ConvNeXt V2

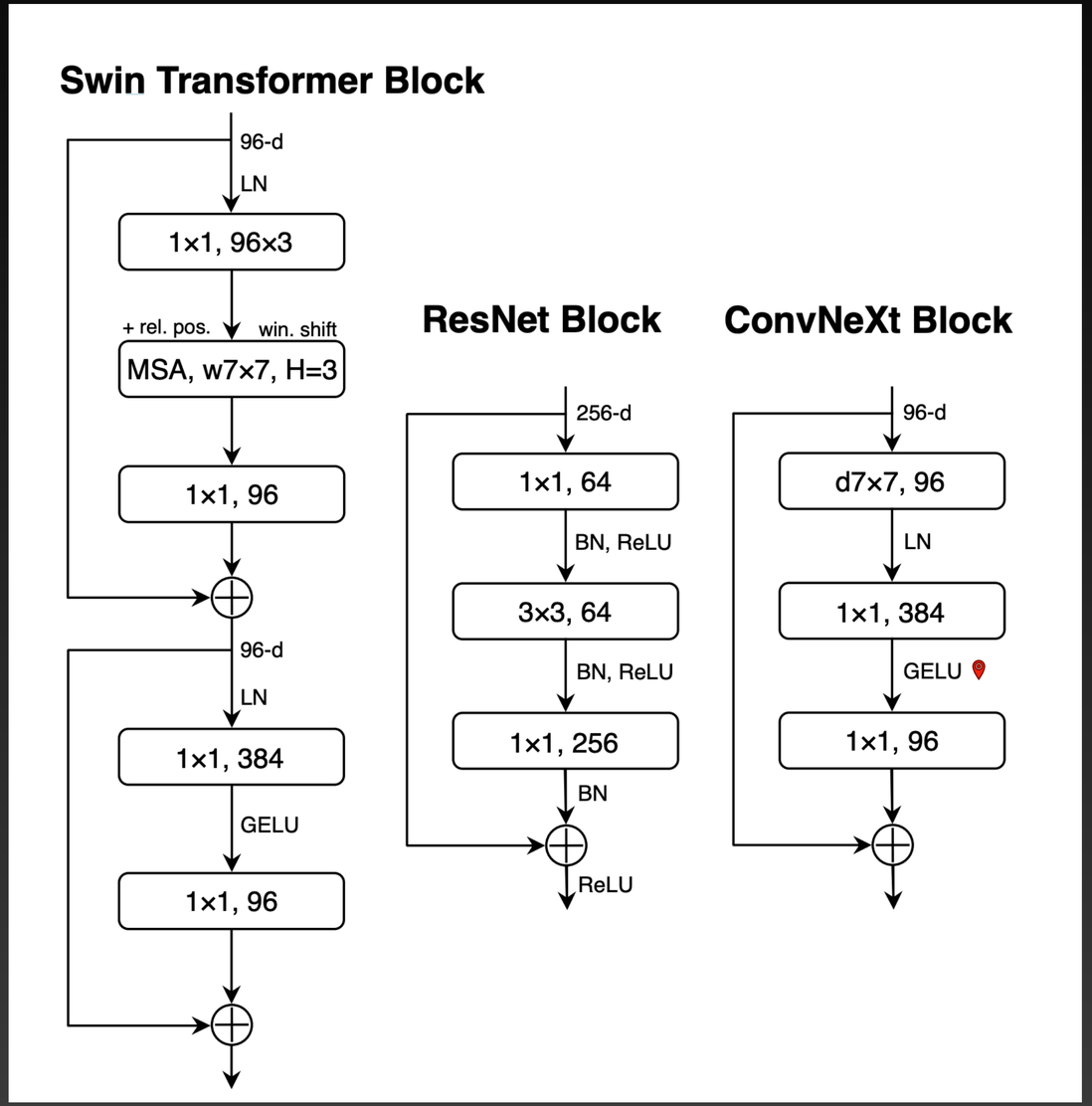
CMP 719 - Computer Vision

Serkan UYSAL - 25.04.2024

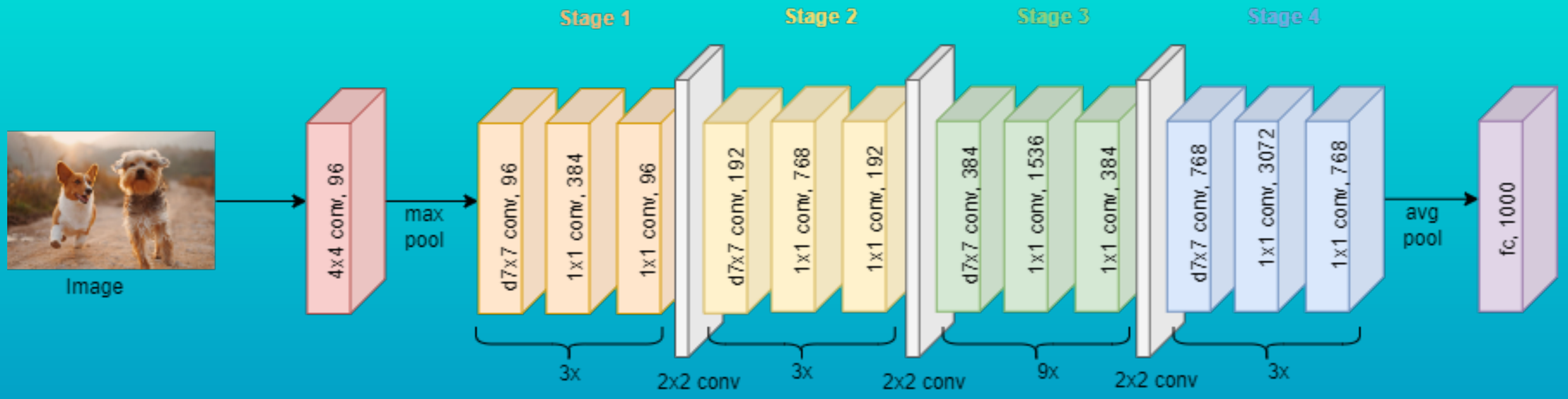
ConvNext v1

2022

Blocks


Swin Transformer Block
96-d
LN
1×1, 96×3
+ rel. pos.
win. shift
MSA, w7×7, H=3
1×1, 96
96-d
LN
1×1, 384
GELU
1×1, 96
ResNet Block
256-d
1×1, 64
BN, ReLU
3×3, 64
BN, ReLU
1×1, 256
BN
ReLU
ConvNeXt Block
96-d
d7×7, 96
LN
1×1, 384
GELU
1×1, 96

Structure


Stage 1
Stage 2
Stage 3
Stage 4
Image
4x4 conv, 96
max pool
d7x7 conv, 96
1x1 conv, 384
1x1 conv, 96
d7x7 conv, 192
1x1 conv, 768
1x1 conv, 192
d7x7 conv, 384
1x1 conv, 1536
1x1 conv, 384
d7x7 conv, 768
1x1 conv, 3072
1x1 conv, 768
avg pool
fc, 1000
3x
2x2 conv
3x
2x2 conv
9x
2x2 conv
3x

