Analyzing Artifacts in Discriminative and Generative Models

Richard Zhang (章睿嘉)

Research Scientist, Adobe SF

GAMES Webinar

Aug 2020

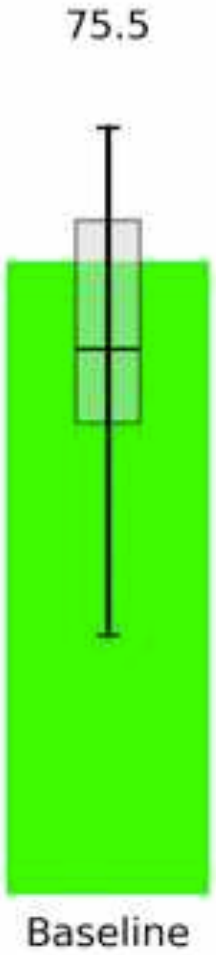
Example classifications

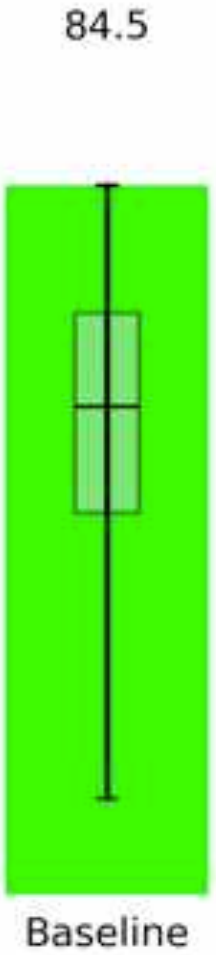
P(correct class)

P(correct class)

Deep Networks are <u>not</u> Shift-Invariant

Deep Networks are <u>not</u> Shift-Invariant





c.f. Azulay and Weiss. **Why do deep convolutional networks generalize so poorly to small image transformations?** Arxiv, 2018. JMLR, 2019.
Engstrom, Tsipras, Schmidt, Madry. **Exploring the Landscape of Spatial Robustness.** Arxiv, 2017. ICML, 2019.

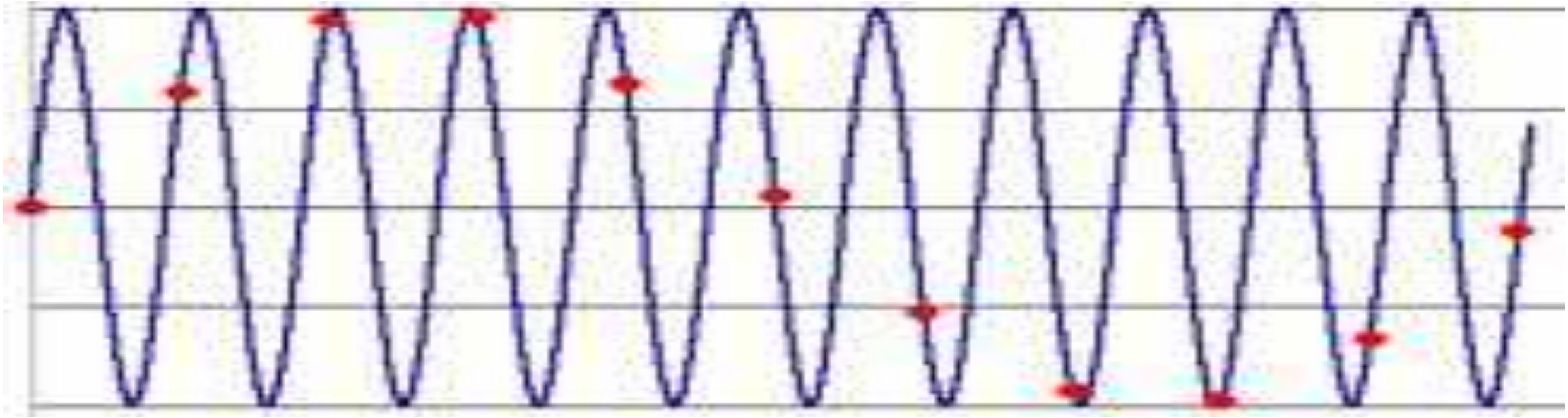
Why is shift-invariance lost?

"Convolutions are shift-equivariant"

"Pooling builds up shift-invariance"

…but striding ignores Nyquist sampling theorem and aliases

Nyquist-Shannon theorem



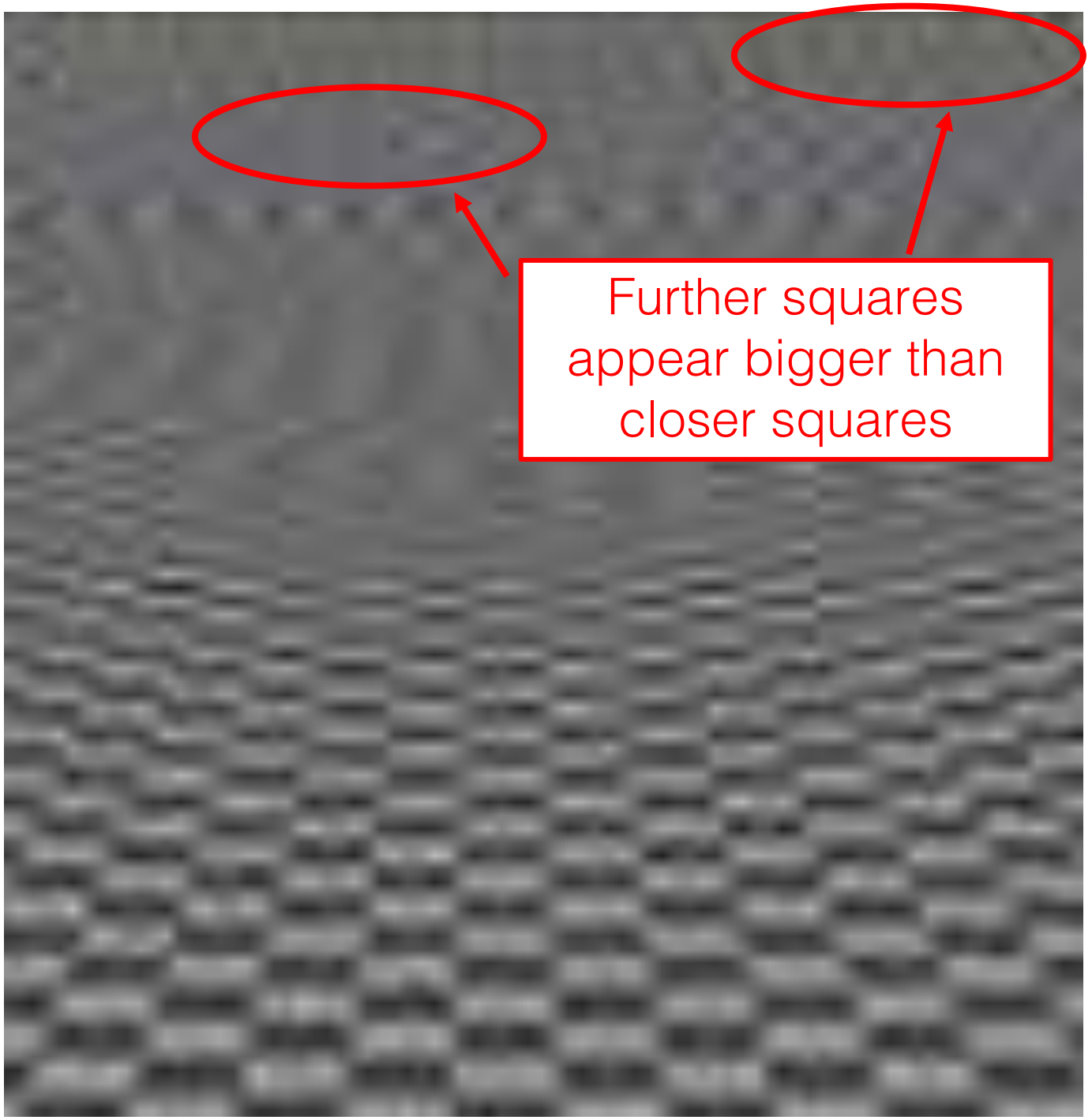
If the sampling frequency is less than twice the underlying signal frequency, the samples **cannot faithfully represent** the underlying signal…

