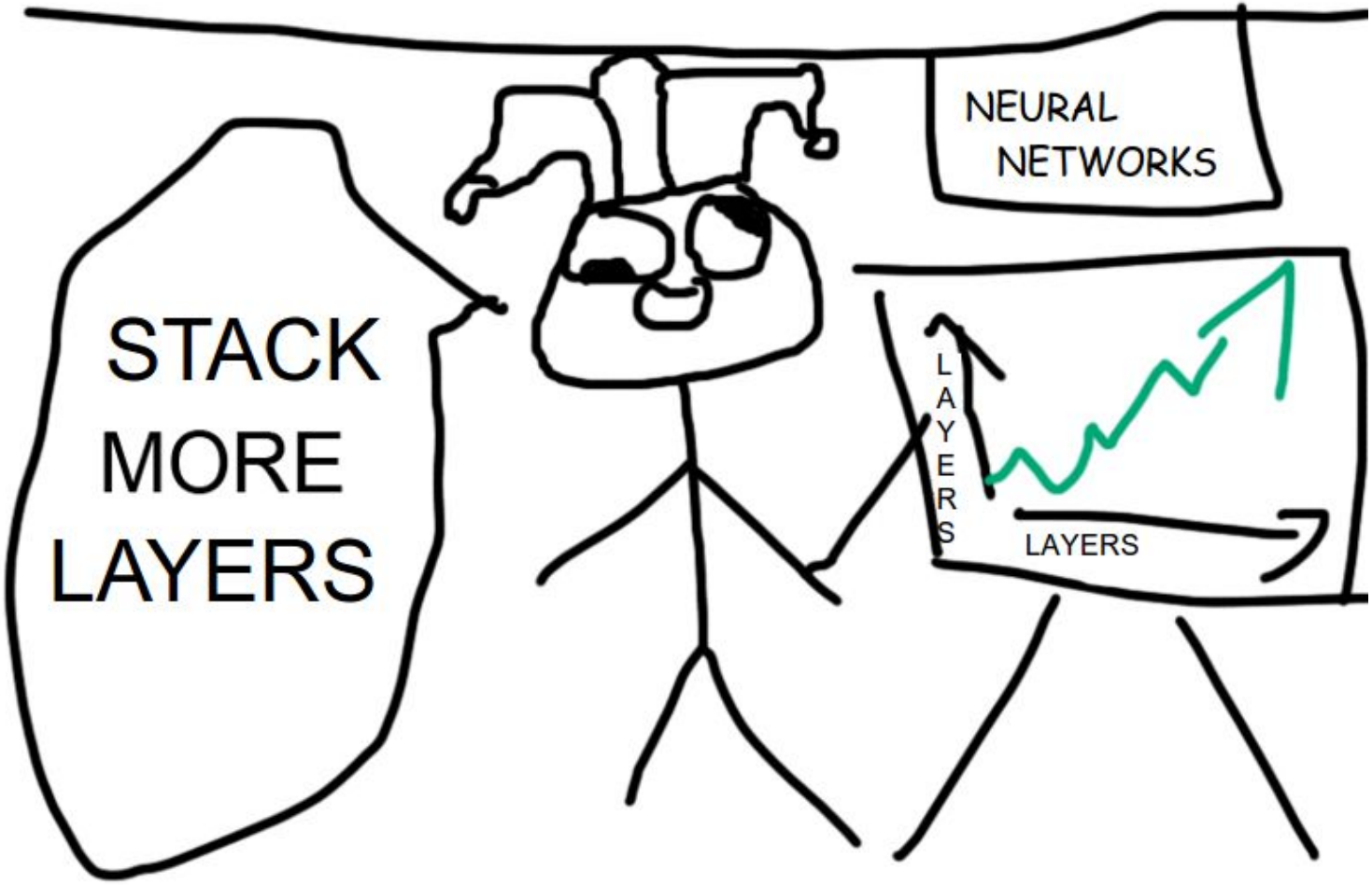
Deep Learning Year in Review 2016: Computer Vision Perspective

Alex Kalinin, PhD Candidate
Bioinformatics @ UMich
alxndrkalinin@gmail.com

@alxndrkalinin

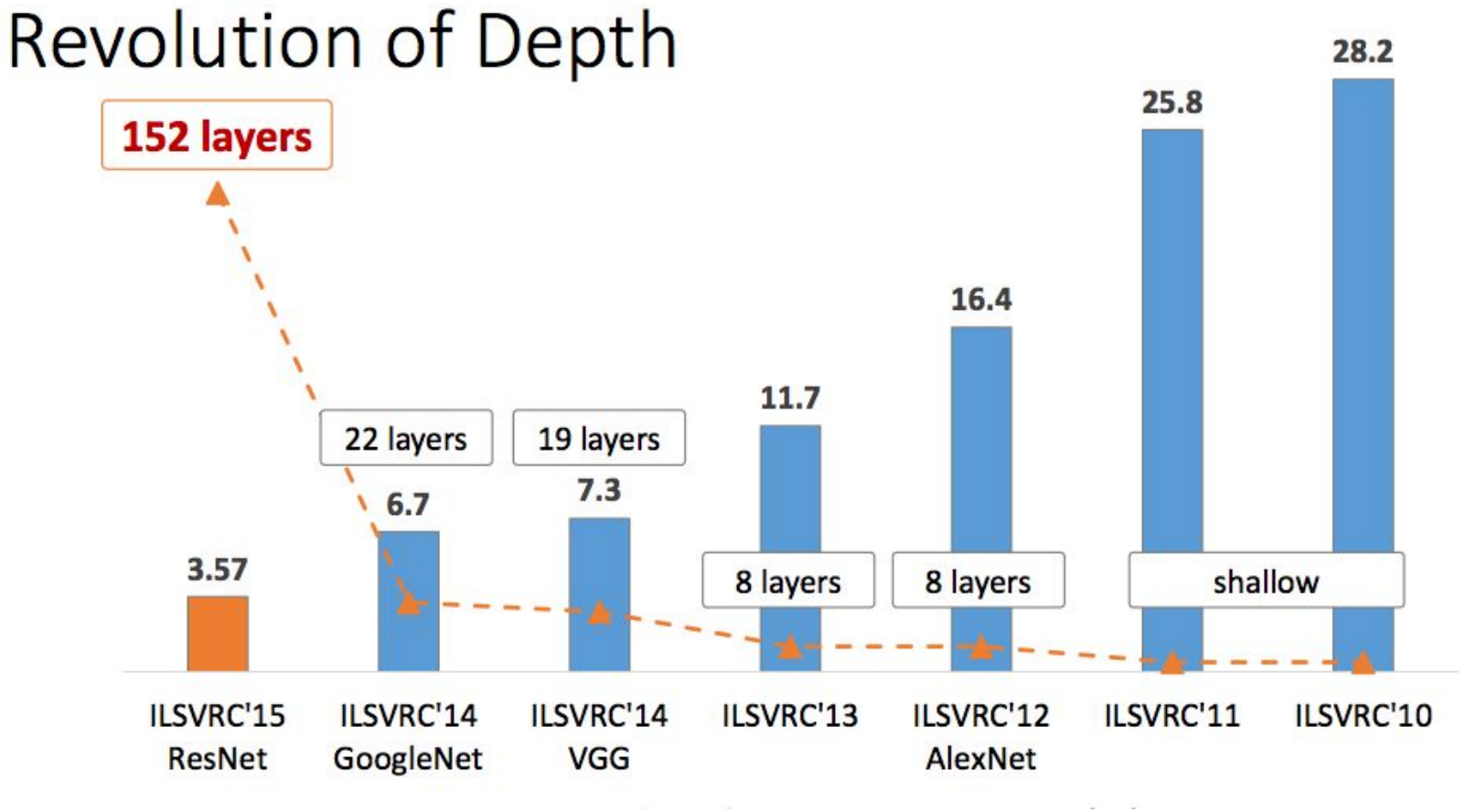
Architectures

Summary of CNN architecture development in 2015



Evolution of ResNets

VGG-19 (2014) – “very deep convolutional neural network”
One of the most important outcomes of 2015: Residual Networks [1]