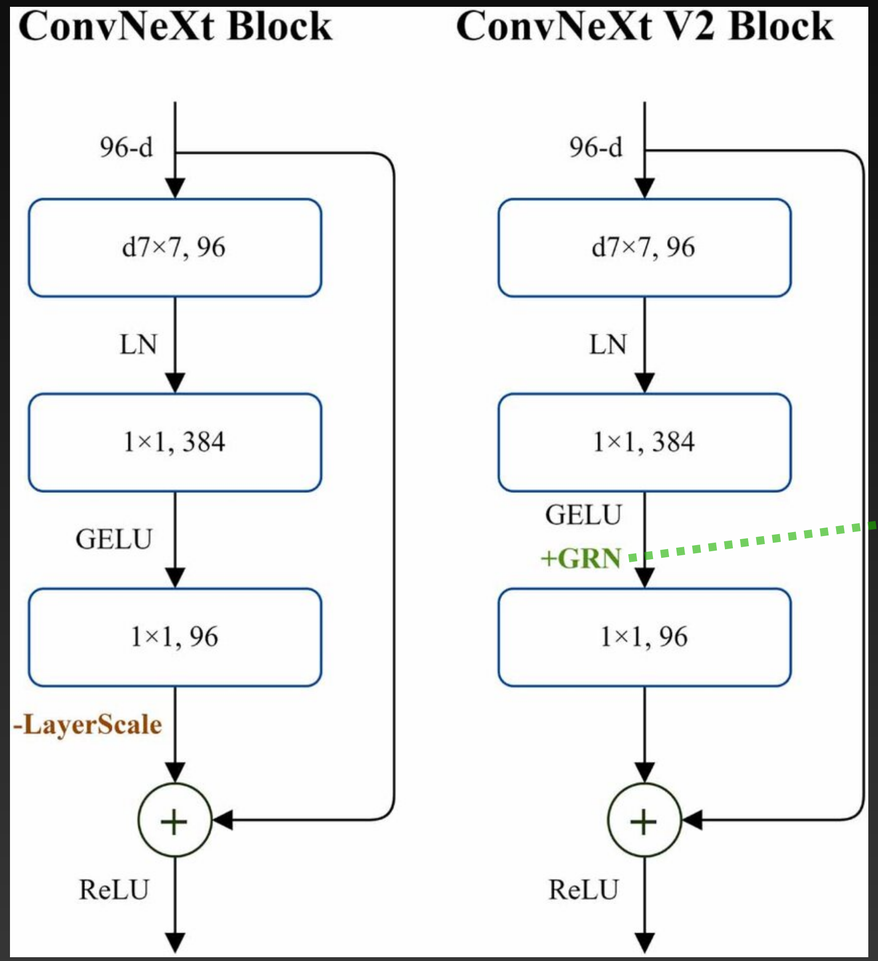
Results

model	image size	#param.	FLOPs	throughput (image / s)	IN-1K top-1 acc.
	ImageNet-1K trained models				
● RegNetY-16G [54]	224^2	84M	16.0G	334.7	82.9
● EffNet-B7 [71]	600^2	66M	37.0G	55.1	84.3
● EffNetV2-L [72]	480^2	120M	53.0G	83.7	85.7
○ DeiT-S [73]	224^2	22M	4.6G	978.5	79.8
○ DeiT-B [73]	224^2	87M	17.6G	302.1	81.8
○ Swin-T	224^2	28M	4.5G	757.9	81.3
● ConvNeXt-T	224^2	29M	4.5G	774.7	**82.1**
○ Swin-S	224^2	50M	8.7G	436.7	83.0
● ConvNeXt-S	224^2	50M	8.7G	447.1	**83.1**
○ Swin-B	224^2	88M	15.4G	286.6	83.5
● ConvNeXt-B	224^2	89M	15.4G	292.1	**83.8**
○ Swin-B	384^2	88M	47.1G	85.1	84.5
● ConvNeXt-B	384^2	89M	45.0G	95.7	**85.1**
● ConvNeXt-L	224^2	198M	34.4G	146.8	**84.3**
● ConvNeXt-L	384^2	198M	101.0G	50.4	**85.5**