What is aliasing?

…worse, the samples will **misrepresent** the underlying signal

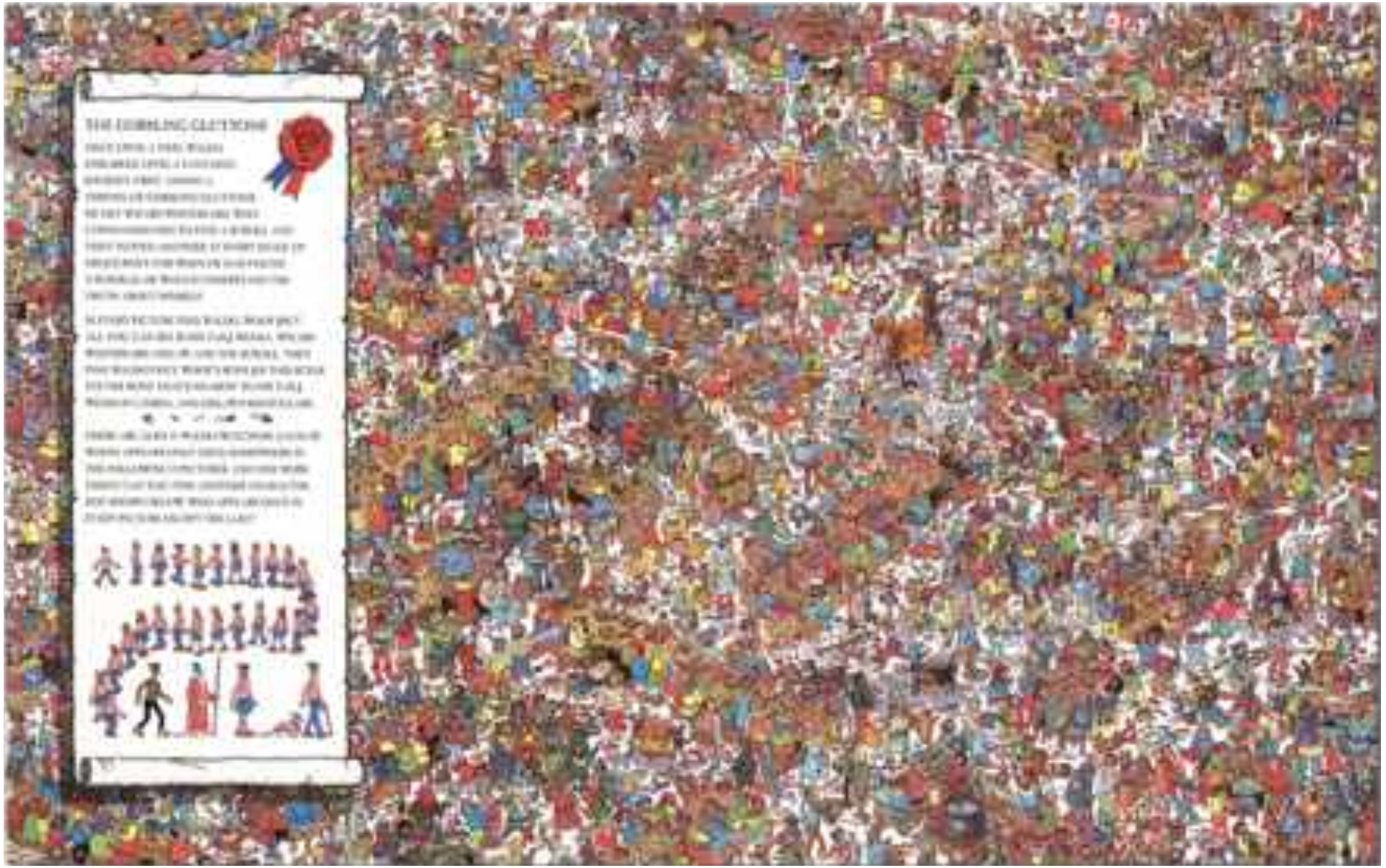
Rendering

Further away, decreased sampling rate



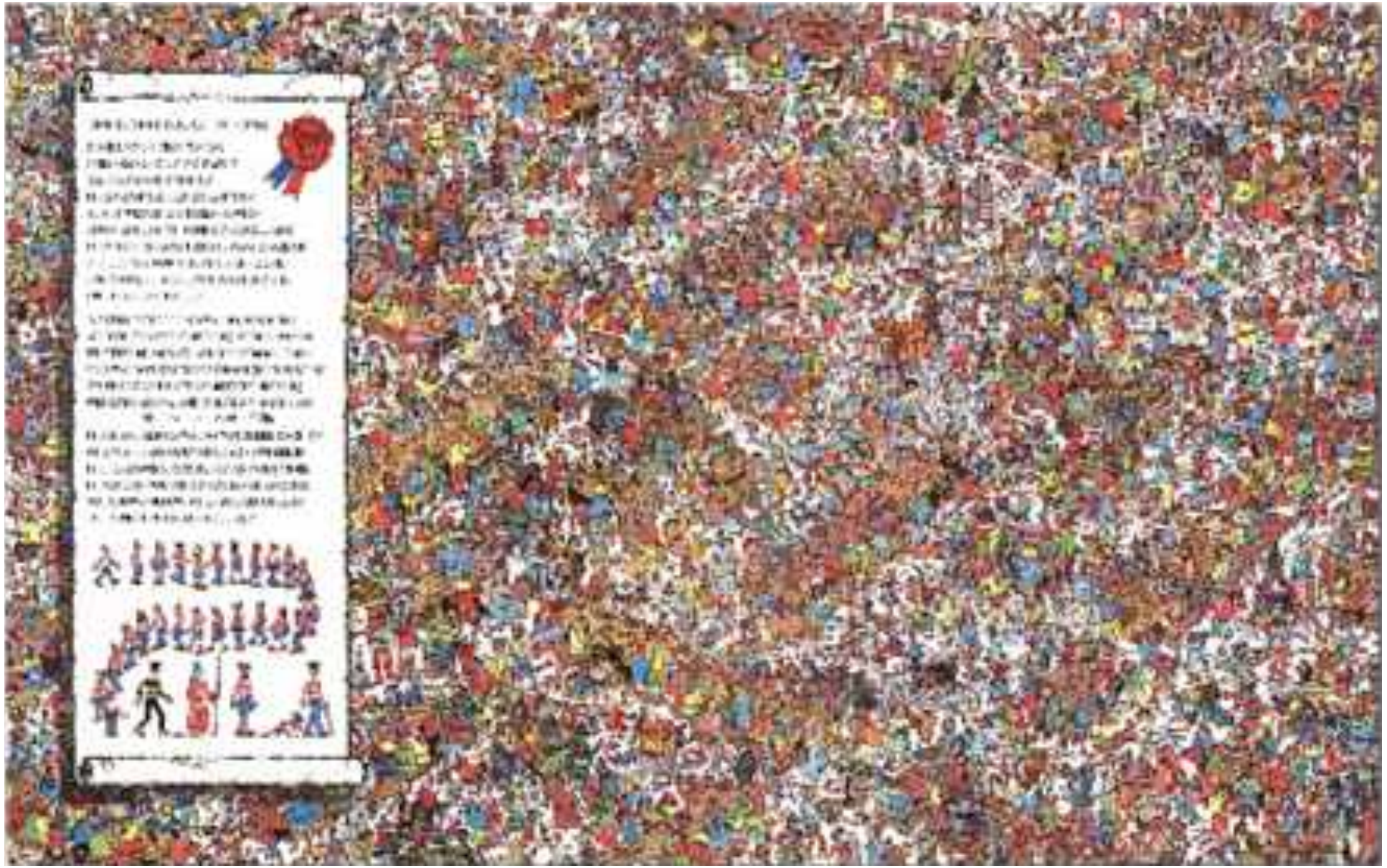
Temporal aliasing

Image



Adapted from Lectures on Digital Photography. Marc Levoy. https://sites.google.com/site/marclevoylectures/schedule

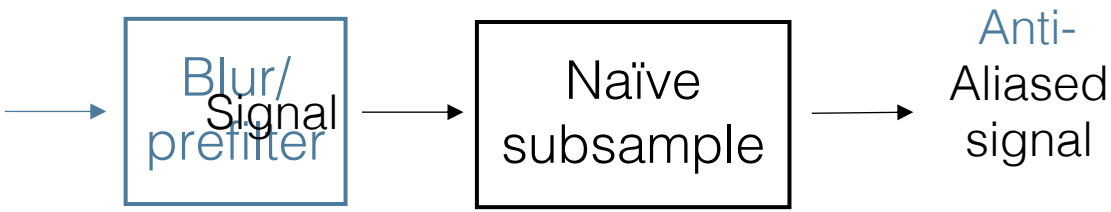
4x naive sub-sampling



Adapted from Lectures on Digital Photography. Marc Levoy. https://sites.google.com/site/marclevoylectures/schedule

