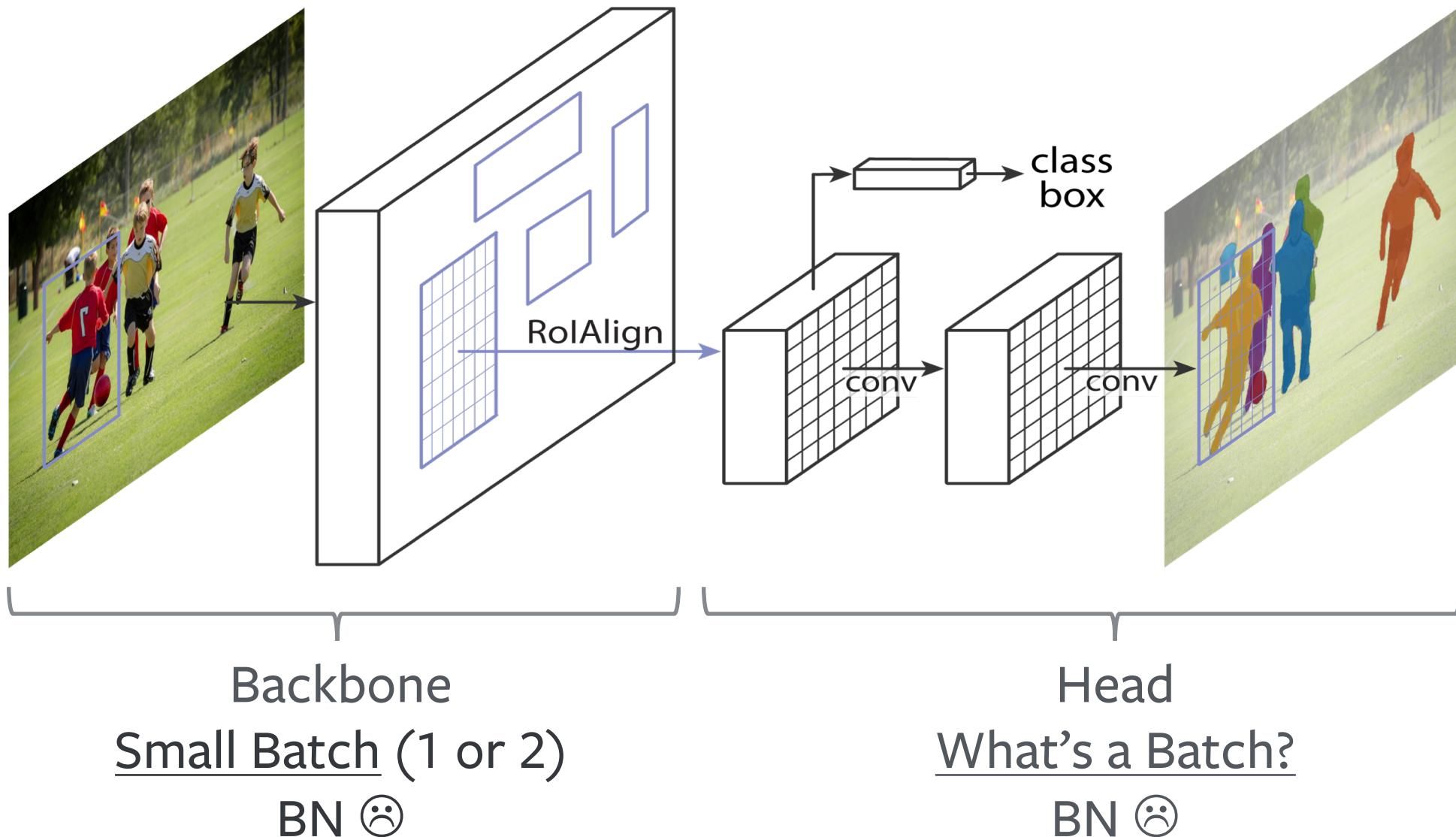
Batch: also source of drawbacks

- Small batch
- Varying batch
 - train vs. test
 - pre-train vs. fine-tune
 - backbone vs. head