ConvNext v2

2023 CVPR

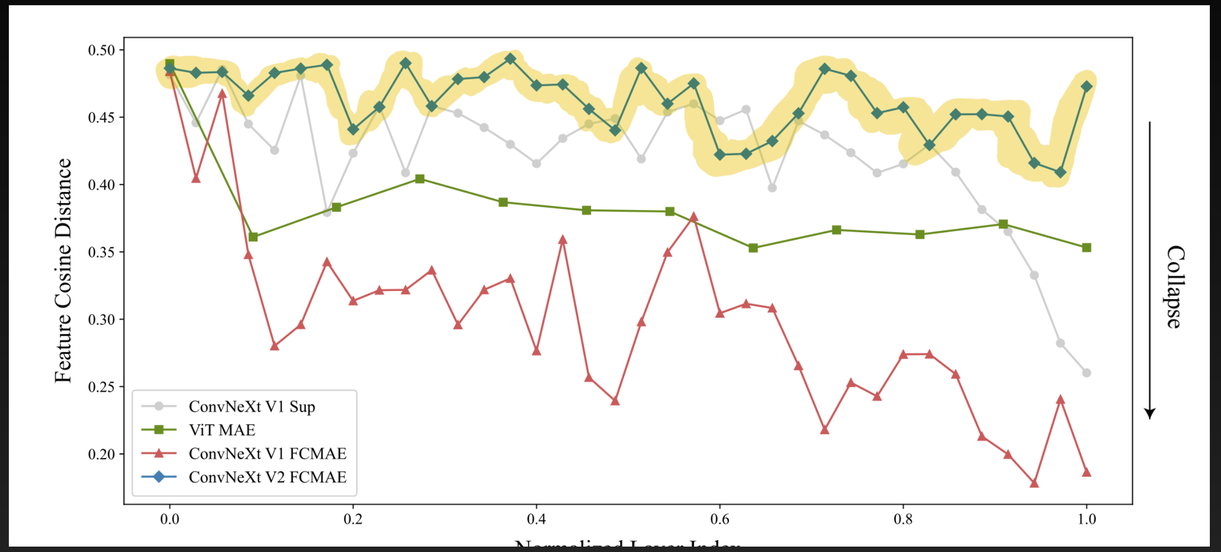
Global Response Normalization



1. Global Feature Aggregation
2. Feature Normalization
3. Feature Calibration

```
# gamma, beta: learnable affine transform parameters
# X: input of shape (N,H,W,C)
gx = torch.norm(X, p=2, dim=(1,2), keepdim=True)
nx = gx / (gx.mean(dim=-1, keepdim=True)+1e-6)
return gamma * (X * nx) + beta + X
```

Parameter Tuning



case	ft
g.avg.	83.7
L1	84.3
L2	**84.6**

(a) **Global aggregation** $G(\cdot)$. L2 Norm-based aggregation function produces the best result.

case	ft
$(\|\|X_i\|\| - \mu)/\sigma$	84.5
$1/\sum \|\|X_i\|\|$	83.8
$\|\|X_i\|\|/\sum \|\|X_i\|\|$	**84.6**

(b) **Normalization operator,** $N(\cdot)$. Divisive normalization is an effective channel importance calibrator.

case	ft
w/o skip	84.0
w/ skip	**84.6**

(c) **Residual connection** helps with GRN optimization and leads to better performance.

case	ft
Baseline	83.7
LRN [26]	83.2
BN [22]	80.5
LN [1]	83.8
GRN	**84.6**

(d) **Feature normalization**. GRN outperforms other normalizations through global contrasting.