Main idea of a ResNet

Adding residual connections helps with degradation problem [2]

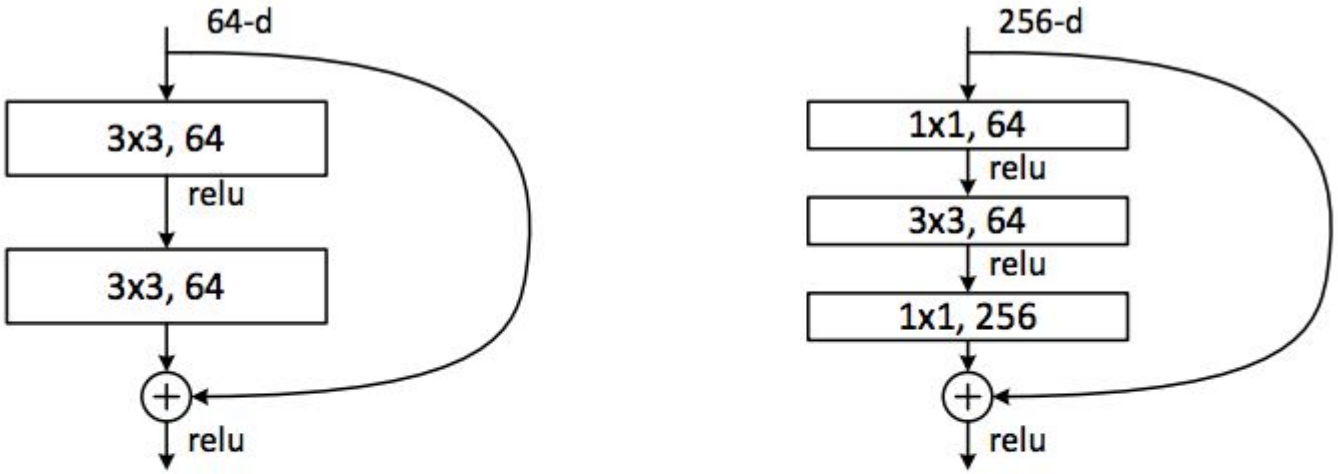


Figure 5. A deeper residual function $\mathcal{F}$ for ImageNet. Left: a building block (on 56×56 feature maps) as in Fig. 3 for ResNet-34. Right: a "bottleneck" building block for ResNet-50/101/152.

Summary of CNN architecture development in 2016

Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning (23 Feb 2016)

Inception ResNet > Inception v4 > Inception v3 [3]

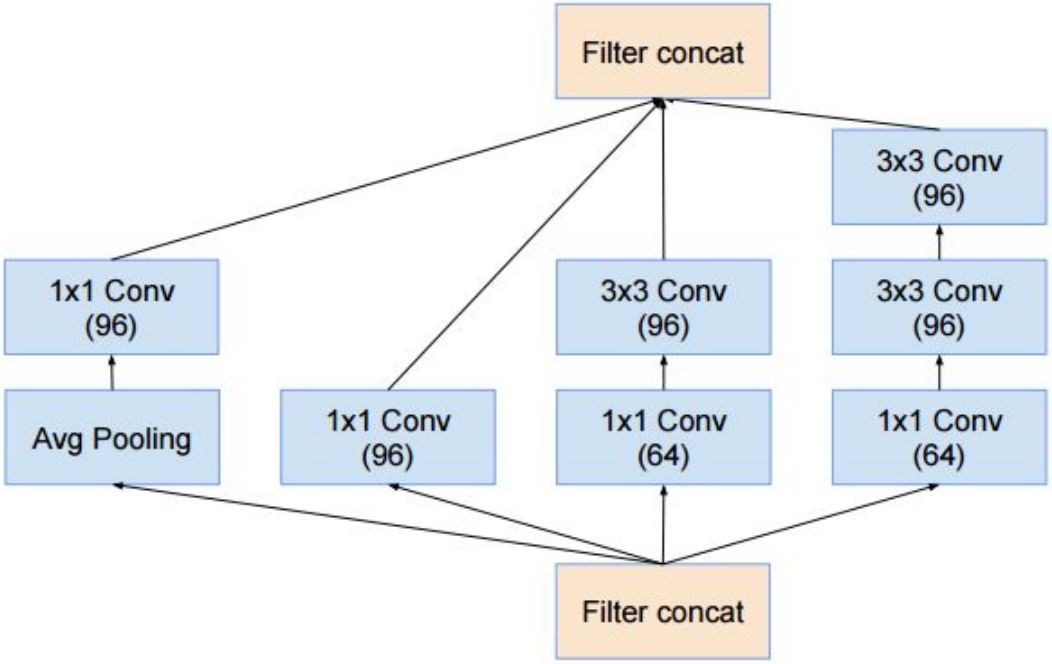


Figure 4. The schema for 35 × 35 grid modules of the pure Inception-v4 network. This is the Inception-A block of Figure 9.

Resnet in Resnet: Generalizing Residual Architectures (25 Mar 2016)

Generalized residual architecture that combines residual networks and standard convolutional networks in parallel [4]

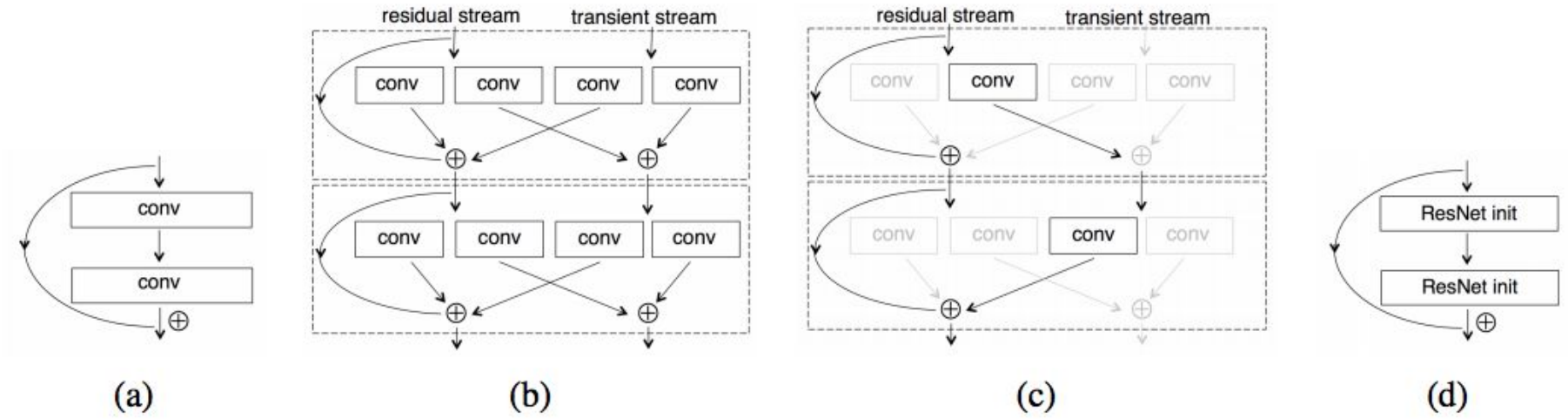


Figure 1: (a) 2-layer ResNet block. (b) 2 generalized residual blocks (ResNet Init). (c) 2-layer ResNet block from 2 generalized residual blocks (grayed out connections are 0). (d) 2-layer RiR block.