Antialiasing

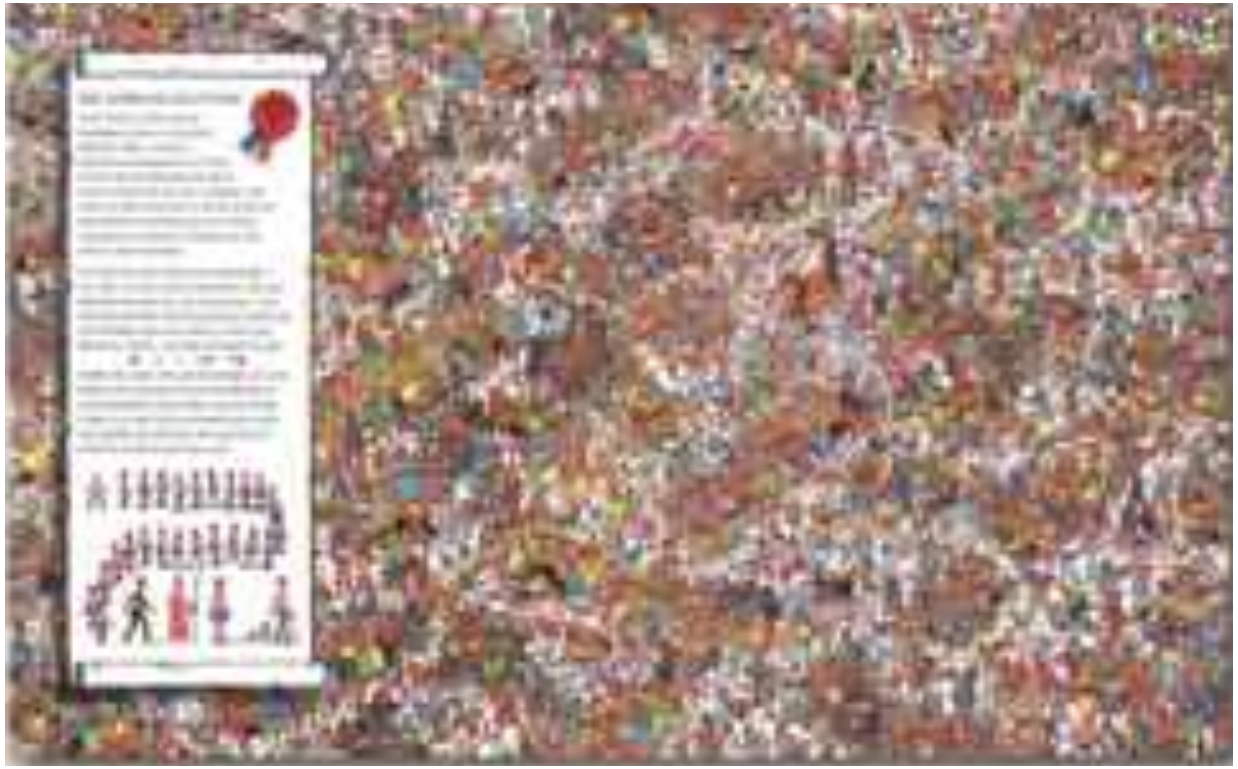
- Low sampling rate → signal cannot be represented
  - Just don't downsample? Expensive…
- Naïve subsampling → high freqs **misrepresented**
- Before subsampling, antialias by blurring → high freqs **unrepresented**
  - Better than aliasing!
  - Blur / prefilter / antialias / low-pass filter



Adapted from Alexei A. Efros CS 194. https://inst.eecs.berkeley.edu//~cs194-26/fa18/