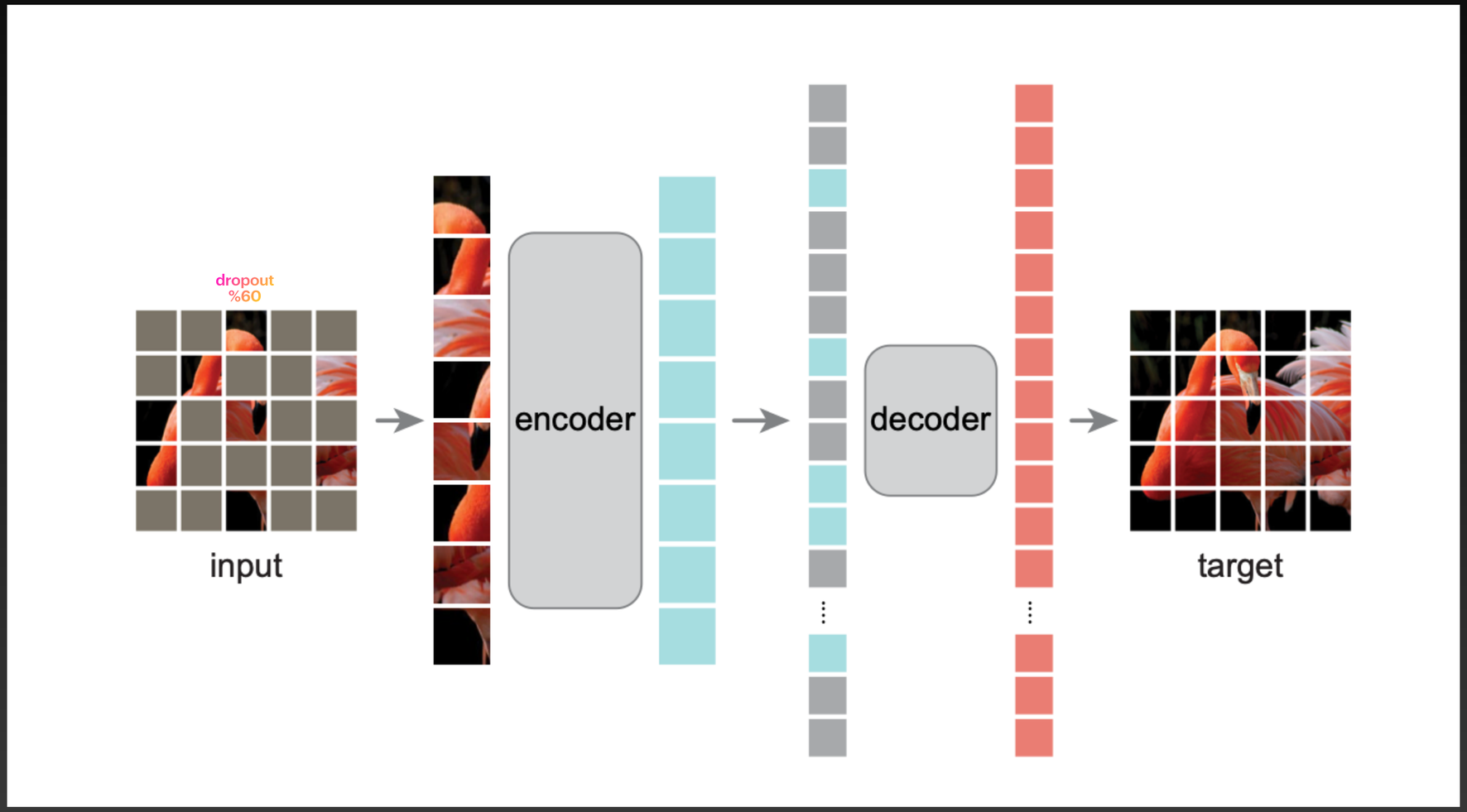
case	ft	#param
Baseline	83.7	89M
SE [19]	84.4	109M
CBAM [48]	84.5	109M
GRN	**84.6**	**89M**

(e) **Feature re-weighting**. GRN does effective and efficient feature re-weighting without parameter overhead.