Deep Networks with Stochastic Depth (30 Mar 2016)

for each mini-batch, randomly drop a subset of layers – more than 1200 layers [5]

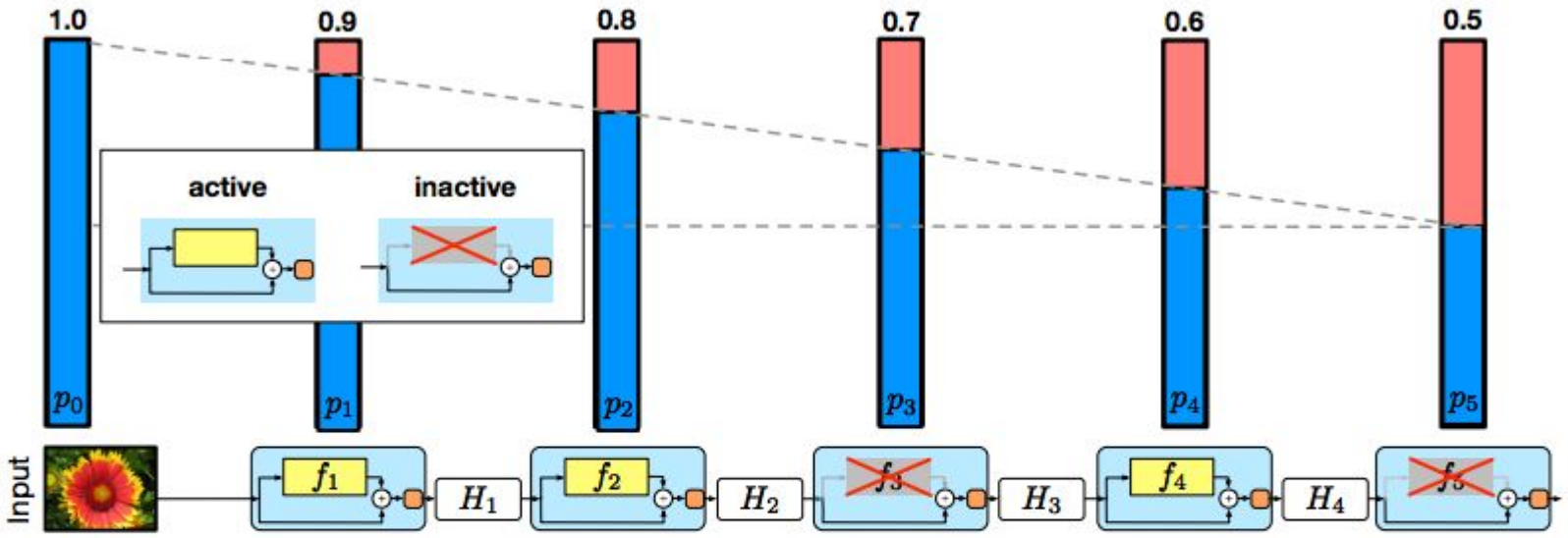


Fig. 2. The linear decay of p_ℓ illustrated on a ResNet with stochastic depth for $p_0 = 1$ and $p_L = 0.5$. Conceptually, we treat the input to the first ResBlock as H_0, which is always active.

Identity Mappings in Deep Residual Networks (12 Apr 2016)

Relocation of ReLU/BN [6]

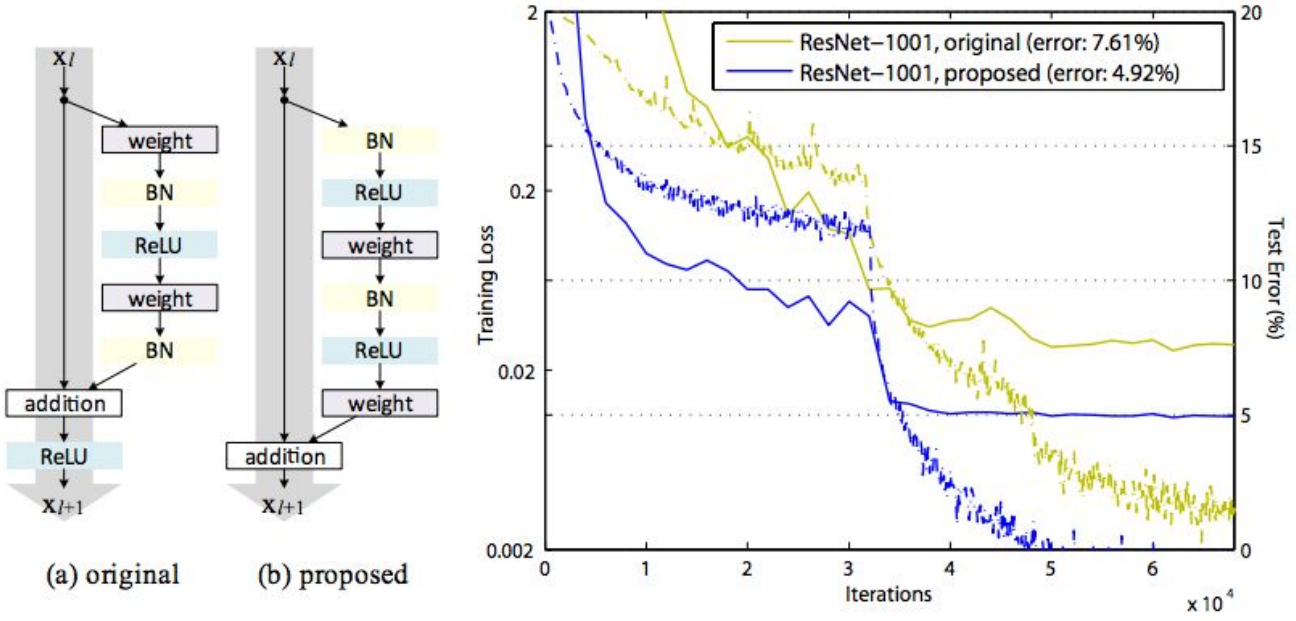
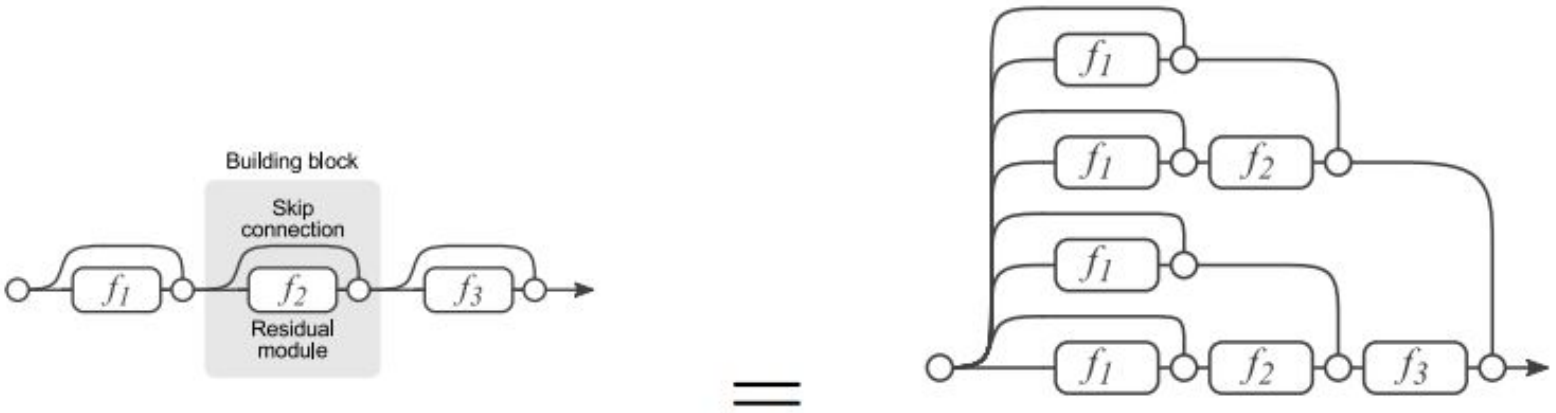


Figure 1. Left: (a) original Residual Unit in [1]; (b) proposed Residual Unit. The grey arrows indicate the easiest paths for the information to propagate, corresponding to the additive term "$\mathbf{x}_l$" in Eqn.(4) (forward propagation) and the additive term "1" in Eqn.(5) (backward propagation). **Right**: training curves on CIFAR-10 of **1001-layer** ResNets. Solid lines denote test error (y-axis on the right), and dashed lines denote training loss (y-axis on the left). The proposed unit makes ResNet-1001 easier to train.