Effect of prefiltering

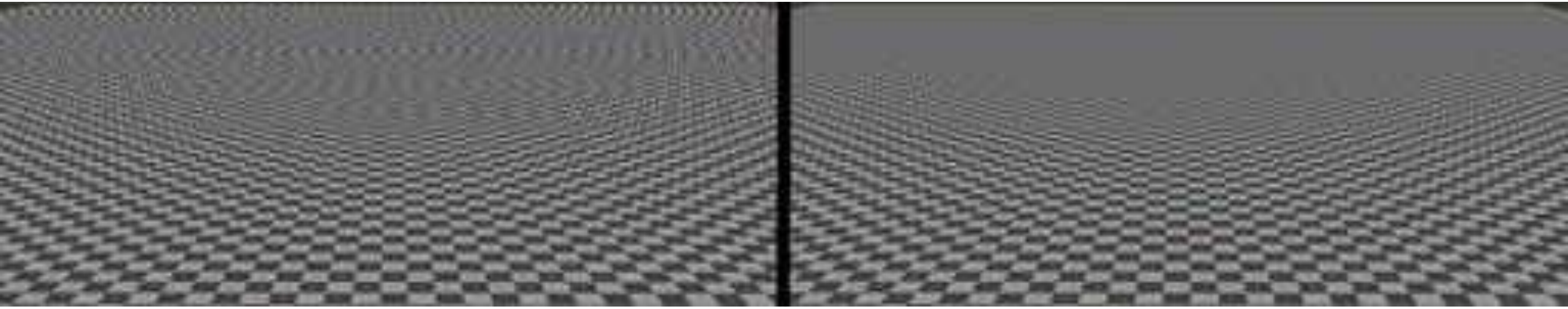
Naïvely subsampled



Prefiltered, then subsampled

Adapted from Lectures on Digital Photography. Marc Levoy. https://sites.google.com/site/marclevoylectures/schedule

Effect of prefiltering



Naïvely subsampled

Prefiltered, then subsampled

Adapted from Lectures on Digital Photography. Marc Levoy. https://sites.google.com/site/marclevoylectures/schedule

Aliasing → Loss of shift-equivariance

Shift to the right

Naive Prefiltered

from Jaakko Lehtinen https://twitter.com/jaakkolehtinen/status/1258102168176951299

Making Convolutional Networks Shift-Invariant Again

Richard Zhang

In **ICML**, 2019

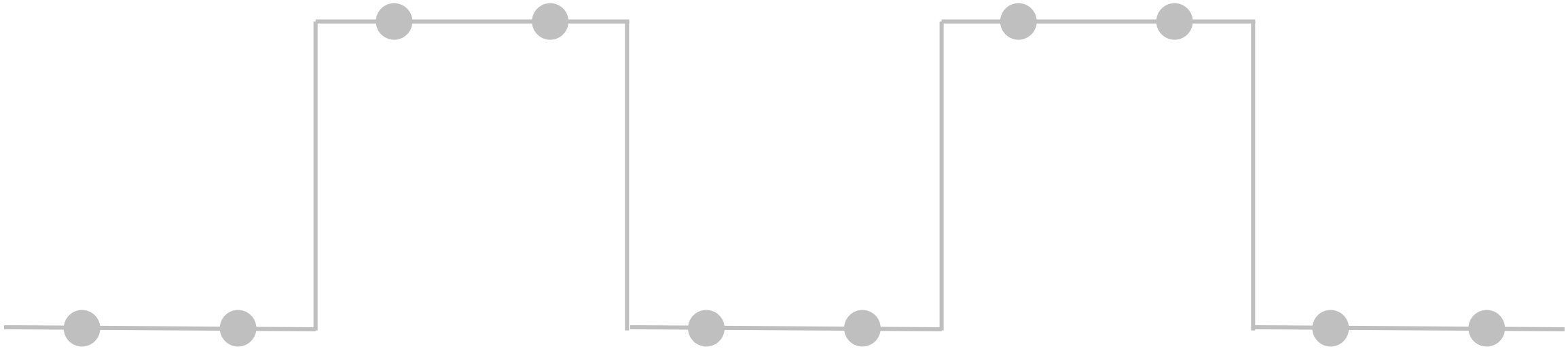
Alternative downsampling methods

- Signal processing; image processing; graphics
  - Prefilter before subsampling

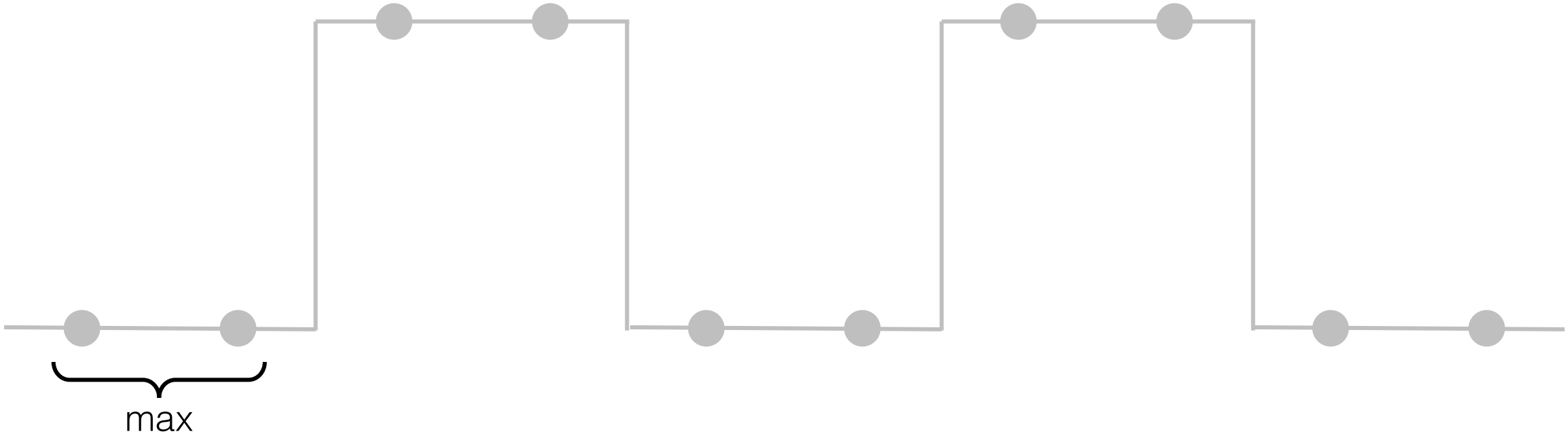
- Deep learning; computer vision
  - Szeliski vision book, Sec 2.3.1 “Sampling and aliasing”
  - Average pooling [LeNet 1989] does some antialiasing
  - Max-pooling gets better accuracy [Scherer 2010]