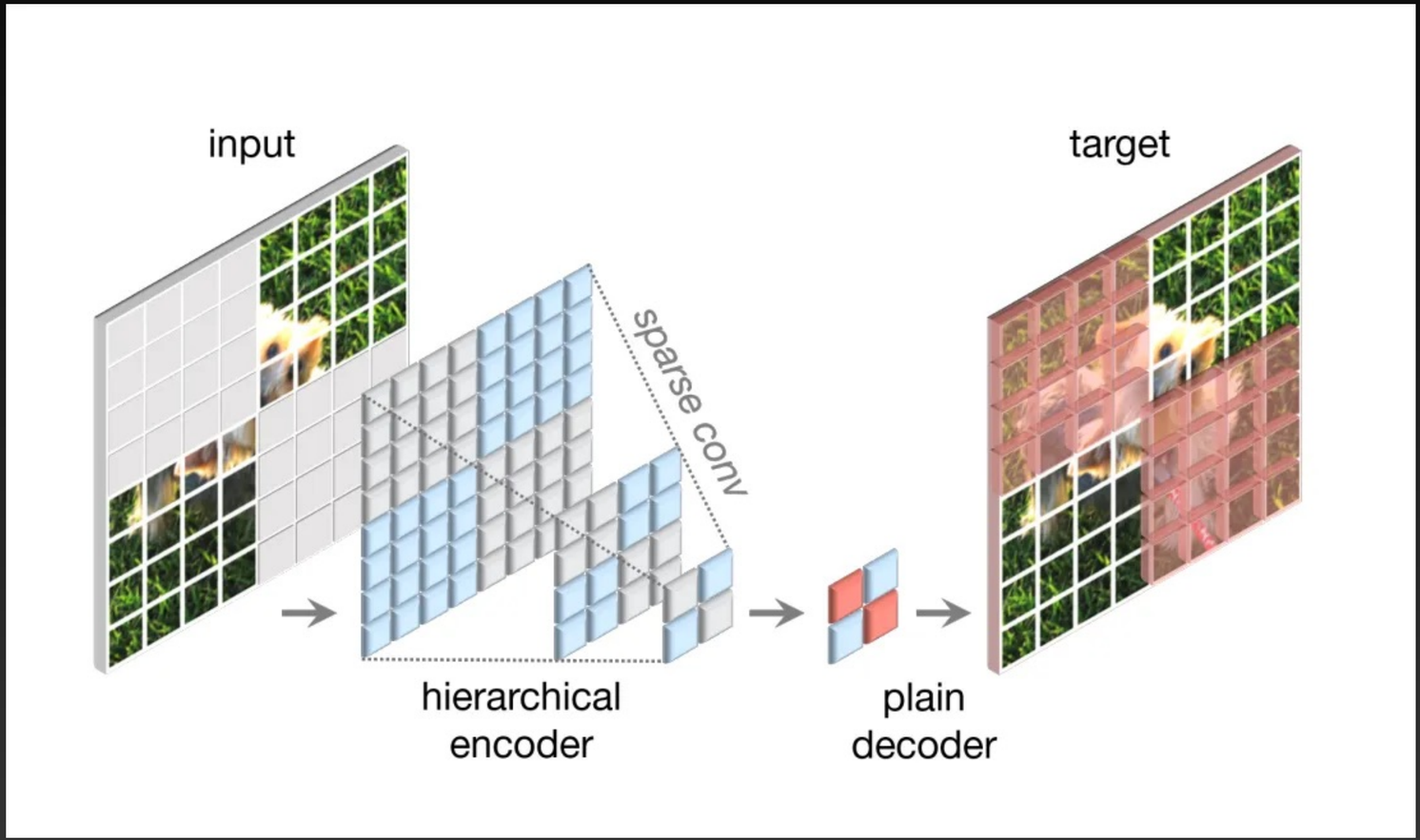
case	ft
Baseline	83.7
drop at ft.	78.8
add at ft.	80.6
both	**84.6**

(f) **GRN in pre-training/fine-tuning**. To be effective, GRN should be used in both stages.