Residual Networks are Exponential Ensembles of Relatively Shallow Networks (20 May 2016)

Describes “multiplicity” of ResNets and shows that removing layers leads to smooth increase in error rates [7]



(a) Conventional 3-block residual network

(b) Unraveled view of (a)

Figure 1: Residual Networks are conventionally shown as (a), which is a natural representation of Equation (1). When we expand this formulation to Equation (6), we obtain an *unraveled view* of a 3-block residual network (b). From this view, it is apparent that residual networks have $O(2^n)$ implicit paths connecting input and output and that adding a block doubles the number of paths.

Wide Residual Networks (23 May 2016)

Decrease depth and increase width: new SOTA on on CIFAR-10, CIFAR-100 and SVHN [8]

group name	output size	block type = $B(3,3)$
conv1	32×32	[3×3, 16]
conv2	32×32	$\begin{bmatrix} 3\times3, 16\times k \\ 3\times3, 16\times k \end{bmatrix} \times N$
conv3	16×16	$\begin{bmatrix} 3\times3, 32\times k \\ 3\times3, 32\times k \end{bmatrix} \times N$
conv4	8×8	$\begin{bmatrix} 3\times3, 64\times k \\ 3\times3, 64\times k \end{bmatrix} \times N$
avg-pool	1×1	$[8 \times 8]$

Table 1: Structure of wide residual networks. Network width is determined by factor k. Original architecture [▭] is equivalent to $k = 1$. Groups of convolutions are shown in brackets where N is a number of blocks in group, downsampling performed by the first layers in groups `conv3` and `conv4`. Final classification layer is omitted for clearance. In the particular example shown, the network uses a ResNet block of type $B(3,3)$.

FractalNet: Ultra-Deep Neural Networks without Residuals (24 May 2016)

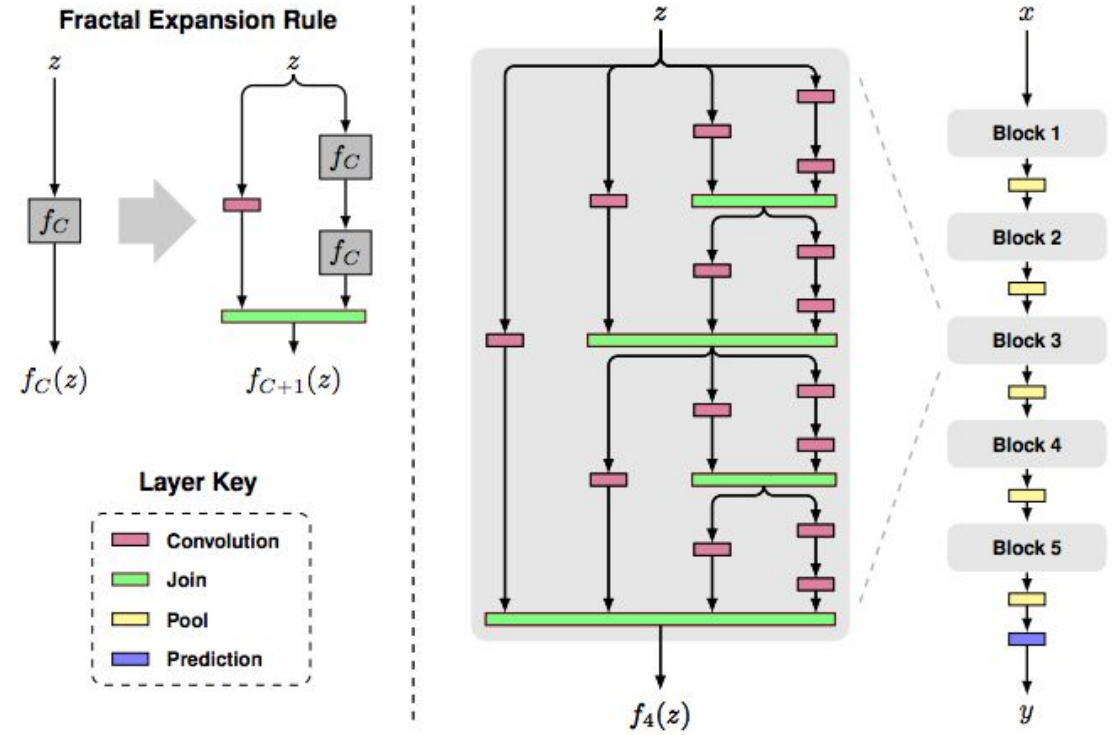


Figure 1: **Fractal architecture.** *Left:* A simple expansion rule generates a fractal architecture with C intertwined columns. The base case, $f_1(z)$, has a single layer of the chosen type (*e.g.* convolutional) between input and output. Join layers compute element-wise mean. *Right:* Deep convolutional networks periodically reduce spatial resolution via pooling. A fractal version uses f_C as a building block between pooling layers. Stacking B such blocks yields a network whose total depth, measured in terms of convolution layers, is $B \cdot 2^{C-1}$. This example has depth 40 ($B = 5, C = 4$).

Residual Networks of Residual Networks: Multilevel Residual Networks (9 Aug 2016)

New SOTA on CIFAR 10/100 and SVHN by combining wide ResNets with additional level-wise shortcut connections [10]

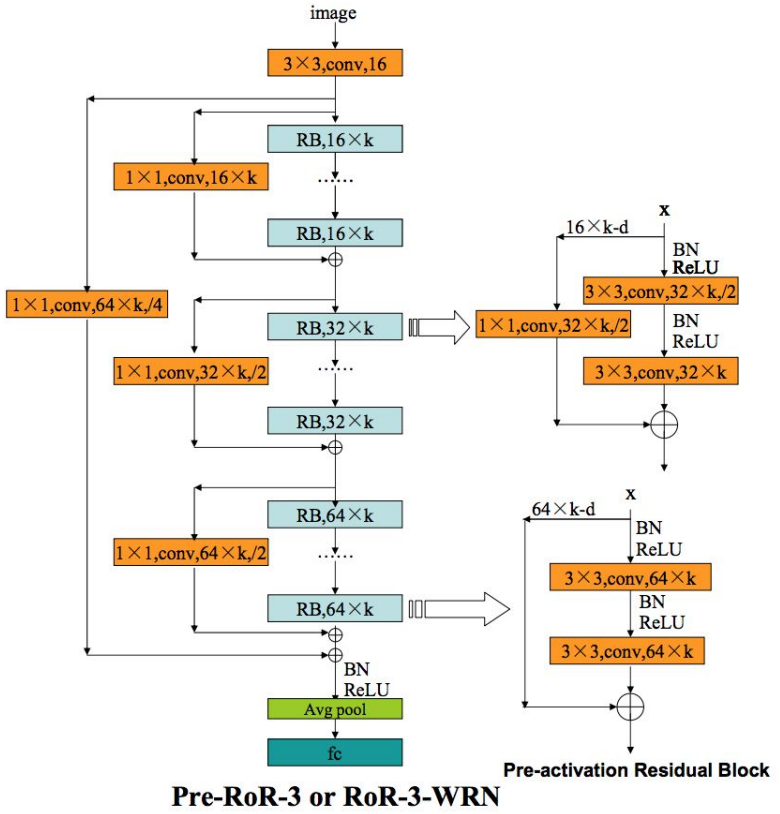


Fig. 3. Pre-RoR-3 and RoR-3-WRN architectures. m=3. The addition is followed by the ReLU. Projection shortcut is done by 1×1 convolutions. BN-ReLU-conv order in residual blocks is adopted. If k=1, this is a Pre-RoR-3 architecture, otherwise this is a RoR-3-WRN architecture. There are several direct paths for propagating information created by identity mappings.