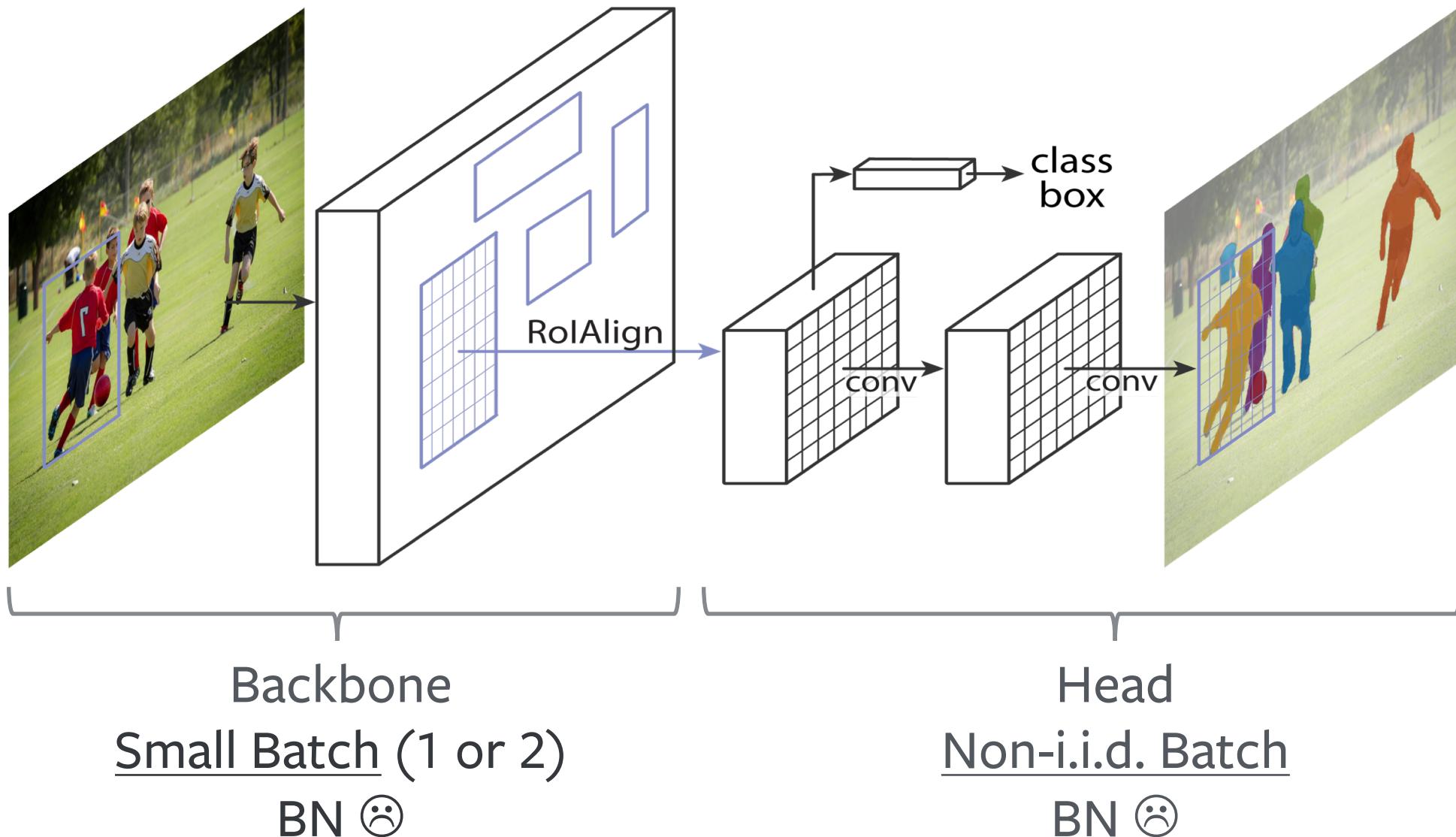
Experiments: Object Detection

What's a Batch?