Masked Auto-Encoder

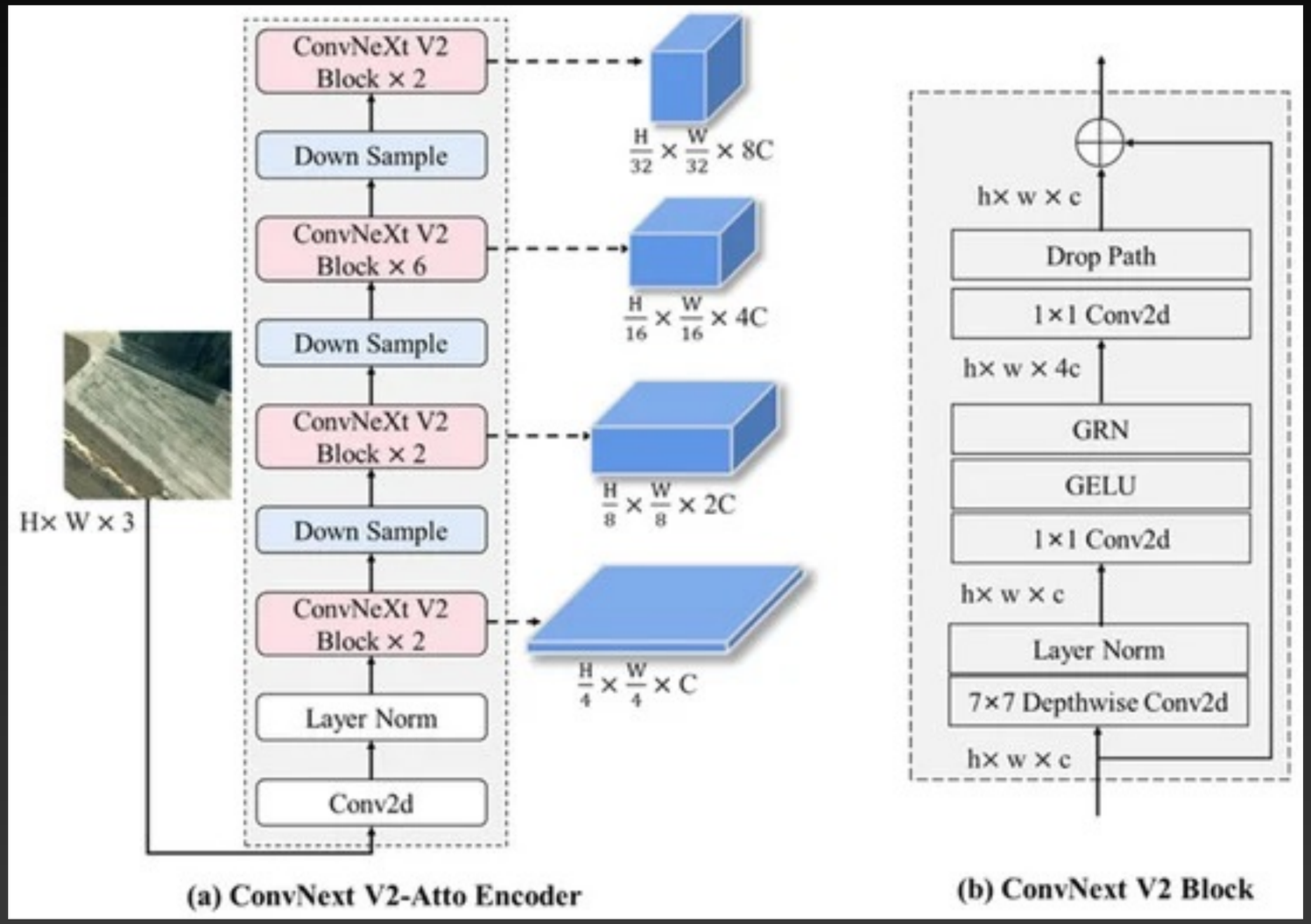


Fully Convolutional Masked AutoEncoder (FCMAE)