Densely Connected Convolutional Networks (25 Aug 2016)

Each layer is directly connected to every other layer in a feed-forward fashion

New SOTA on CIFAR 10/100 and SVHN. [11]

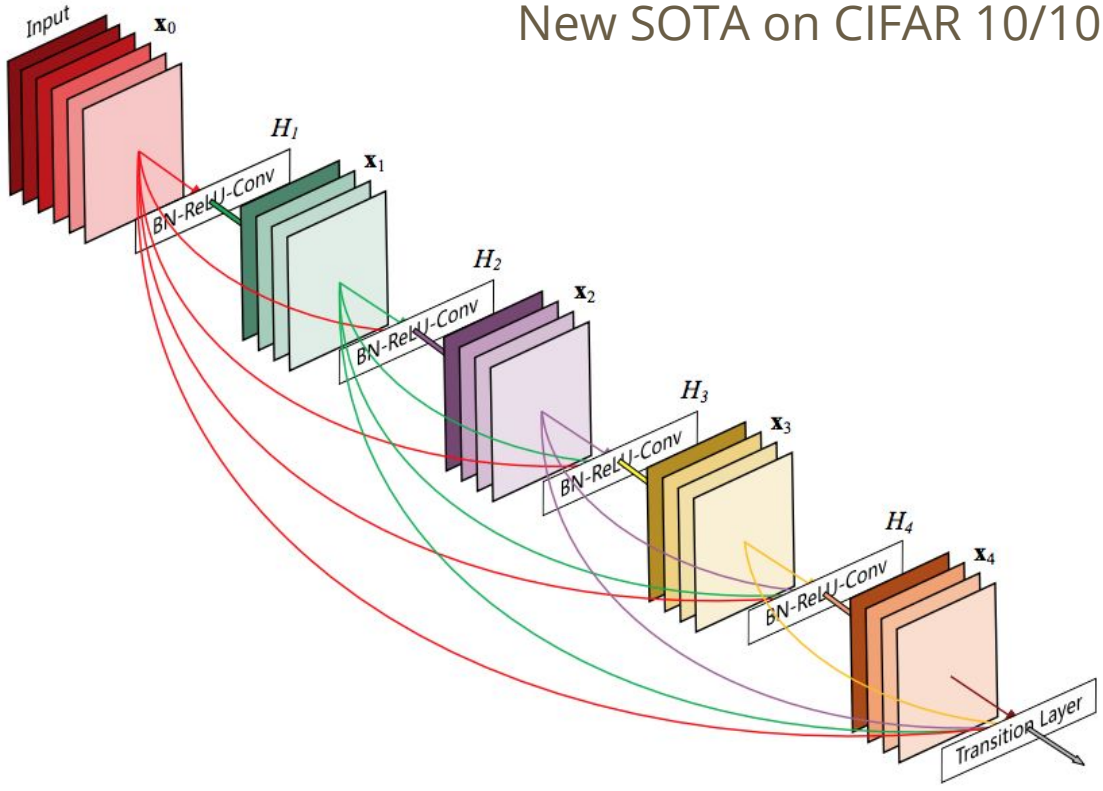


Figure 1: A 5-layer dense block with a growth rate of $k = 4$. Each layer takes all preceding feature maps as input.

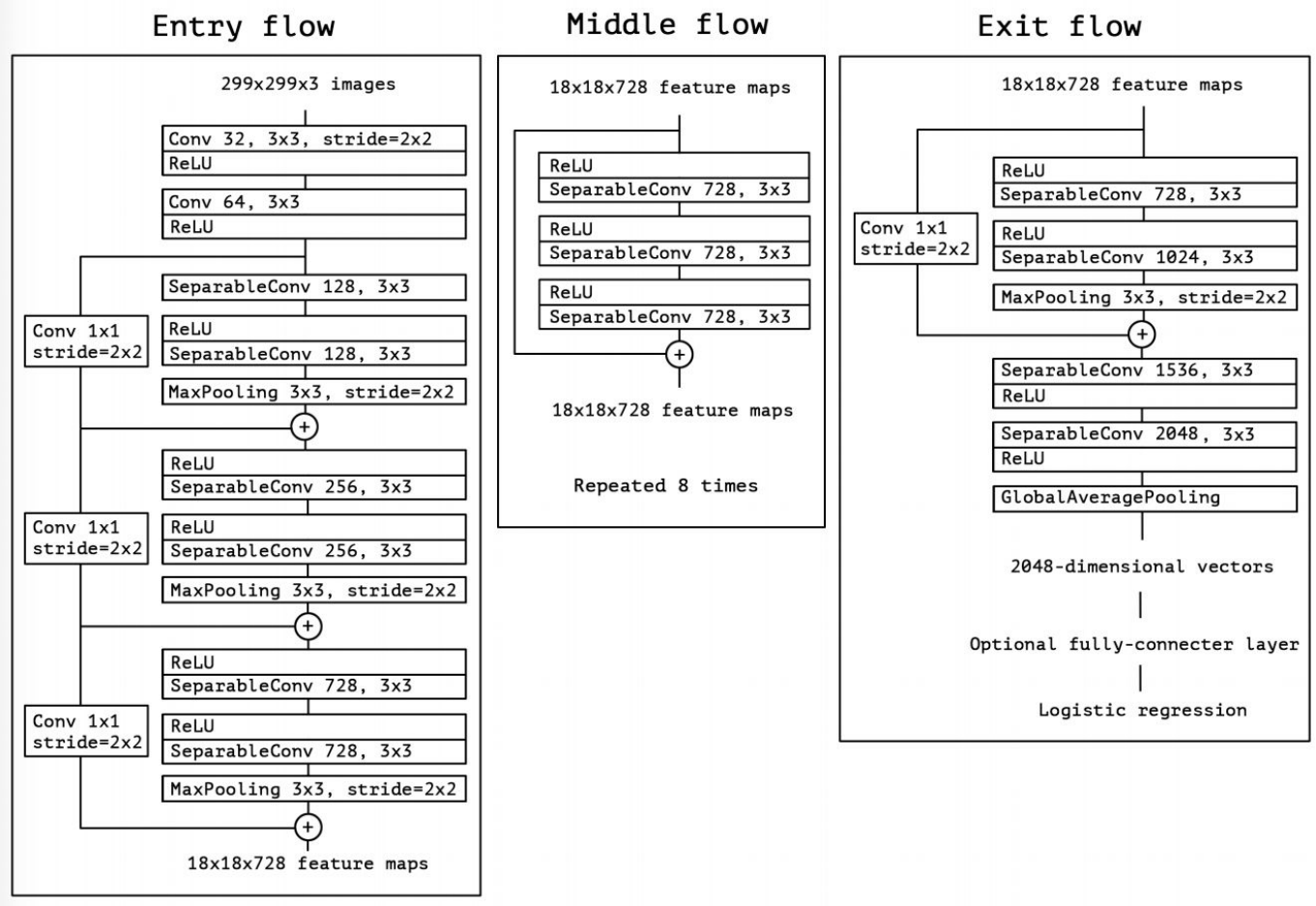
Large Scale Visual Recognition Challenge 2016, ILSVRC2016 (26 Sep 2016)

- Inception, Inception-Resnet, ResNet & Wide Residual Network (WRN)
- Faster R-CNN
- Ensembling [12]

Xception: Deep Learning with Depthwise Separable Convolutions (7 Oct 2016)

Replaces Inception modules with depthwise separable convolutions.

Outperforms V3 with the same # of parameters [13]



Deep Pyramidal Residual Networks (10 Oct 2016)

instead of using downsampling increases the width at all the units to involve as many locations as possible [14]

