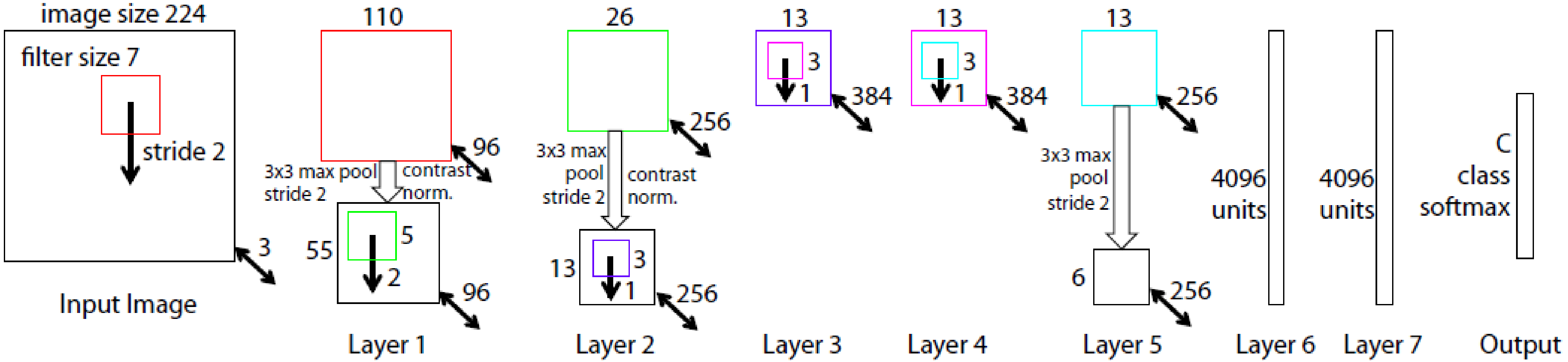
Visualizing and Understanding Convolutional Networks