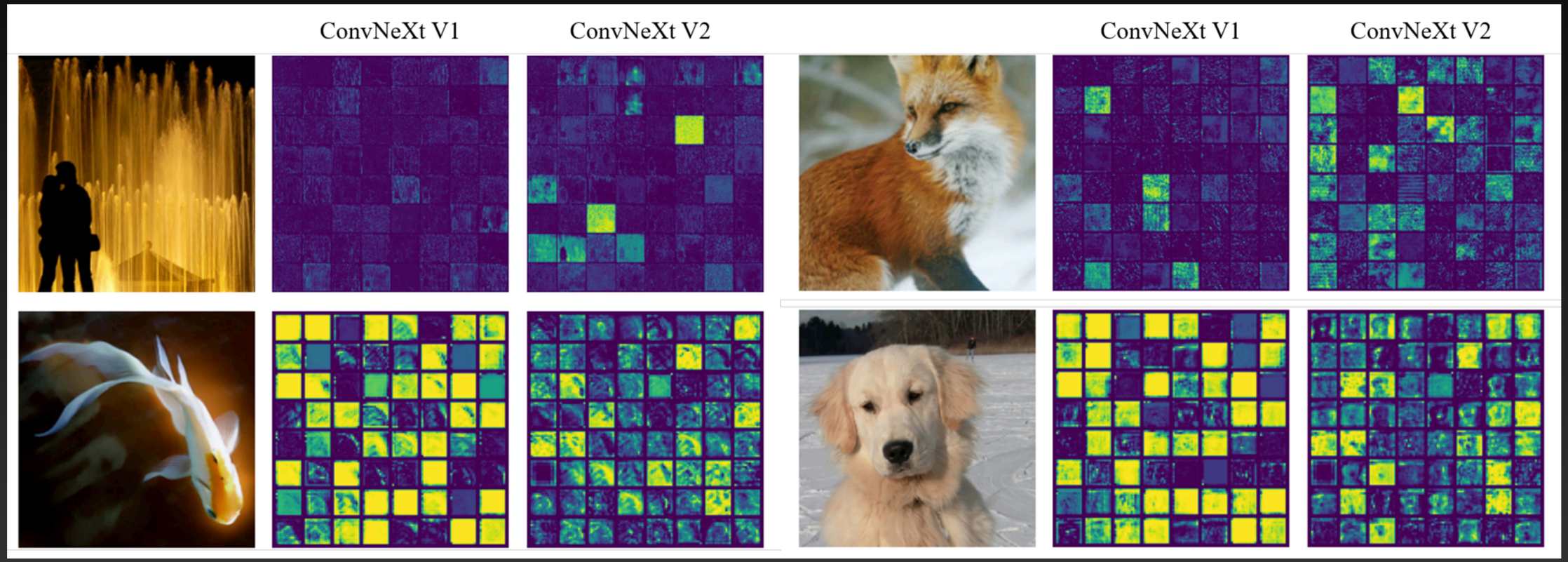


Encoder Decoder

Encoder & Decoder Block



Results



Results on ImageNet

Backbone	Method	#param	PT epoch	FT acc.
ViT-B	BEiT	88M	800	83.2
ViT-B	MAE	88M	1600	83.6
Swin-B	SimMIM	88M	800	84.0
ConvNeXt V2-B	FCMAE	89M	800	84.6
ConvNeXt V2-B	FCMAE	89M	1600	**84.9**
ViT-L	BEiT	307M	800	85.2
ViT-L	MAE	307M	1600	85.9
Swin-L	SimMIM	197M	800	85.4
ConvNeXt V2-L	FCMAE	198M	800	85.6
ConvNeXt V2-L	FCMAE	198M	1600	**85.8**
ViT-H	MAE	632M	1600	86.9
Swin V2-H	SimMIM	658M	800	85.7
ConvNeXt V2-H	FCMAE	659M	800	85.8
ConvNeXt V2-H	FCMAE	659M	1600	**86.3**

References

1. ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders - arXiv:2301.00808}
2. ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders | CVPR 2023
3. https://huggingface.co/docs/transformers/en/model_doc/convnextv2
4. https://medium.com/thedeephub/papers-explained-94-convnext-v2-2ecdabf2081c
5. https://ritvik19.medium.com/papers-explained-92-convnext-d13385d9177d

End.

ConvNext v2

ConvNext v1