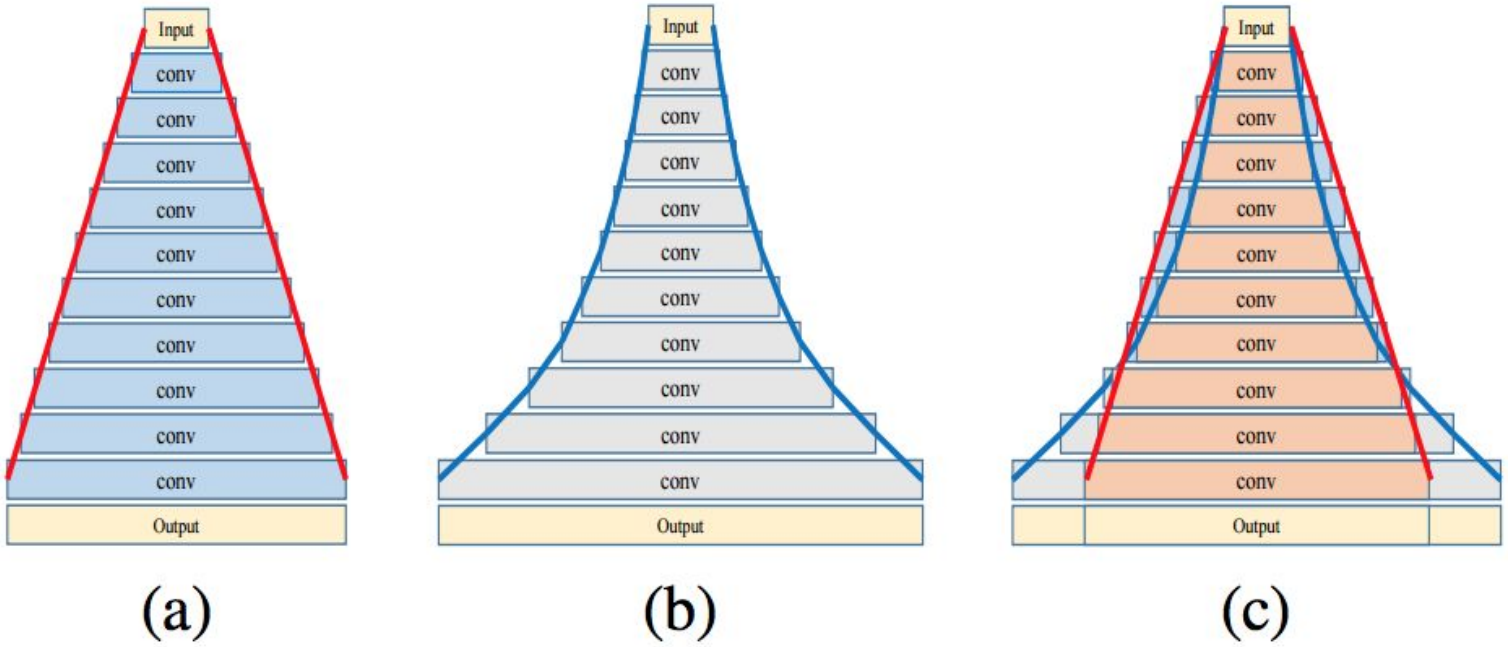


Figure 2. Visual illustrations of (a) additive PyramidNet, (b) multiplicative PyramidNet, and (c) a comparison of (a) and (b).

Generative models in 2016

- VAE [15]
- GAN [16,17]
- PixelRNN & PixelCNN [18,19]

Autoencoding beyond pixels using a learned similarity metric (10 Feb 2016)

Combining VAE with a GAN to use learned feature representations in the GAN discriminator as basis for the VAE reconstruction objective [20]

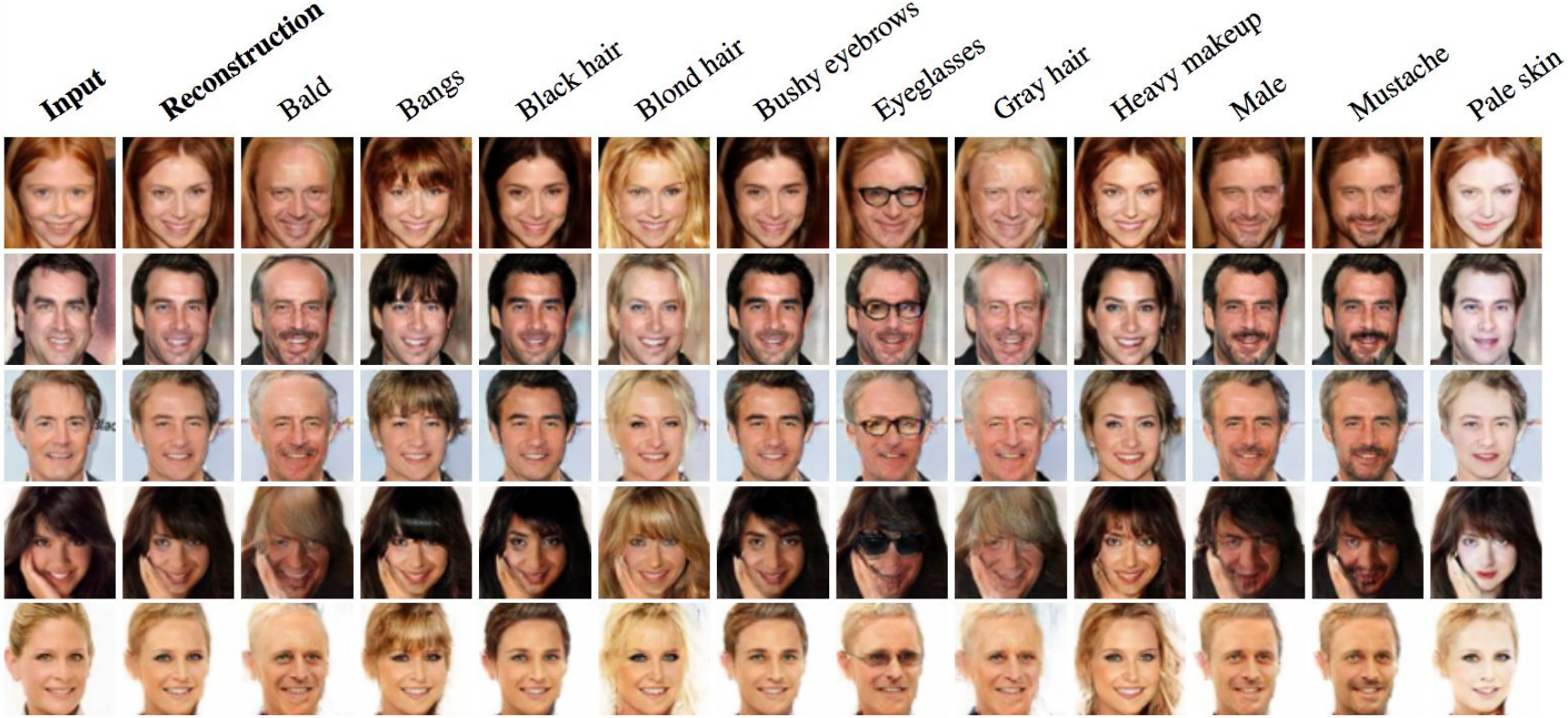


Figure 5. Using the VAE/GAN model to reconstruct dataset samples with visual attribute vectors added to their latent representations.

Texture Networks: Feed-forward Synthesis of Textures and Stylized Images (10 Mar 2016)

compact feed-forward CNNs to generate multiple samples of the same texture of arbitrary size to transfer artistic style from a given image to any other image (100x speed up) [21]