Matthew D. Zeiler, Rob Fergus, New York University

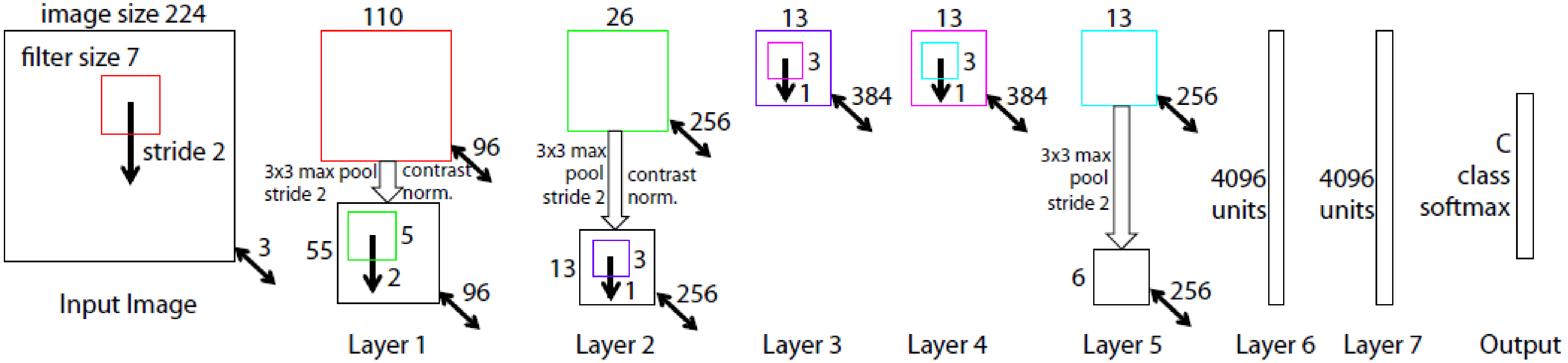
Presenter: Wanli Ma, Oct 14 2015

Convolution Neural Network



- Krizhevsky et. al. NIPS 2012

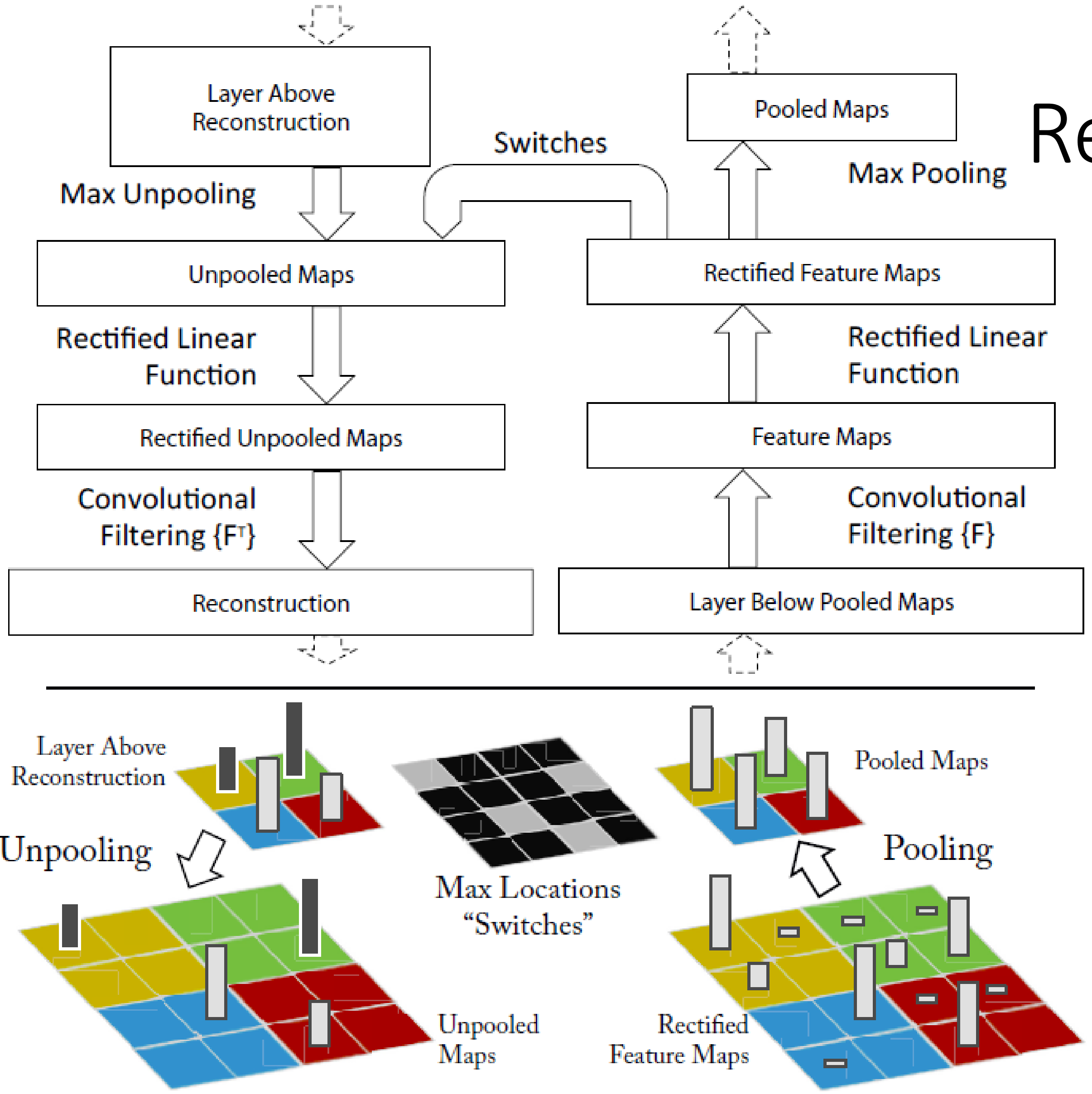
Convolution Neural Network



  - Krizhevsky et. al. NIPS 2012

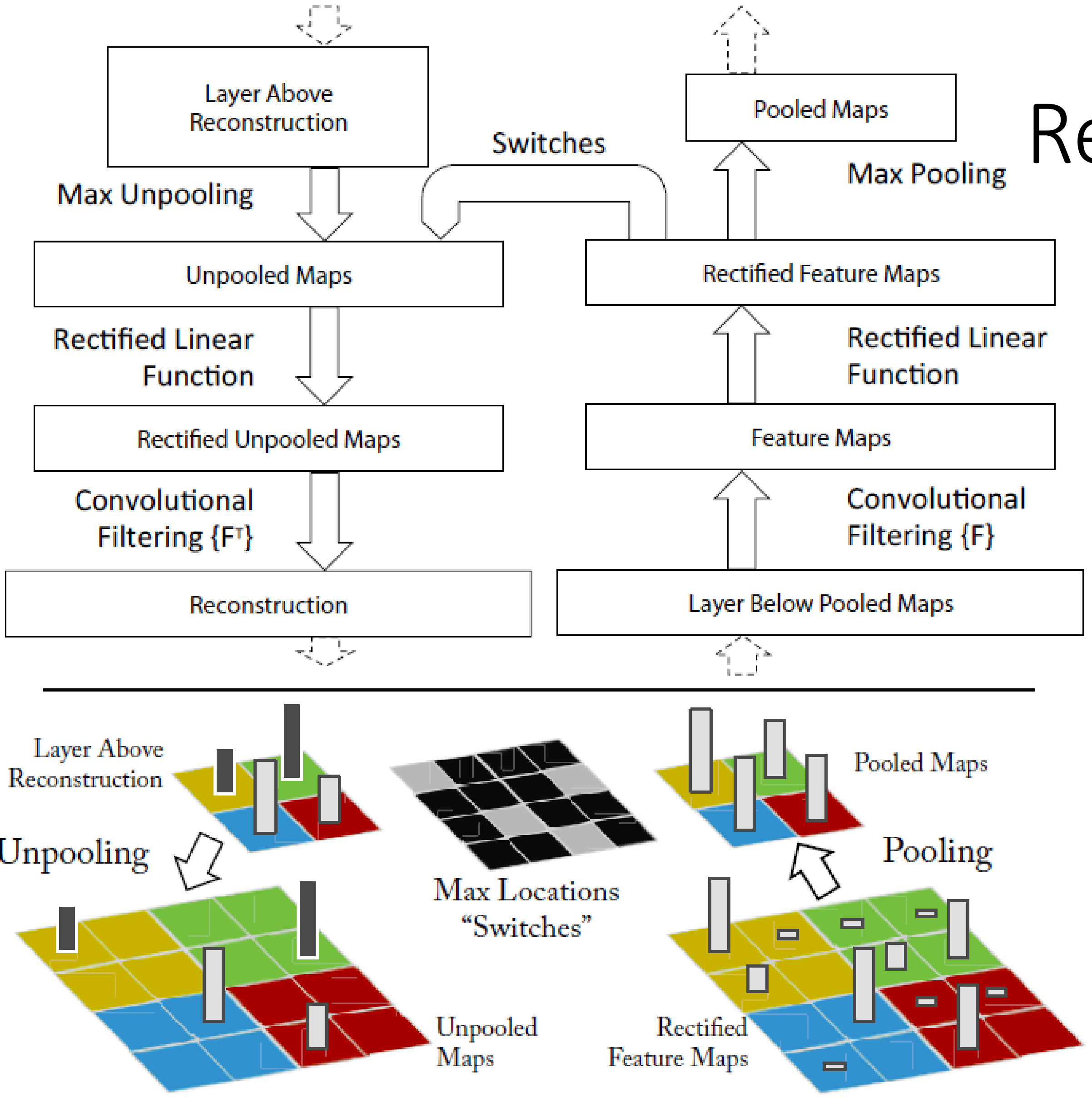
- Visualize hidden layer
  - 1st layer: map back to the input pixel space
  - Higher (convolutional) layers: how?
  - How to understand?

Reverse Operations

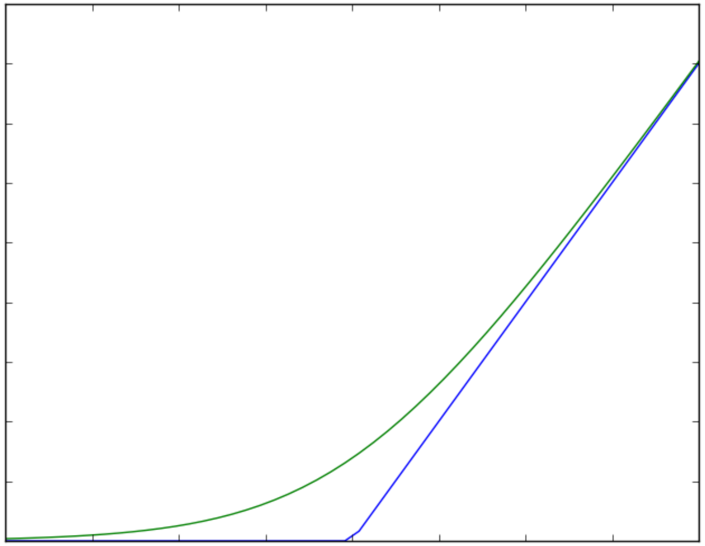


- For each layer
  - Convolution -> deconvolution
  - Max-pooling -> unpool
  - *relu -> relu*