Antialiasing abandoned and forgotten

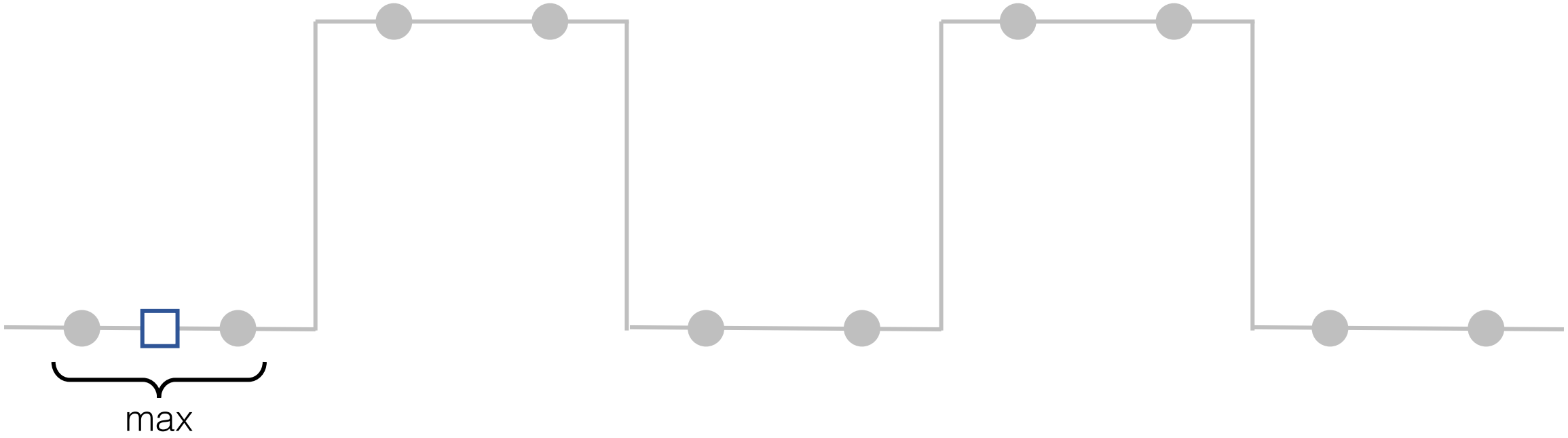
Re-examining Max-Pooling



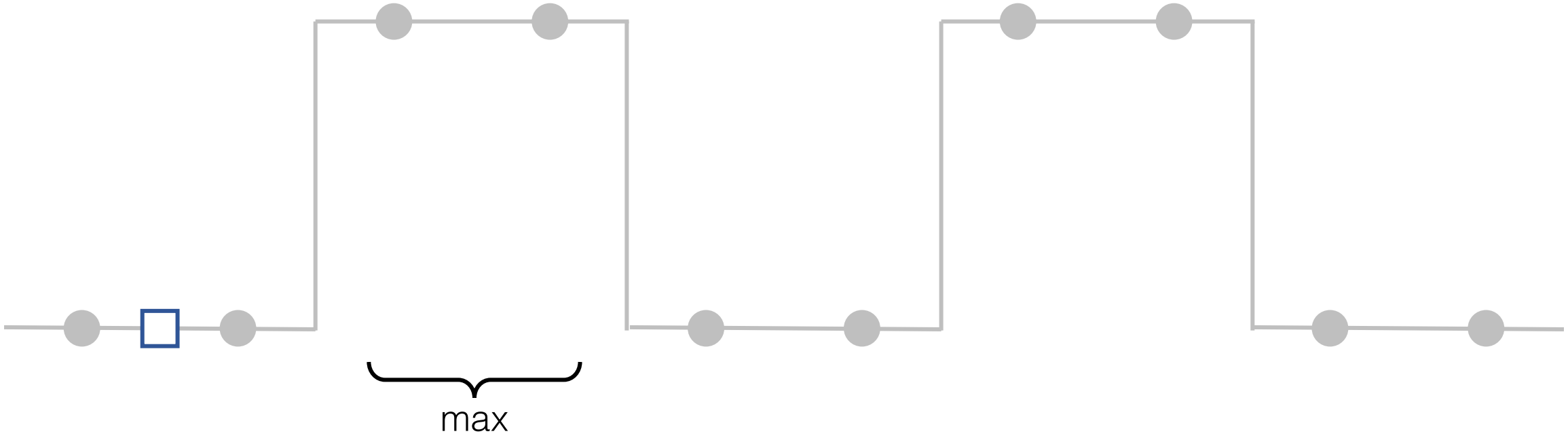
Re-examining Max-Pooling



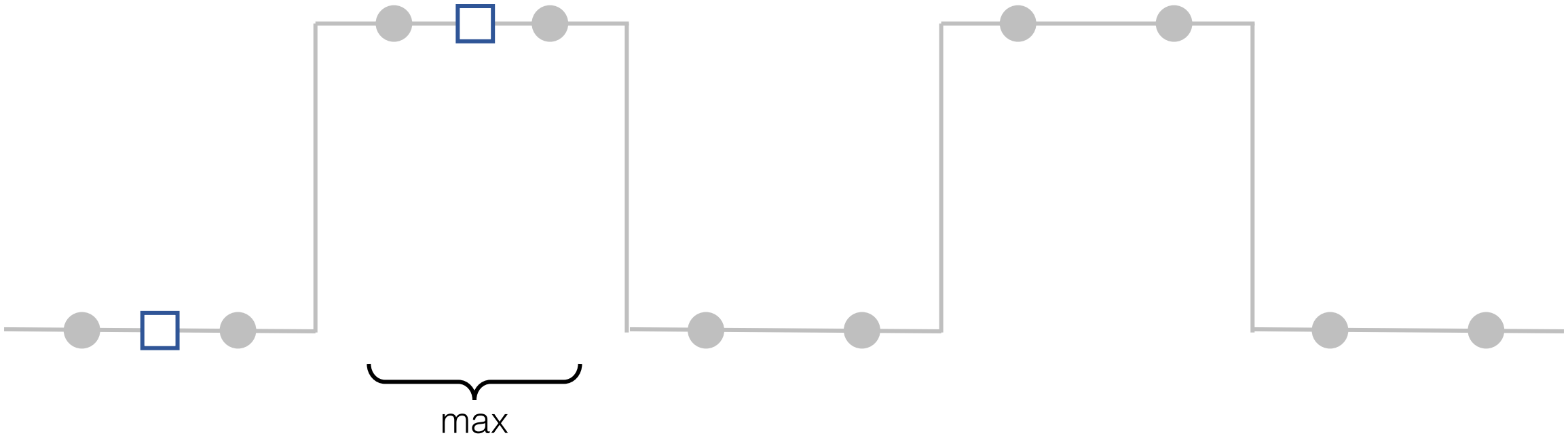
Re-examining Max-Pooling