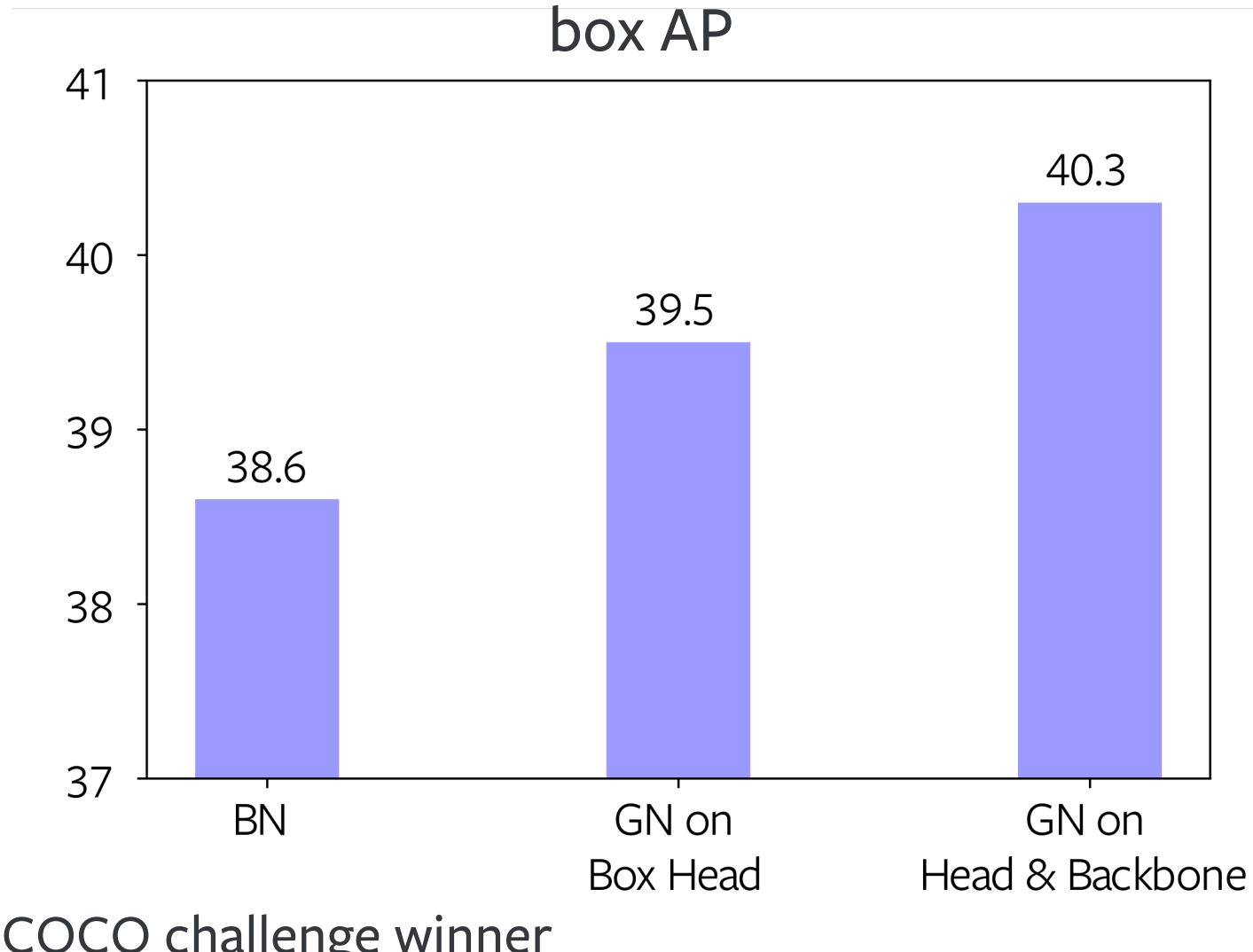
Mask R-CNN



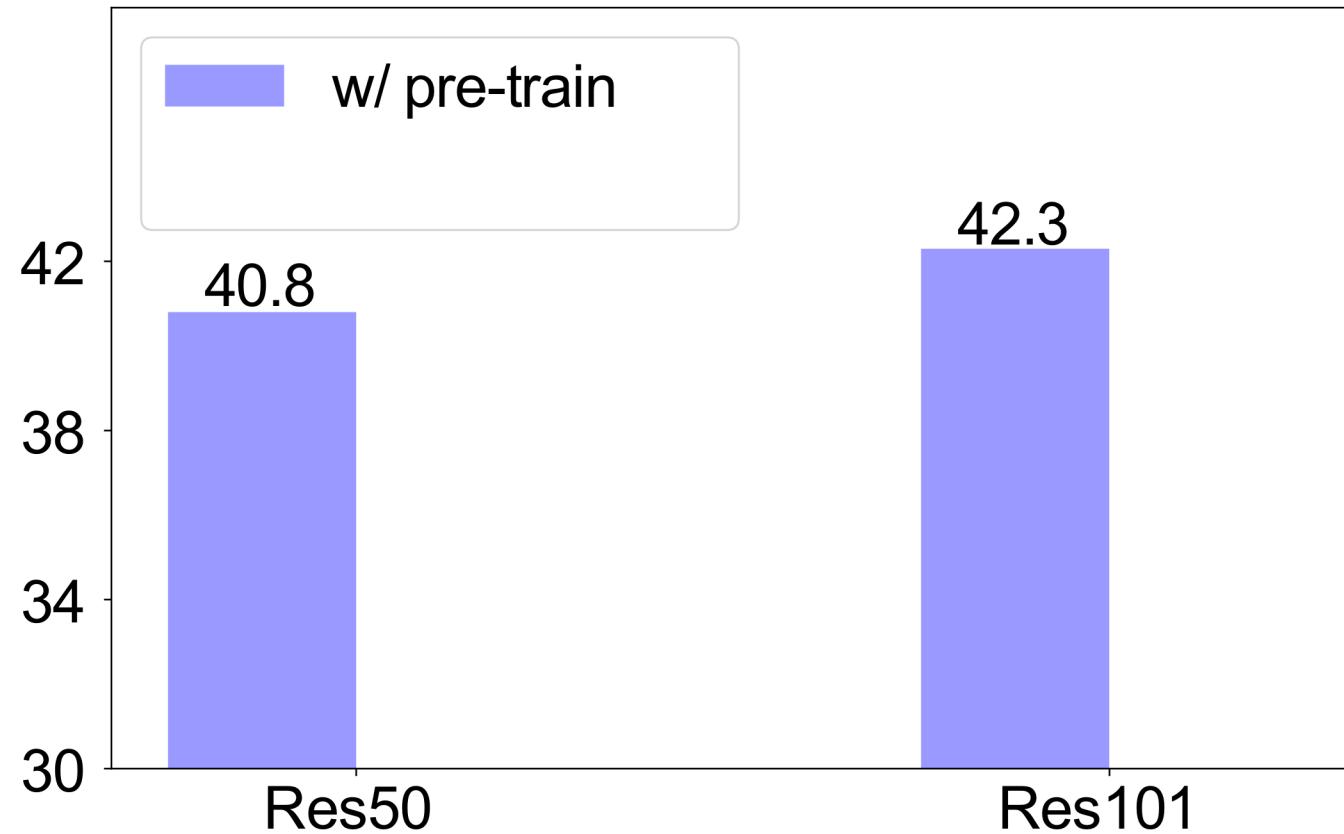
Mask R-CNN



Mask R-CNN on COCO

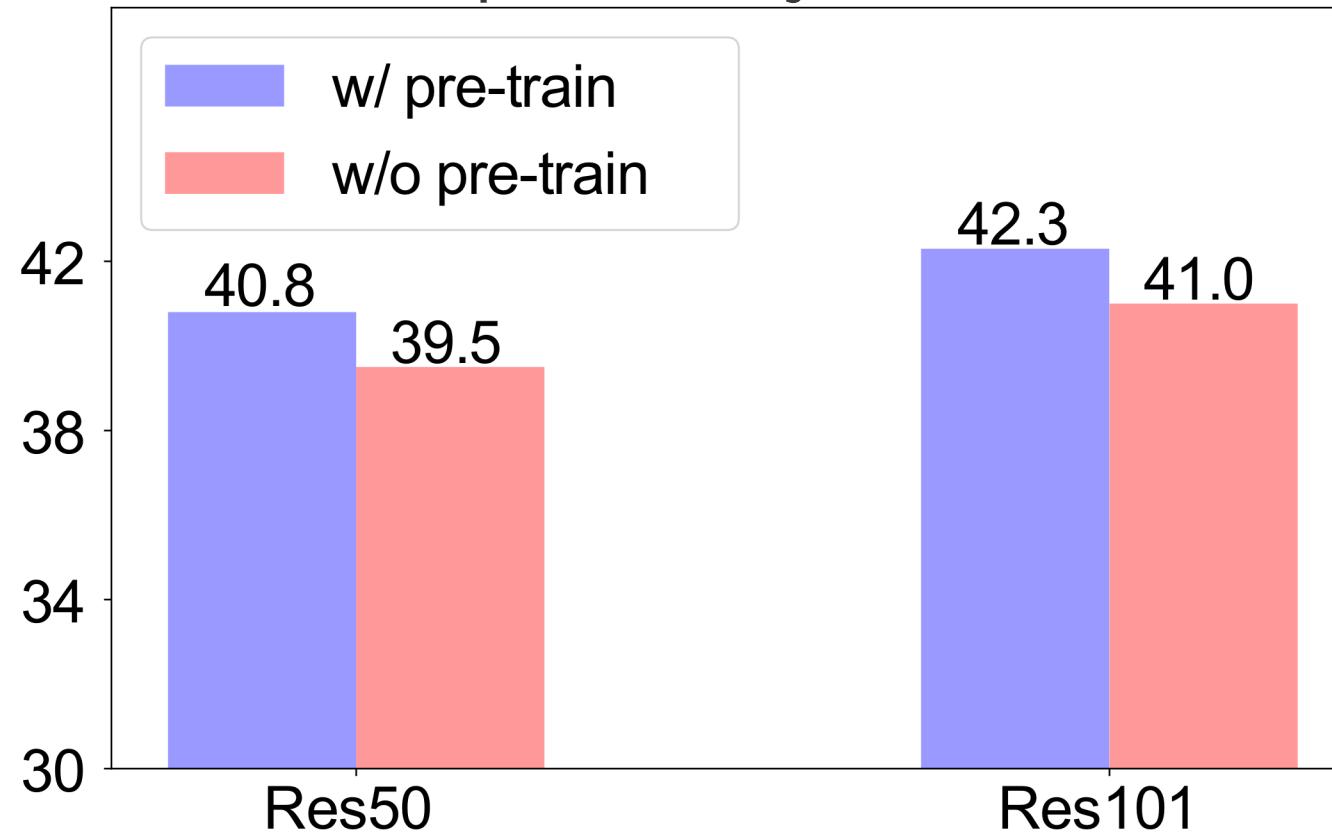