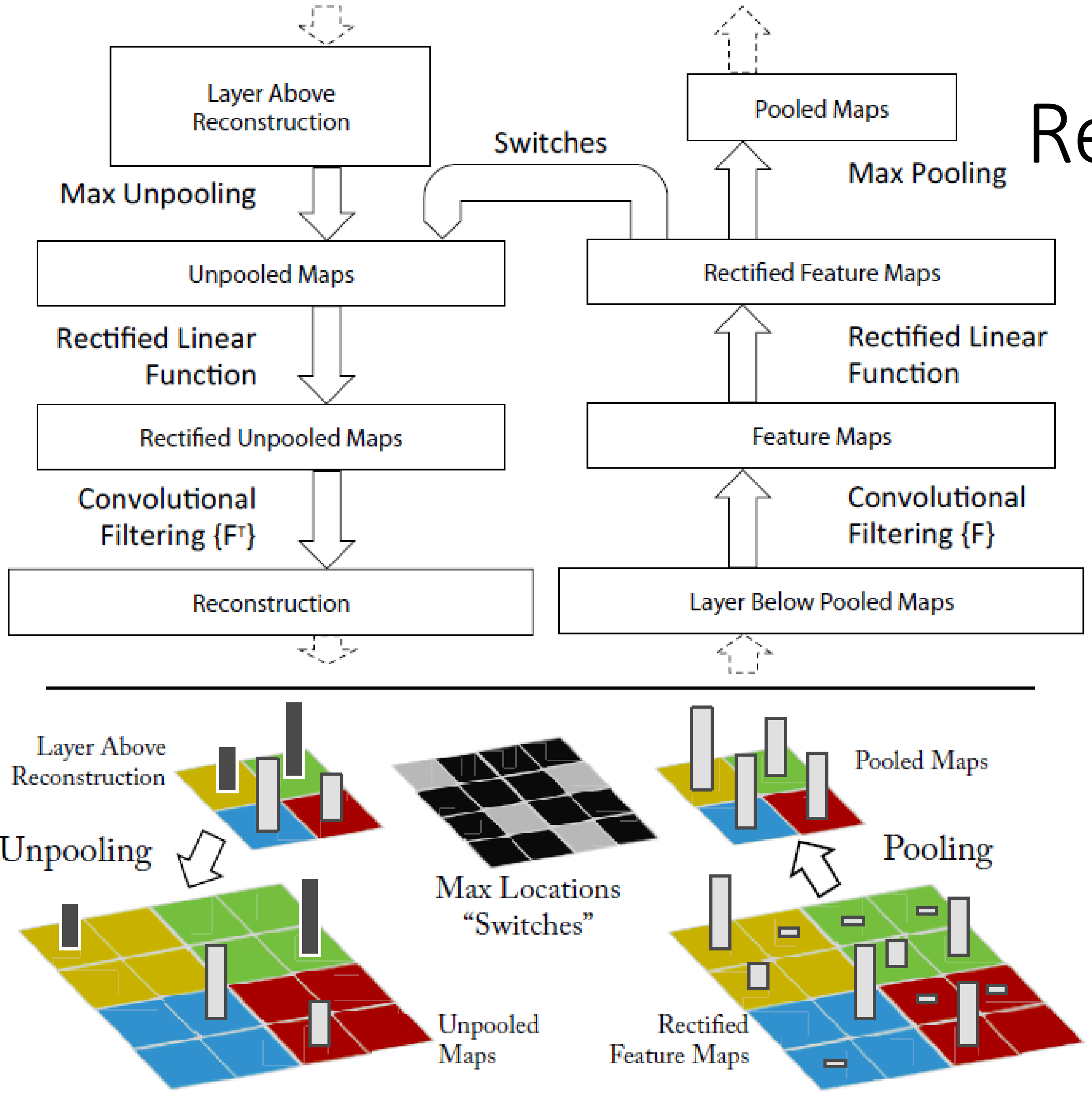
Reverse Operations



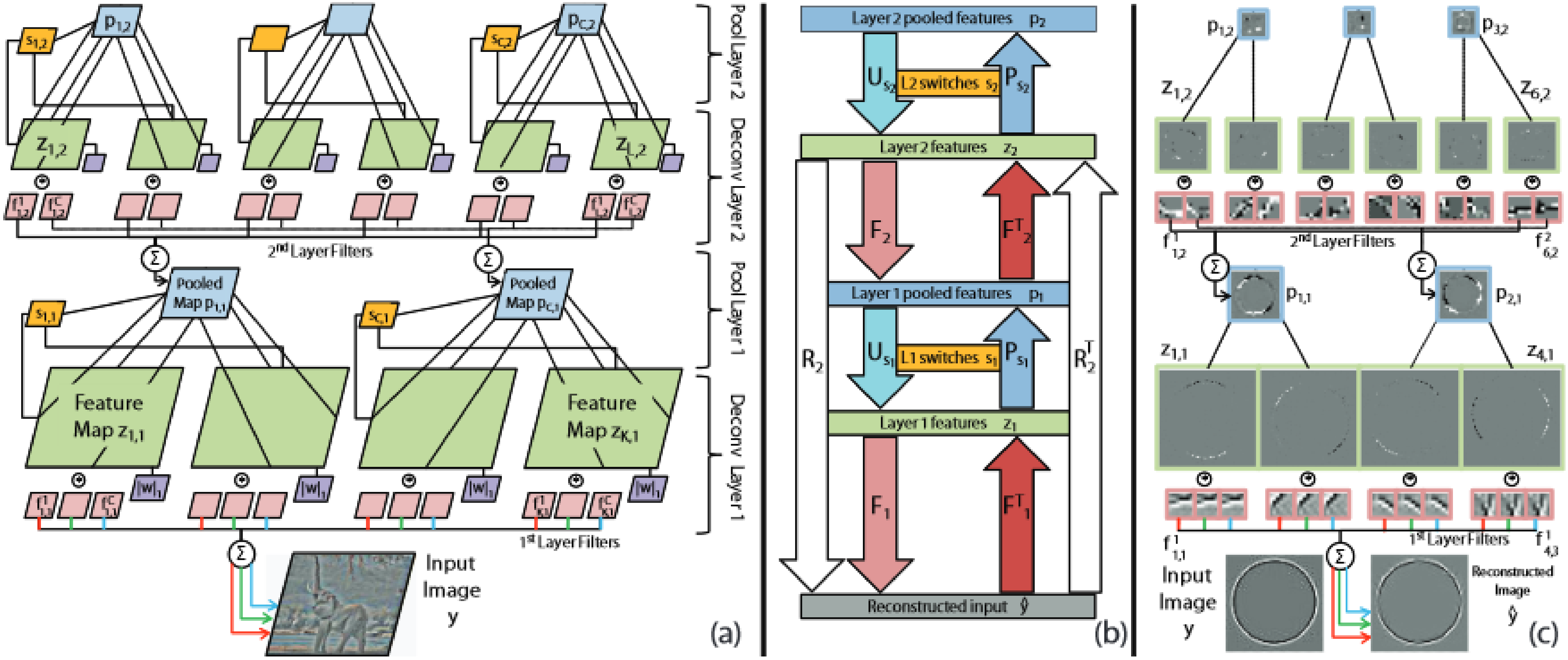
- For each layer
  - Convolution -> deconvolution
  - Max-pooling -> unpool
  - *relu -> relu*