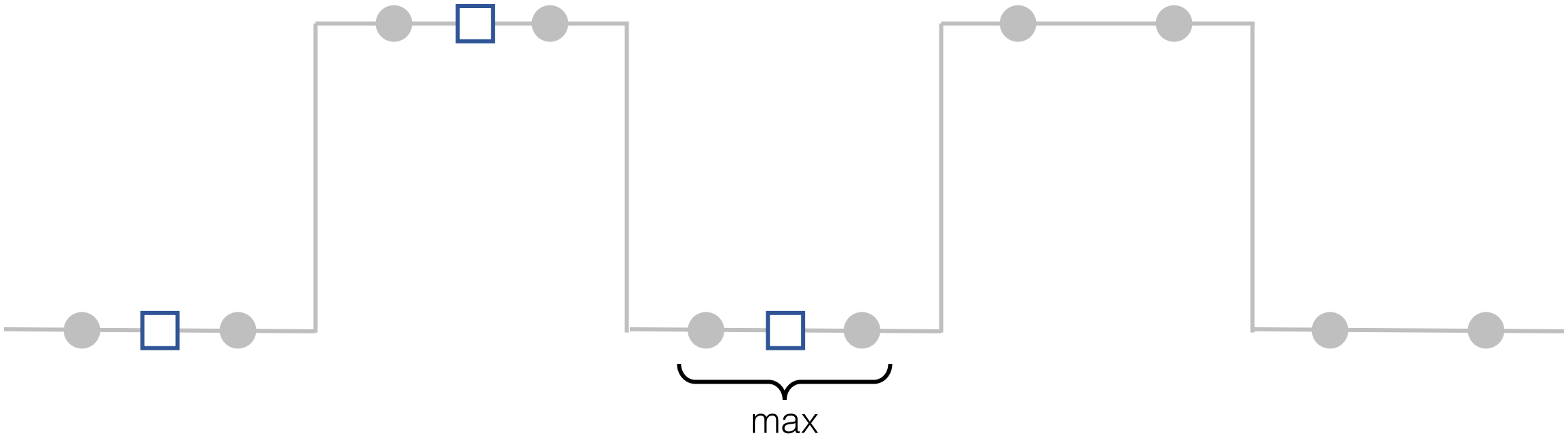
Re-examining Max-Pooling



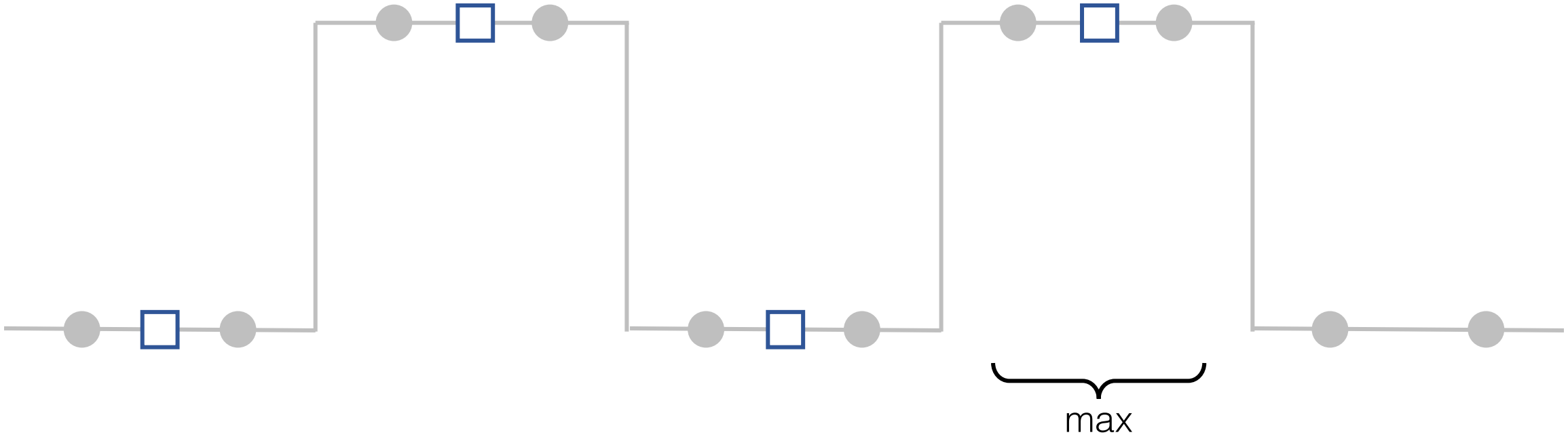
Re-examining Max-Pooling



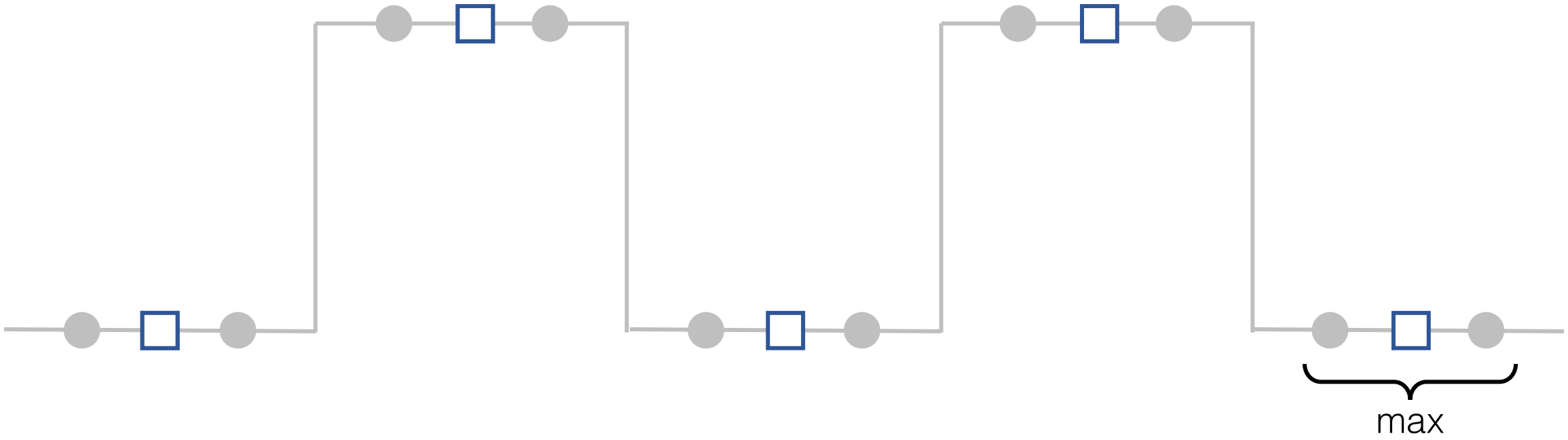
Re-examining Max-Pooling