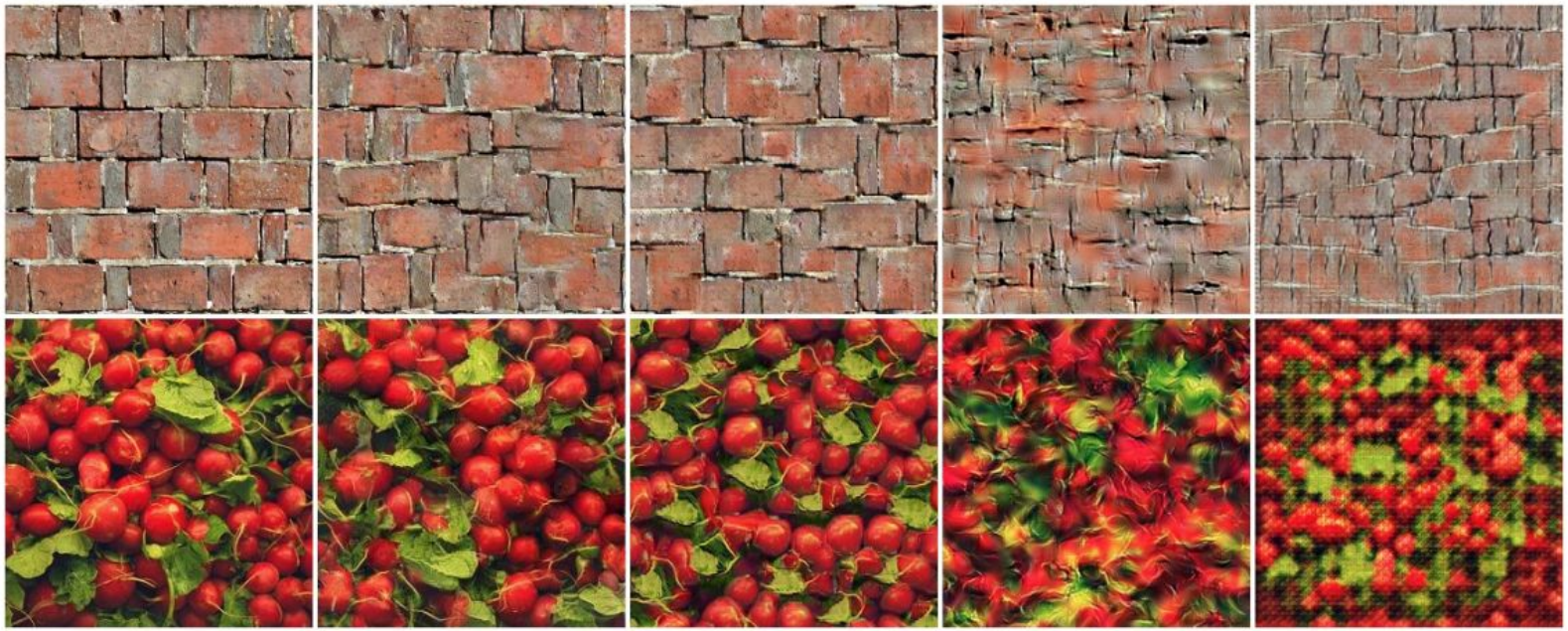


Figure 4. Further comparison of textures generated with several methods including the original statistics matching method (Portilla & Simoncelli, 2000) and the DCGAN (Radford et al., 2015) approach. Overall, our method and (Gatys et al., 2015a) provide better results, our method being hundreds times faster.

Perceptual Losses for Real-Time Style Transfer and Super-Resolution (27 Mar 2016)

Suggests 2 perceptual loss functions combining per-pixel loss between the output and ground-truth images with high-level features [22]

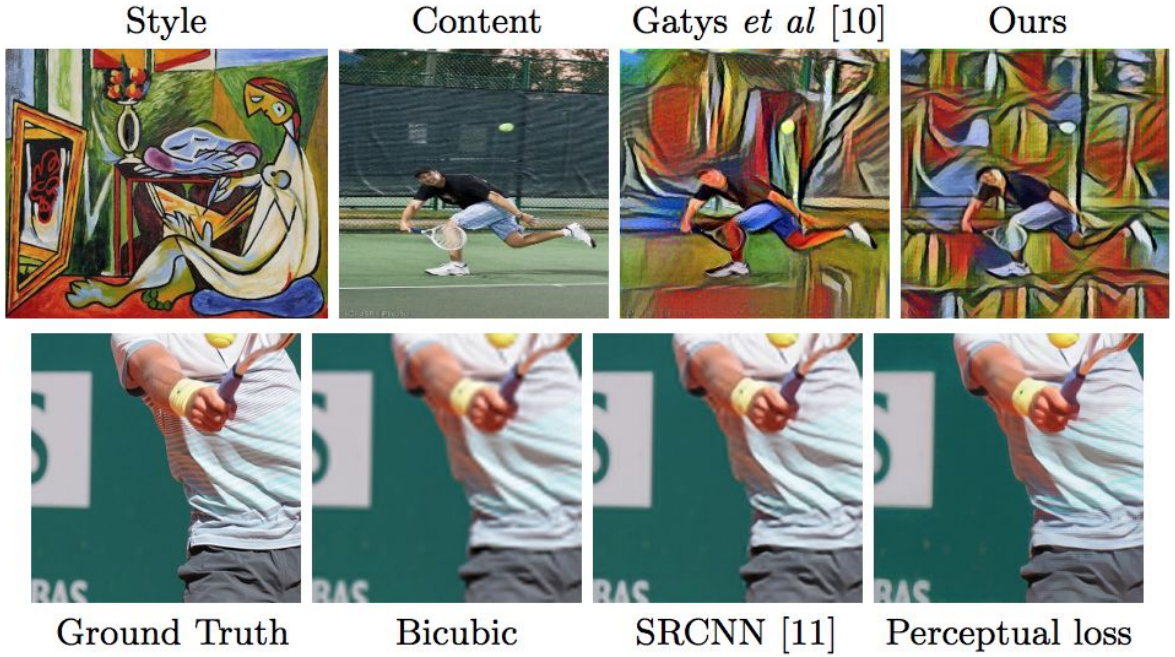


Fig. 1. Example results for style transfer (top) and ×4 super-resolution (bottom). For style transfer, we achieve similar results as Gatys *et al* [10] but are three orders of magnitude faster. For super-resolution our method trained with a perceptual loss is able to better reconstruct fine details compared to methods trained with per-pixel loss.

Context Encoders: Feature Learning by Inpainting (25 Apr 2016

Unsupervised visual feature learning algorithm driven by context-based pixel prediction [23]

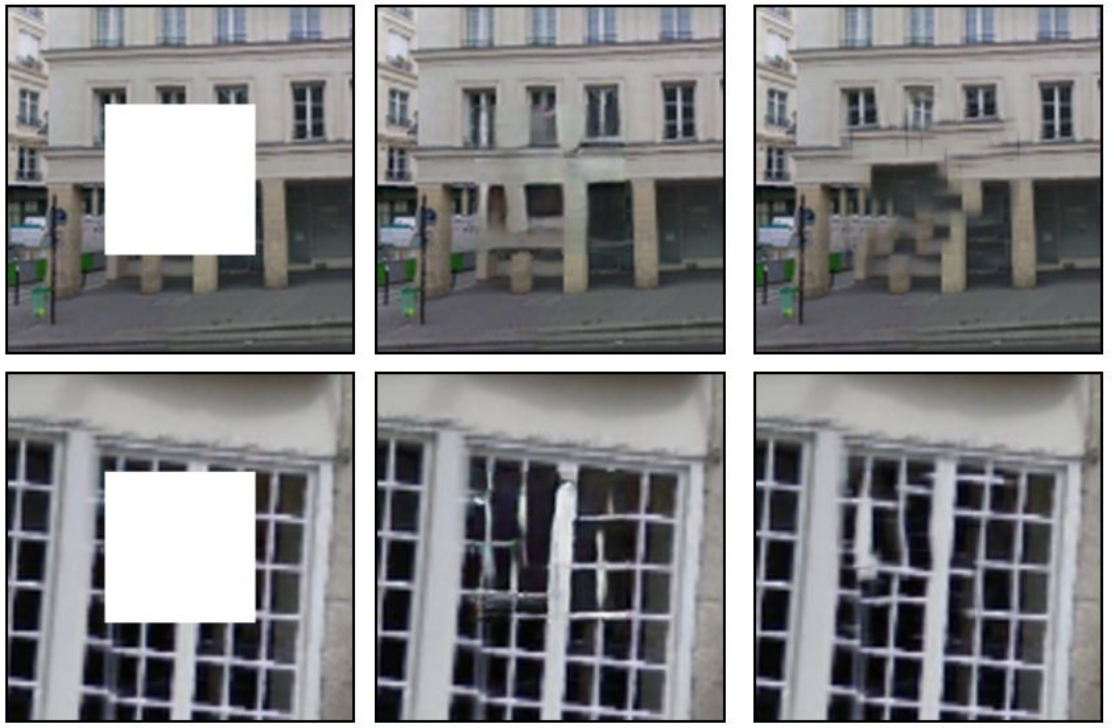


Figure 5: Comparison with Content-Aware Fill (Photoshop feature based on [2]) on *held-out* images. Our method works better in semantic cases (top row) and works slightly worse in textured settings (bottom row).

Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network (15 Sep 2016)

superresolution generative adversarial network (SRGAN) with a perceptual loss function which consists of an adversarial loss and a content loss [24]

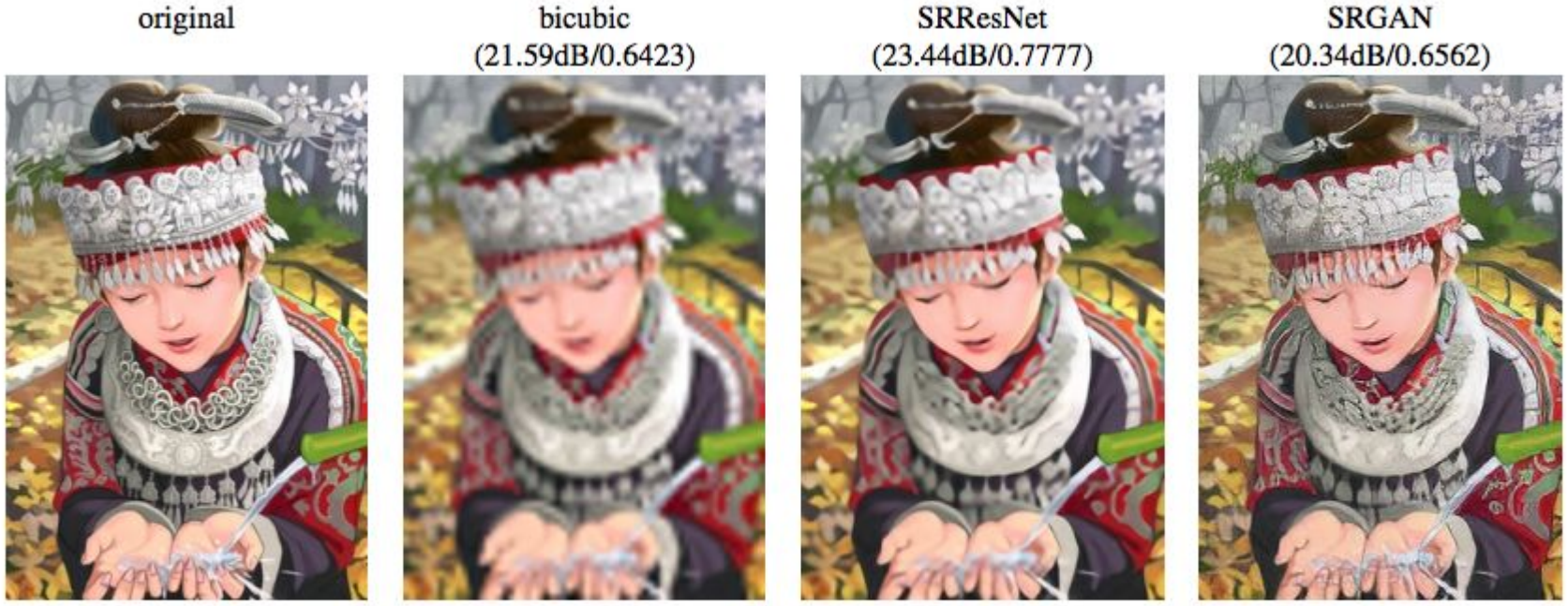


Figure 2: Illustration of performance of different SR approaches with downsampling factor: 4×. From left to right: original HR image, bicubic interpolation, deep residual network optimized for MSE, deep residual generative adversarial network optimized for a loss more sensitive to human perception. Corresponding PSNR and SSIM are shown in brackets.

And much, much more...

- SqueezeNet [25] and XNOR-net [26]
- Weight [27] and Layer [28] normalization
- 3D ConvNets [29]
- WaveNet [30]
- etc

References

[1] He, Kaiming, et al. Deep Residual Learning. MSRA @ ILSVRC & COCO 2015 competitions. URL
[2] He, Kaiming, et al. "Deep residual learning for image recognition." arXiv preprint arXiv:1512.03385 (2015).
[3] Szegedy, Christian, Sergey Ioffe, and Vincent Vanhoucke. "Inception-v4, inception-resnet and the impact of residual connections on learning." arXiv preprint arXiv:1602.07261 (2016).
[4] Targ, Sasha, Diogo Almeida, and Kevin Lyman. "Resnet in Resnet: Generalizing Residual Architectures." arXiv preprint arXiv:1603.08029 (2016).
[5] Huang, Gao, et al. "Deep networks with stochastic depth." arXiv preprint arXiv:1603.09382 (2016).
[6] He, Kaiming, et al. "Identity mappings in deep residual networks." arXiv preprint arXiv:1603.05027 (2016).
[7] Veit, Andreas, Michael Wilber, and Serge Belongie. "Residual Networks are Exponential Ensembles of Relatively Shallow Networks." arXiv preprint arXiv:1605.06431 (2016).
[8] Zagoruyko, Sergey, and Nikos Komodakis. "Wide Residual Networks." arXiv preprint arXiv:1605.07146 (2016).
[9] Larsson, Gustav, Michael Maire, and Gregory Shakhnarovich. "FractalNet: Ultra-Deep Neural Networks without Residuals." arXiv preprint arXiv:1605.07648 (2016).

References

[10] Zhang, Ke, et al. "Residual Networks of Residual Networks: Multilevel Residual Networks." arXiv preprint arXiv:1608.02908 (2016).
[11] Huang, Gao, Zhuang Liu, and Kilian Q. Weinberger. "Densely connected convolutional networks." arXiv preprint arXiv:1608.06993 (2016).
[12] http://image-net.org/challenges/LSVRC/2016/
[13] Chollet, François. "Xception: Deep Learning with Separable Convolutions." arXiv preprint arXiv:1610.02357 (2016).
[14] Han, Dongyoon, Jiwhan Kim, and Junmo Kim. "Deep Pyramidal Residual Networks." arXiv preprint arXiv:1610.02915 (2016).
[15] Eslami, S. M., et al. "Attend, Infer, Repeat: Fast Scene Understanding with Generative Models." arXiv preprint arXiv:1603.08575 (2016).
[16] Salimans, Tim, et al. "Improved techniques for training gans." arXiv preprint arXiv:1606.03498 (2016).
[17] Zhao, Junbo, Michael Mathieu, and Yann LeCun. "Energy-based Generative Adversarial Network." arXiv preprint arXiv:1609.03126 (2016).
[18] van den Oord, Aaron, Nal Kalchbrenner, and Koray Kavukcuoglu. "Pixel Recurrent Neural Networks." arXiv preprint arXiv:1601.06759 (2016).
[19] Oord, Aaron van den, et al. "Conditional image generation with pixelcnn decoders." arXiv preprint arXiv:1606.05328 (2016).

References

[20] Larsen, Anders Boesen Lindbo, Søren Kaae Sønderby, and Ole Winther. "Autoencoding beyond pixels using a learned similarity metric." arXiv preprint arXiv:1512.09300 (2015).
[21] Ulyanov, Dmitry, et al. "Texture Networks: Feed-forward Synthesis of Textures and Stylized Images." arXiv preprint arXiv:1603.03417 (2016).
[22] Johnson, Justin, Alexandre Alahi, and Li Fei-Fei. "Perceptual losses for real-time style transfer and super-resolution." arXiv preprint arXiv:1603.08155 (2016).
[23] Pathak, Deepak, et al. "Context Encoders: Feature Learning by Inpainting." arXiv preprint arXiv:1604.07379 (2016).
[24] Ledig, Christian, et al. "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network." arXiv preprint arXiv:1609.04802 (2016).
[25] Iandola, Forrest N., et al. "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 1MB model size." arXiv preprint arXiv:1602.07360 (2016).
[26] Rastegari, Mohammad, et al. "XNOR-Net: ImageNet Classification Using Binary Convolutional Neural Networks." arXiv preprint arXiv:1603.05279 (2016).
[27] Salimans, Tim, and Diederik P. Kingma. "Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks." arXiv preprint arXiv:1602.07868 (2016).
[28] Ba, Jimmy Lei, Jamie Ryan Kiros, and Geoffrey E. Hinton. "Layer normalization." arXiv preprint arXiv:1607.06450 (2016).

References

[29] Qi, Charles R., et al. "Volumetric and Multi-View CNNs for Object Classification on 3D Data." arXiv preprint arXiv:1604.03265 (2016).
[30] Oord, Aaron van den, et al. "WaveNet: A generative model for raw audio." arXiv preprint arXiv:1609.03499 (2016).