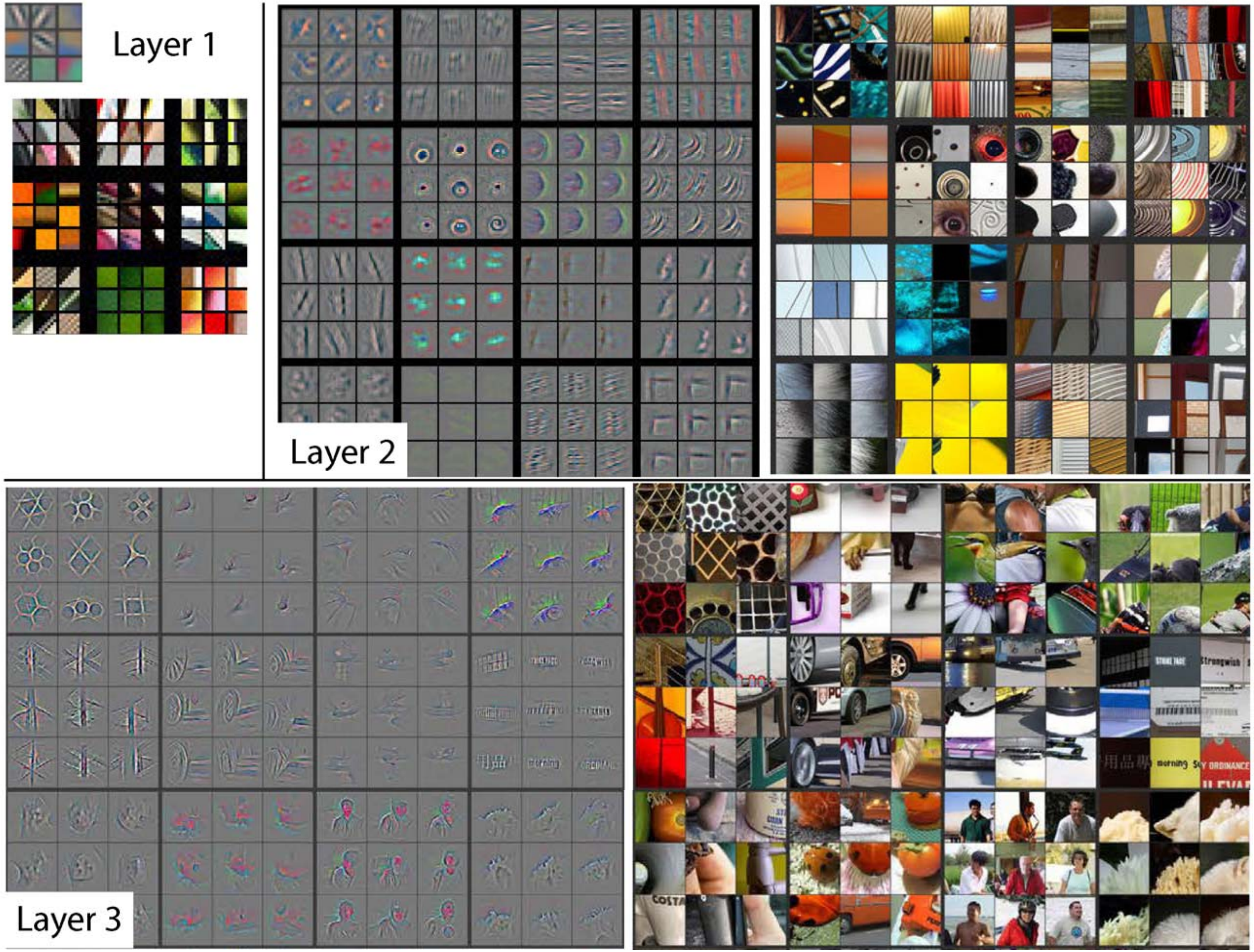


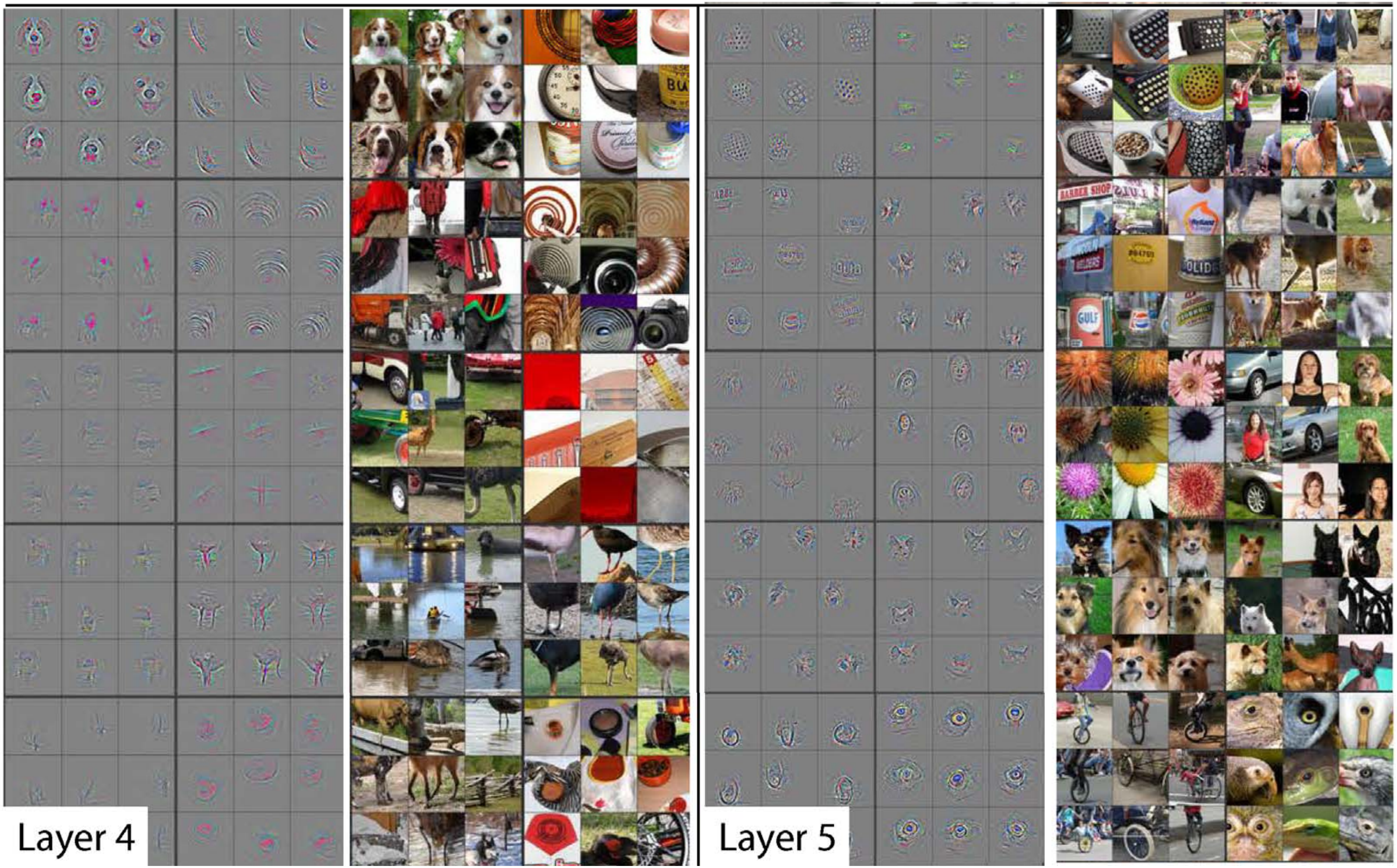
Reverse Operations

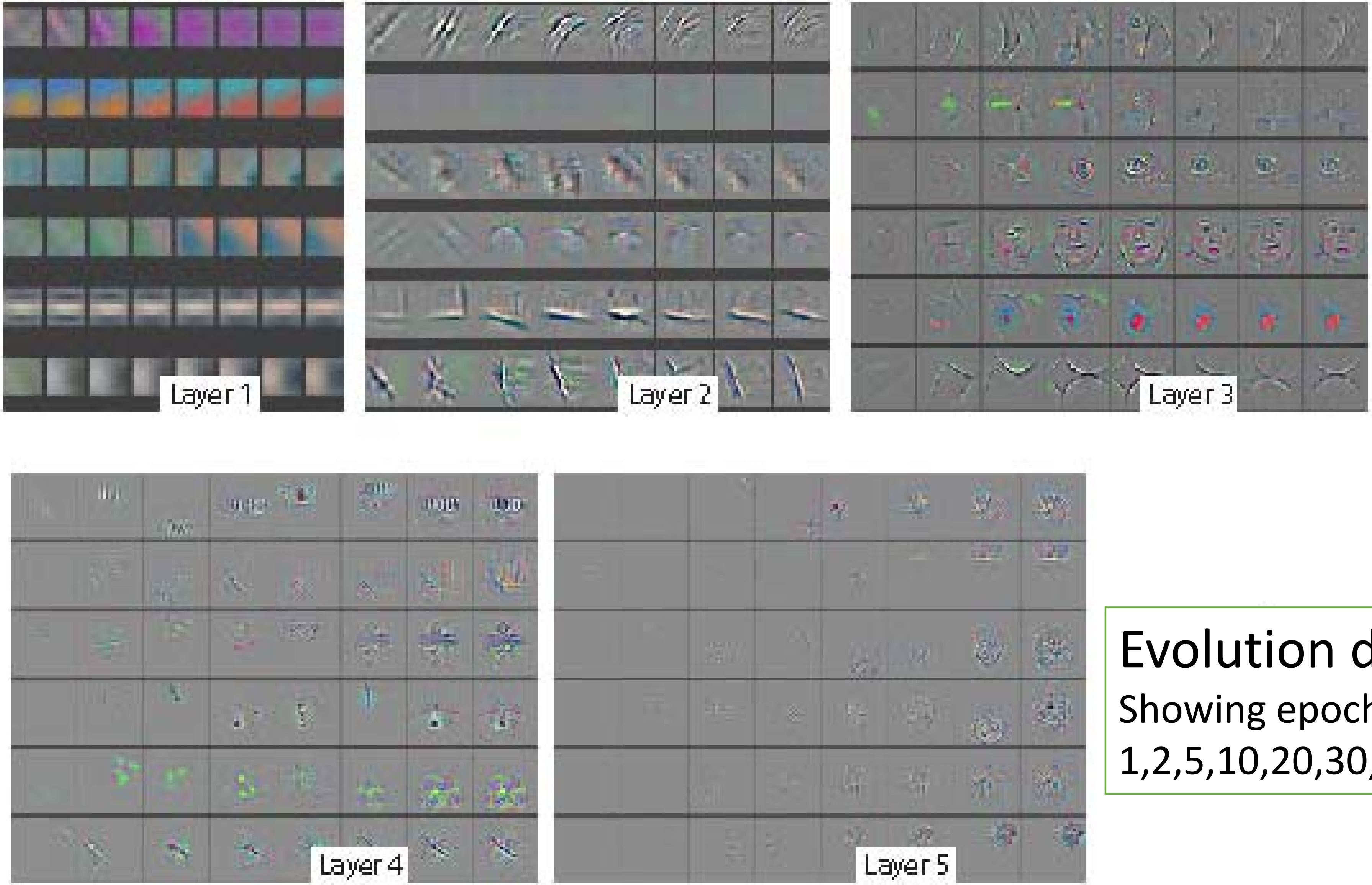


- For each layer
  - Convolution -> deconvolution
  - Max-pooling -> unpool
  - *relu -> relu*
  - No 'contrast normalization'


$p_{1,2}$
$s_{1,2}$
$p_{C,2}$
$s_{C,2}$
$z_{1,2}$
$z_{L,2}$
$f^1_{1,2}$
$f^C_{1,2}$
$f^1_{L,2}$
$f^C_{L,2}$
2^{nd} Layer Filters
Pooled Map $p_{1,1}$
Pooled Map $p_{C,1}$
$s_{1,1}$
$s_{C,1}$
Feature Map $z_{1,1}$
Feature Map $z_{K,1}$
$|w|_1$
$f^1_{1,1}$
$f^C_{1,1}$
$f^1_{K,1}$
$f^C_{K,1}$
1^{st} Layer Filters
Input Image y
Pool Layer 2
Deconv Layer 2
Pool Layer 1
Deconv Layer 1
(a)
Layer 2 pooled features p_2
U_{s_2}
L2 switches s_2
P_{s_2}
Layer 2 features z_2
R_2
F_2
F^T_2
R^T_2
Layer 1 pooled features p_1
U_{s_1}
L1 switches s_1
P_{s_1}
Layer 1 features z_1
F_1
F^T_1
Reconstructed input $\hat{y}$
(b)
$p_{1,2}$
$p_{3,2}$
$z_{1,2}$
$z_{6,2}$
$f^1_{1,2}$
2^{nd} Layer Filters
$f^2_{6,2}$
$p_{1,1}$
$p_{2,1}$
$z_{1,1}$
$z_{4,1}$
$f^1_{1,1}$
1^{st} Layer Filters
$f^1_{4,3}$
Input Image y
Reconstructed Image $\hat{y}$
(c)