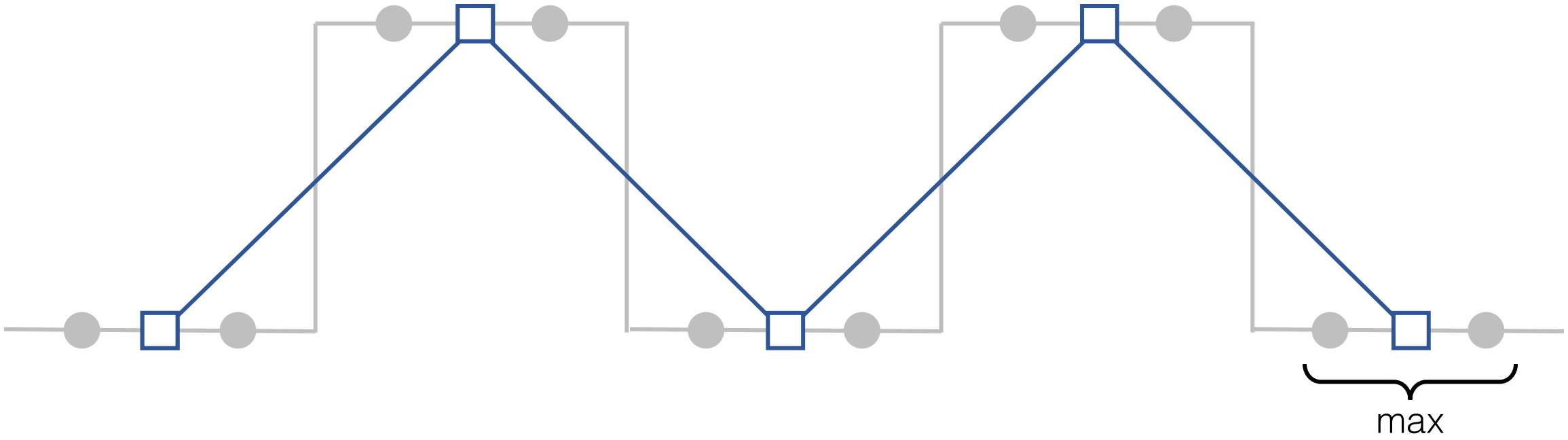
Re-examining Max-Pooling



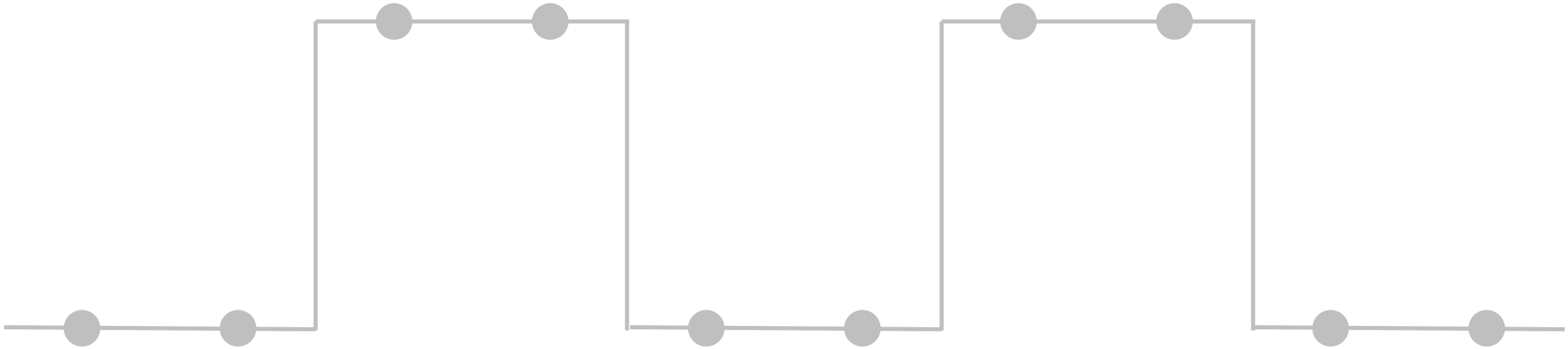
Re-examining Max-Pooling



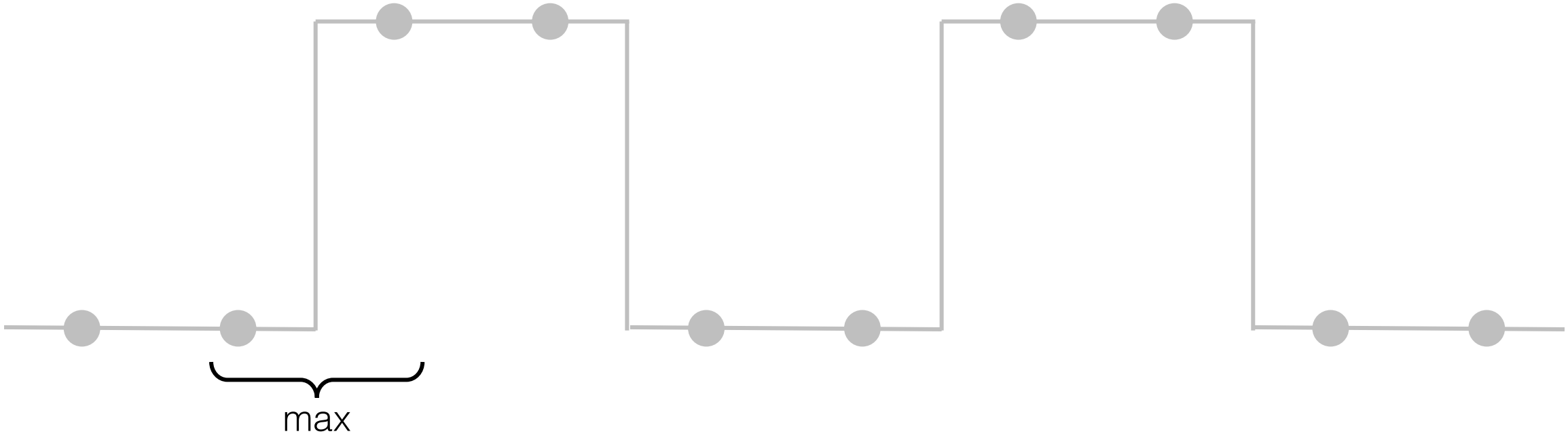
Re-examining Max-Pooling



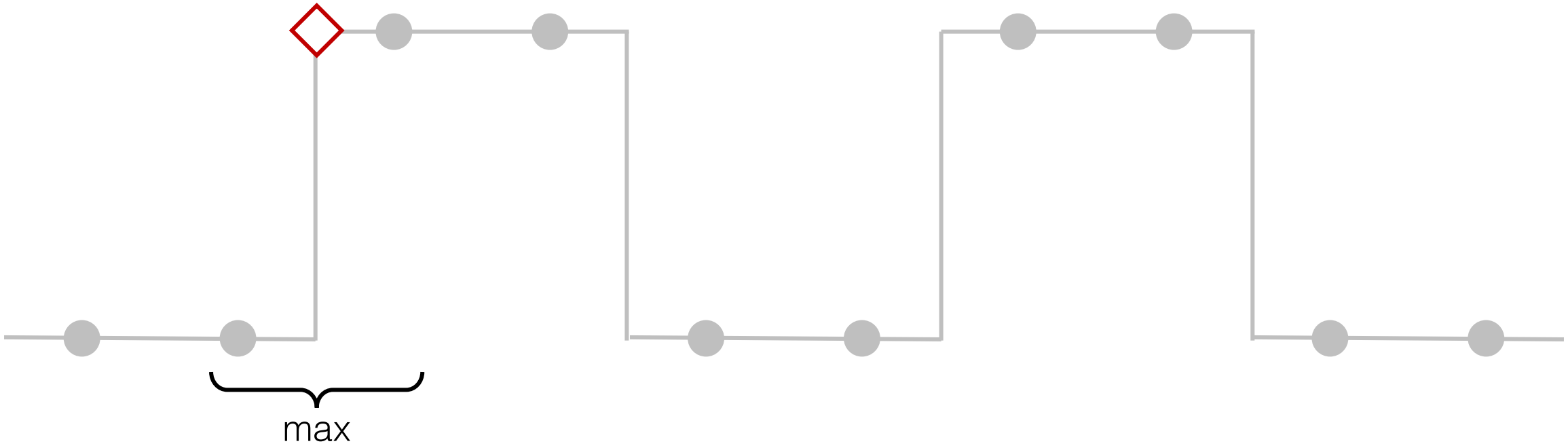