GN Enables Training Mask R-CNN From Scratch



GN Enables Training Mask R-CNN From Scratch

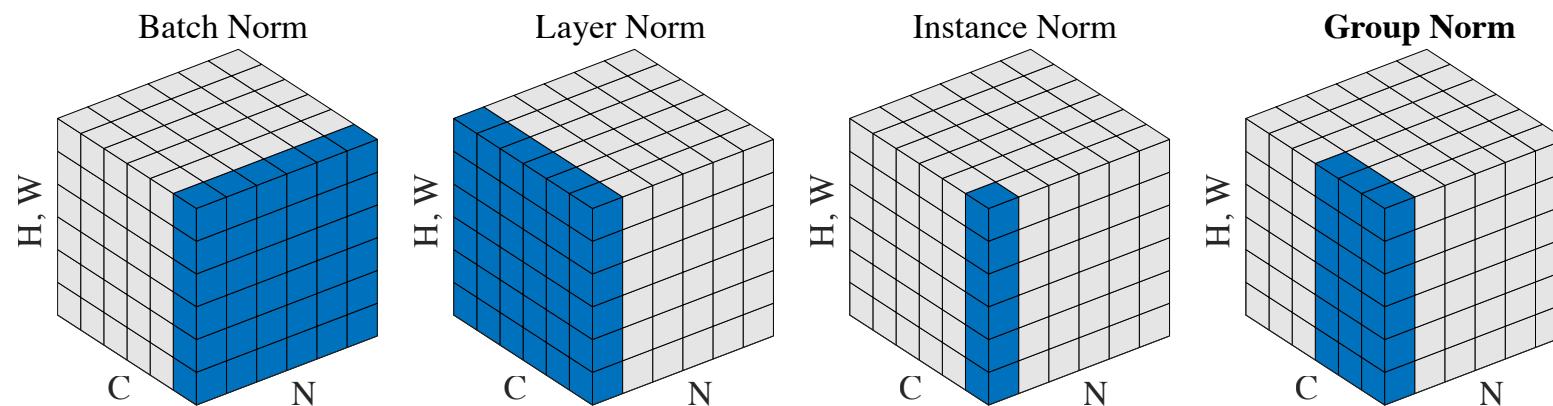
preliminary results



More to come: 50.9 AP w/o pre-training

Conclusion

- normalization matters
- “batch” is not always ideal
- channels can be grouped, and have substructures



Code: <https://github.com/facebookresearch/Detectron/tree/master/projects/GN>