Layer 1
Layer 2
Layer 3

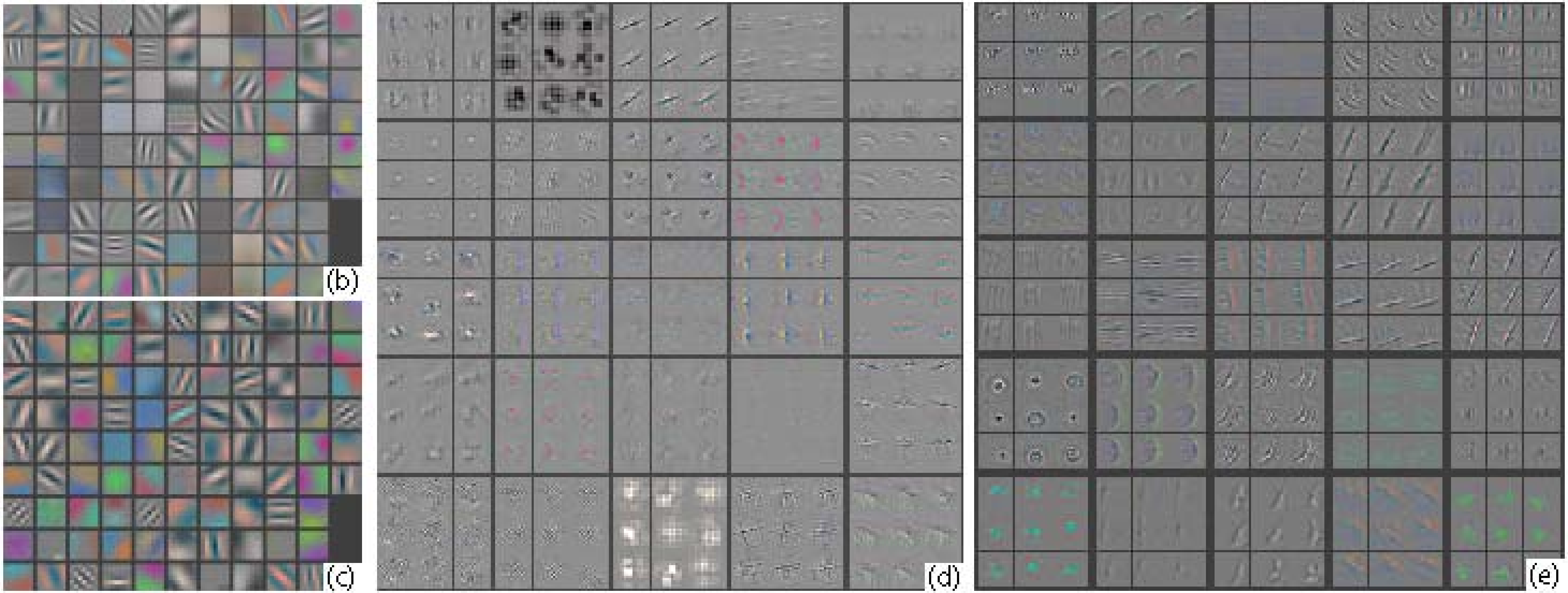

Layer 4
Layer 5



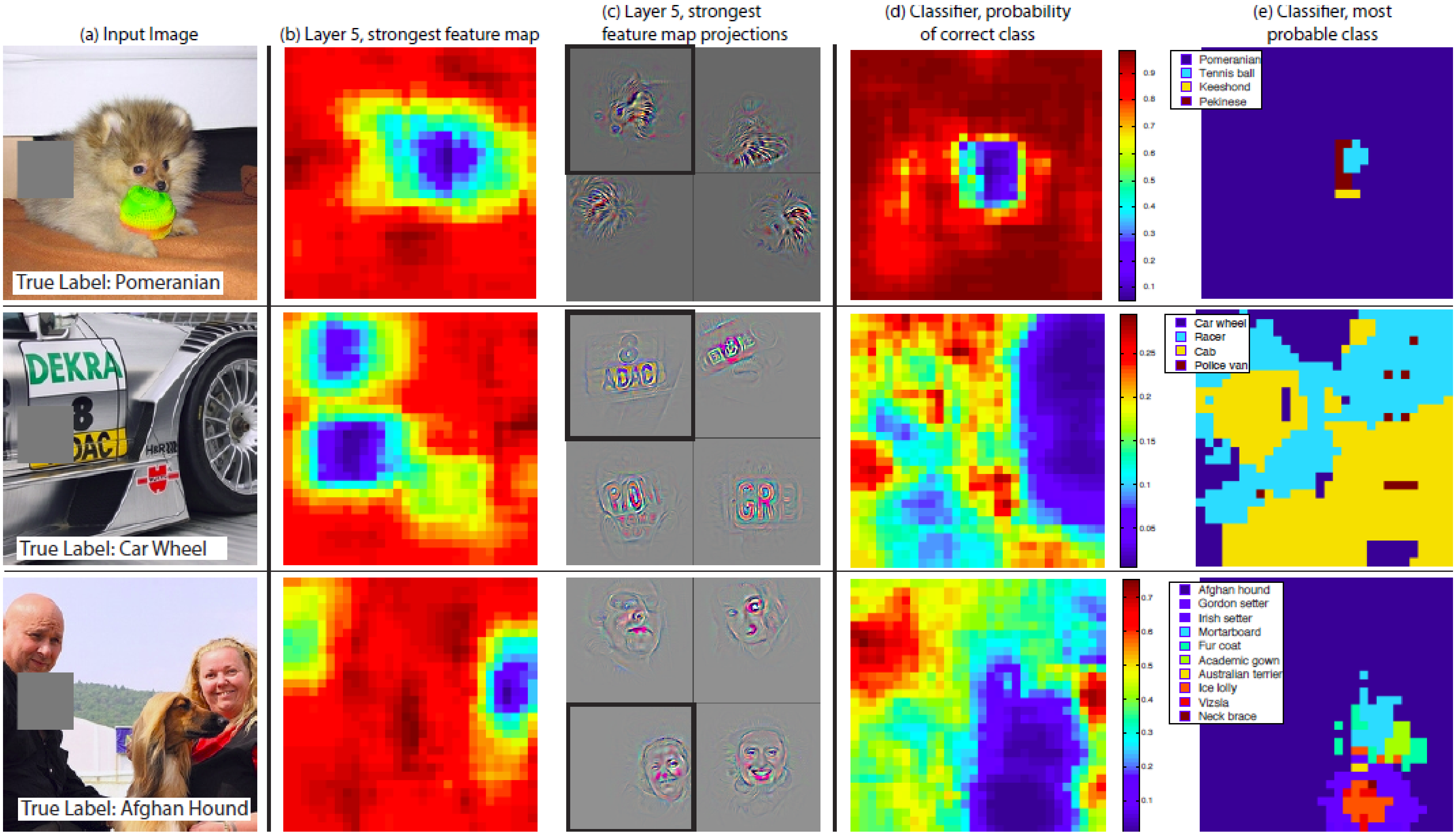
Evolution during training

Showing epochs (total 70)
1,2,5,10,20,30,40,64

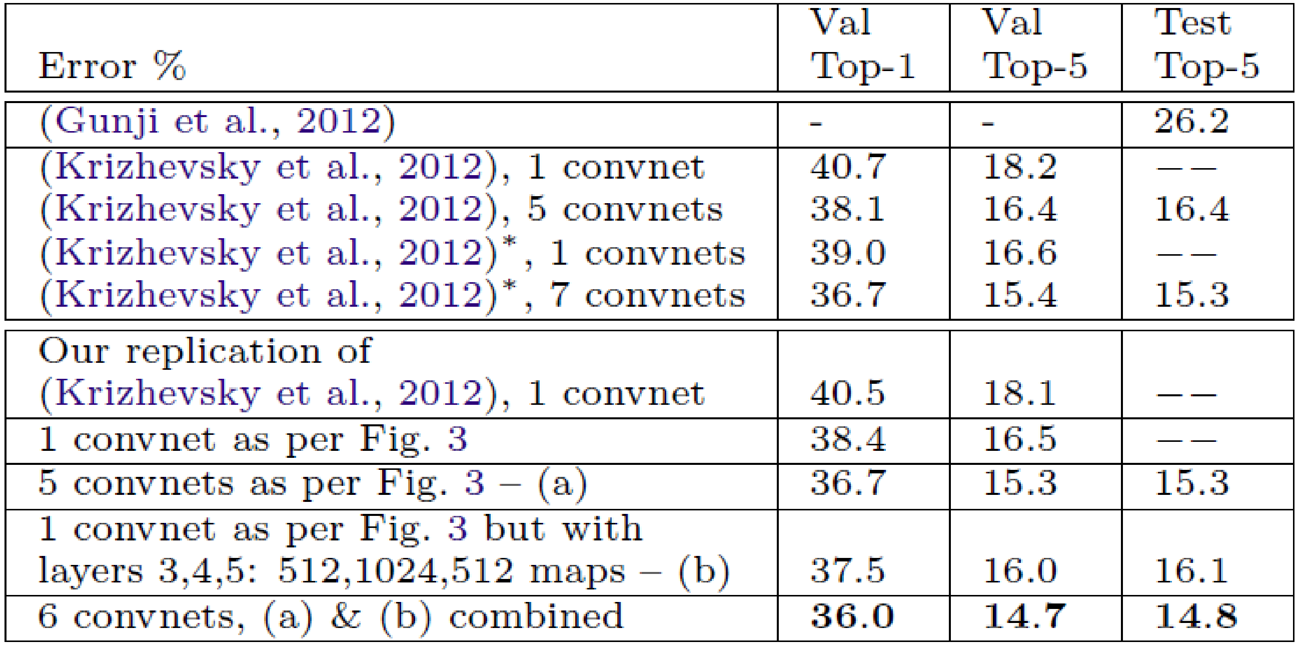
Improvements



- (b)(d)Krizhevski vs (c)(e)this paper

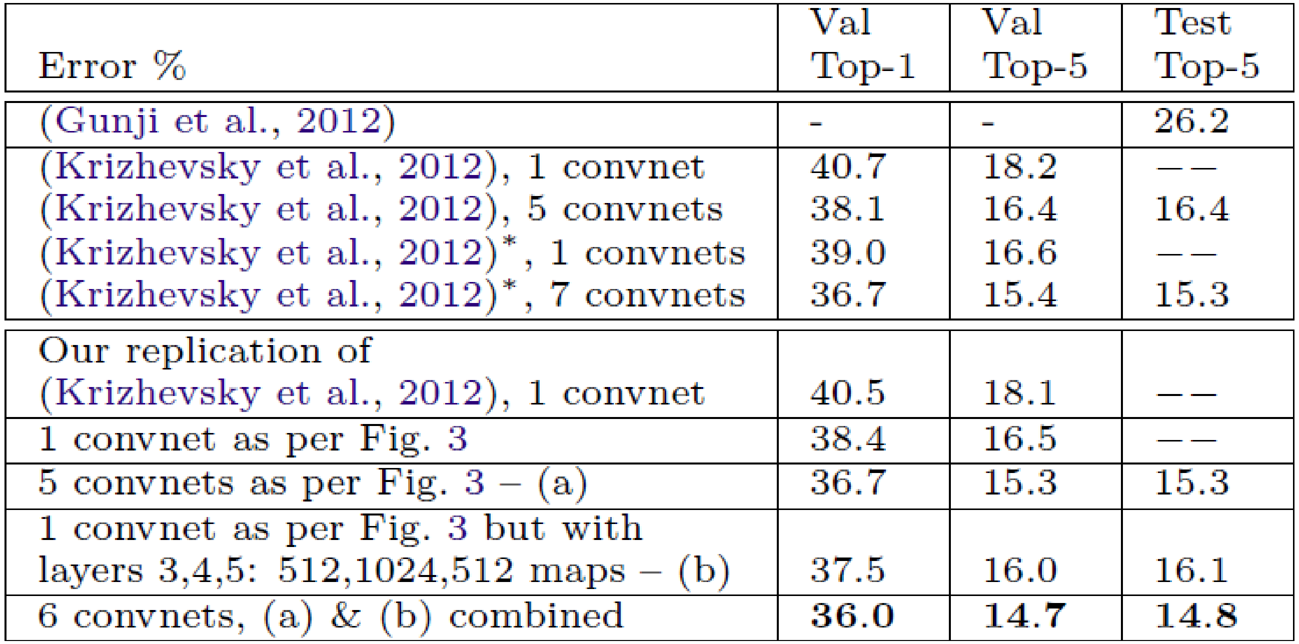

(a) Input Image
(b) Layer 5, strongest feature map
(c) Layer 5, strongest feature map projections
(d) Classifier, probability of correct class
(e) Classifier, most probable class
True Label: Pomeranian
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Pomeranian
Tennis ball
Keeshond
Pekinese
True Label: Car Wheel
0.25
0.2
0.15
0.1
0.05
Car wheel
Racer
Cab
Police van
True Label: Afghan Hound
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Afghan hound
Gordon setter
Irish setter
Mortarboard
Fur coat
Academic gown
Australian terrier
Ice lolly
Vizsla
Neck brace

Results of improvements (ImageNet 2012)

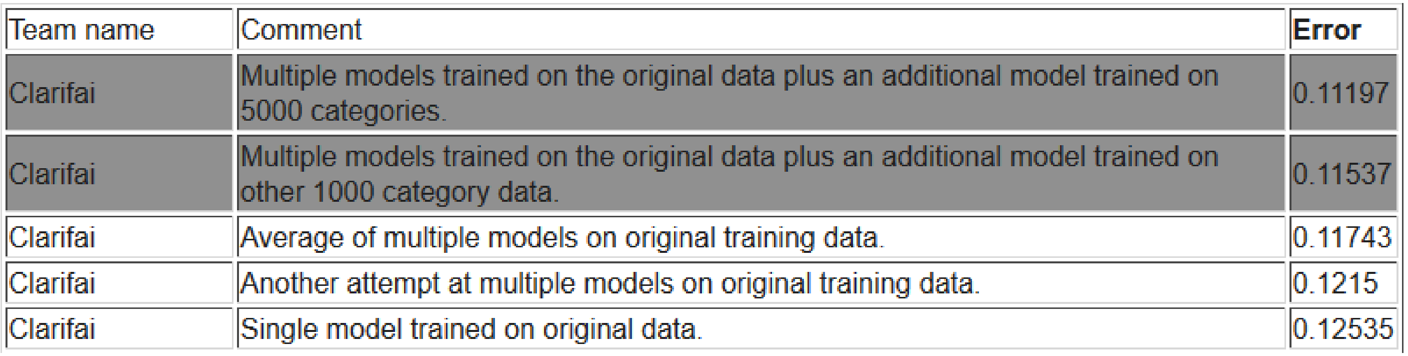


Error %	Val Top-1	Val Top-5	Test Top-5
(Gunji et al., 2012)	-	-	26.2
(Krizhevsky et al., 2012), 1 convnet	40.7	18.2	– –
(Krizhevsky et al., 2012), 5 convnets	38.1	16.4	16.4
(Krizhevsky et al., 2012)*, 1 convnets	39.0	16.6	– –
(Krizhevsky et al., 2012)*, 7 convnets	36.7	15.4	15.3
Our replication of (Krizhevsky et al., 2012), 1 convnet	40.5	18.1	– –
1 convnet as per Fig. 3	38.4	16.5	– –
5 convnets as per Fig. 3 – (a)	36.7	15.3	15.3
1 convnet as per Fig. 3 but with layers 3,4,5: 512,1024,512 maps – (b)	37.5	16.0	16.1
6 convnets, (a) & (b) combined	**36.0**	**14.7**	**14.8**

Results of improvements (ImageNet 2012)

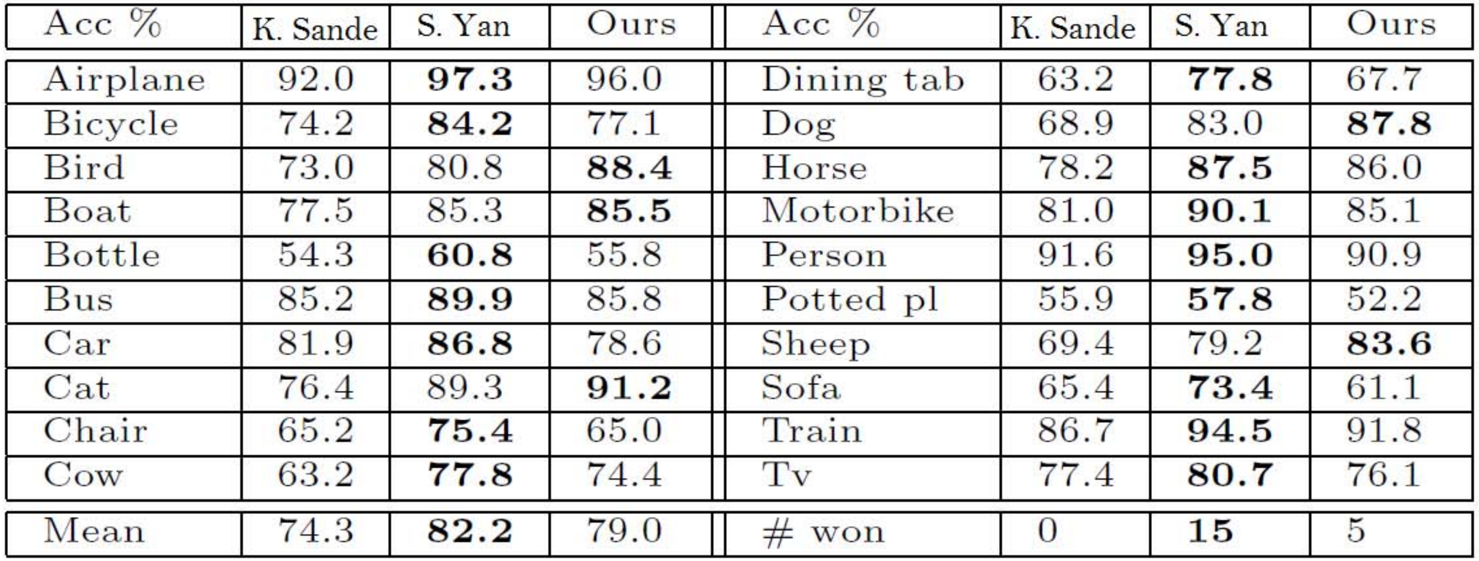


Error %	Val Top-1	Val Top-5	Test Top-5
(Gunji et al., 2012)	-	-	26.2
(Krizhevsky et al., 2012), 1 convnet	40.7	18.2	−−
(Krizhevsky et al., 2012), 5 convnets	38.1	16.4	16.4
(Krizhevsky et al., 2012)*, 1 convnets	39.0	16.6	−−
(Krizhevsky et al., 2012)*, 7 convnets	36.7	15.4	15.3
Our replication of (Krizhevsky et al., 2012), 1 convnet	40.5	18.1	−−
1 convnet as per Fig. 3	38.4	16.5	−−
5 convnets as per Fig. 3 – (a)	36.7	15.3	15.3
1 convnet as per Fig. 3 but with layers 3,4,5: 512,1024,512 maps – (b)	37.5	16.0	16.1
6 convnets, (a) & (b) combined	**36.0**	**14.7**	**14.8**



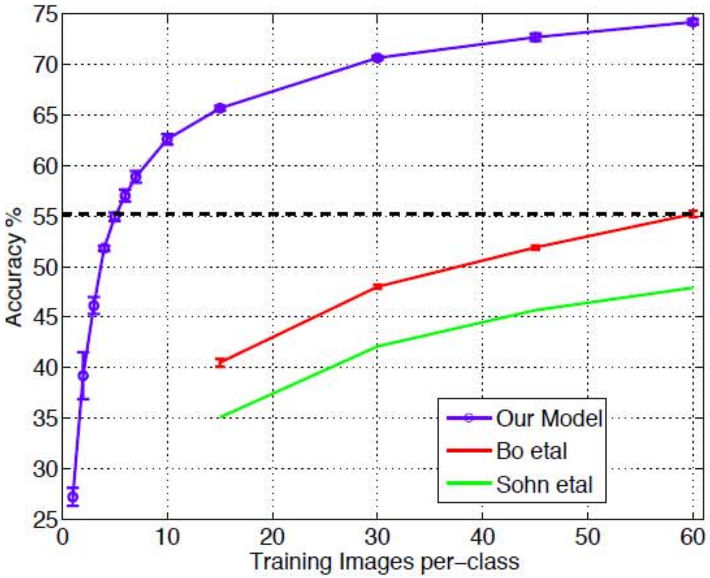
Team name	Comment	**Error**
Clarifai	Multiple models trained on the original data plus an additional model trained on 5000 categories.	0.11197
Clarifai	Multiple models trained on the original data plus an additional model trained on other 1000 category data.	0.11537
Clarifai	Average of multiple models on original training data.	0.11743
Clarifai	Another attempt at multiple models on original training data.	0.1215
Clarifai	Single model trained on original data.	0.12535

Results on other datasets



Acc %	K. Sande	S. Yan	Ours	Acc %	K. Sande	S. Yan	Ours
Airplane	92.0	**97.3**	96.0	Dining tab	63.2	**77.8**	67.7
Bicycle	74.2	**84.2**	77.1	Dog	68.9	83.0	**87.8**
Bird	73.0	80.8	**88.4**	Horse	78.2	**87.5**	86.0
Boat	77.5	85.3	**85.5**	Motorbike	81.0	**90.1**	85.1
Bottle	54.3	**60.8**	55.8	Person	91.6	**95.0**	90.9
Bus	85.2	**89.9**	85.8	Potted pl	55.9	**57.8**	52.2
Car	81.9	**86.8**	78.6	Sheep	69.4	79.2	**83.6**
Cat	76.4	89.3	**91.2**	Sofa	65.4	**73.4**	61.1
Chair	65.2	**75.4**	65.0	Train	86.7	**94.5**	91.8
Cow	63.2	**77.8**	74.4	Tv	77.4	**80.7**	76.1
Mean	74.3	**82.2**	79.0	# won	0	**15**	5

PASCAL 2012



Caltech 256

Thank you

References

- Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." *Computer Vision–ECCV 2014*. Springer International Publishing, 2014. 818-833.
- Zeiler, Matthew D., Graham W. Taylor, and Rob Fergus. "Adaptive deconvolutional networks for mid and high level feature learning." *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011.