Re-examining Max-Pooling



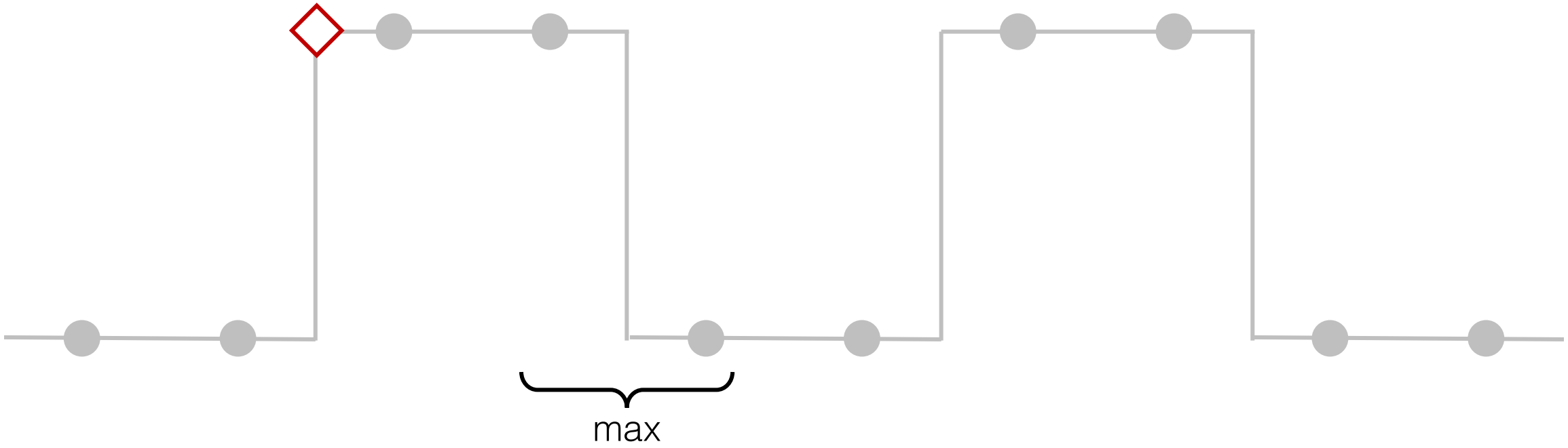
Re-examining Max-Pooling



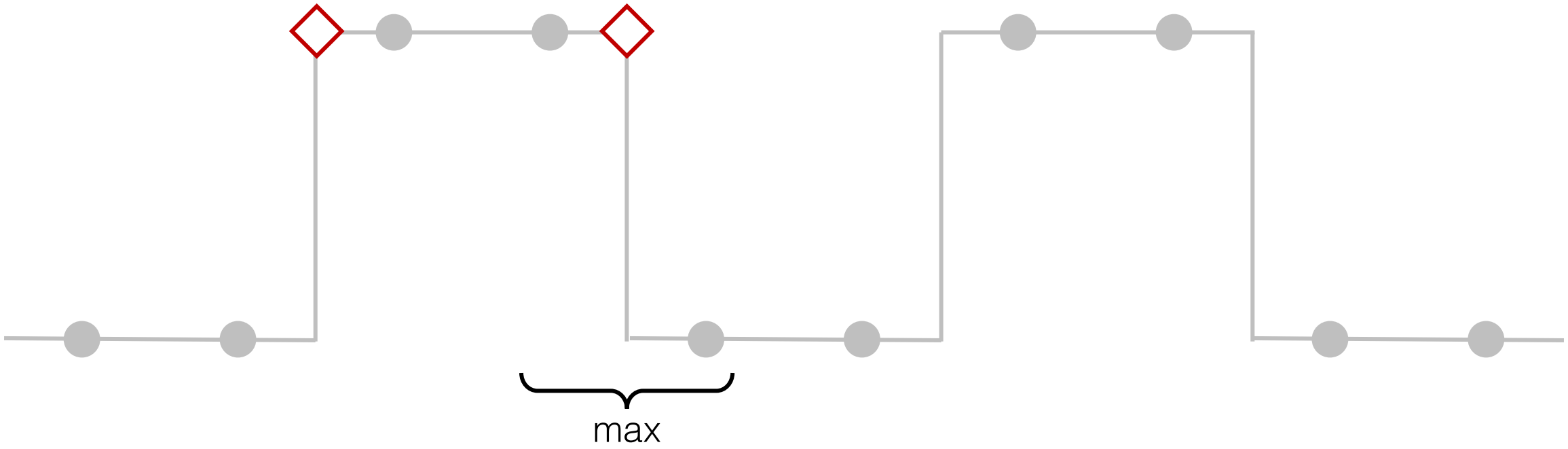
Re-examining Max-Pooling



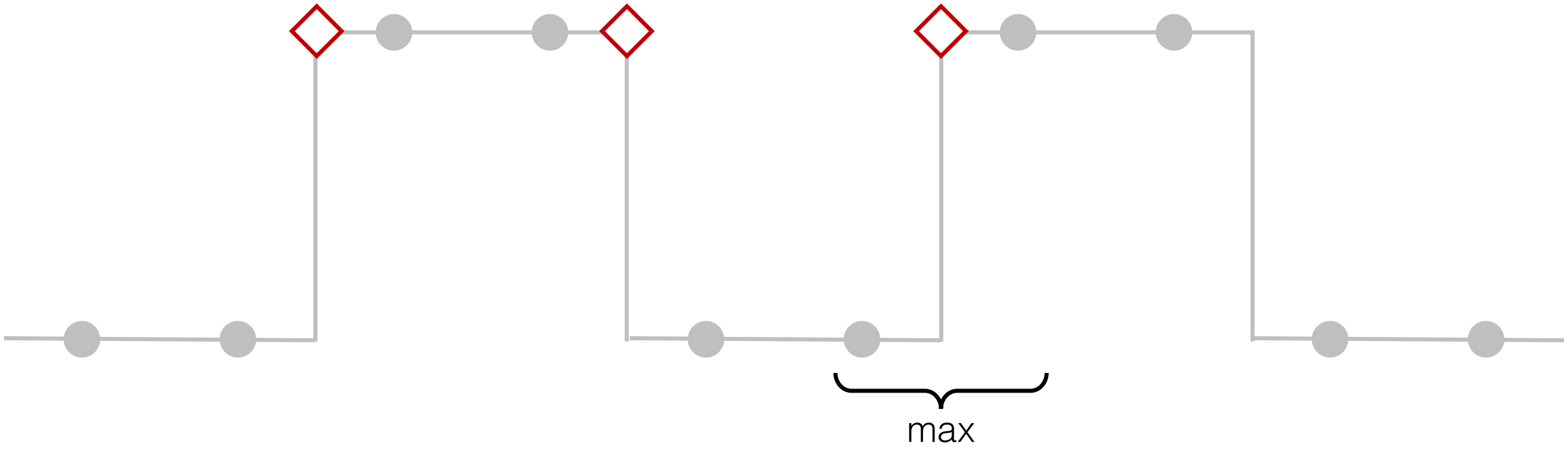
Re-examining Max-Pooling



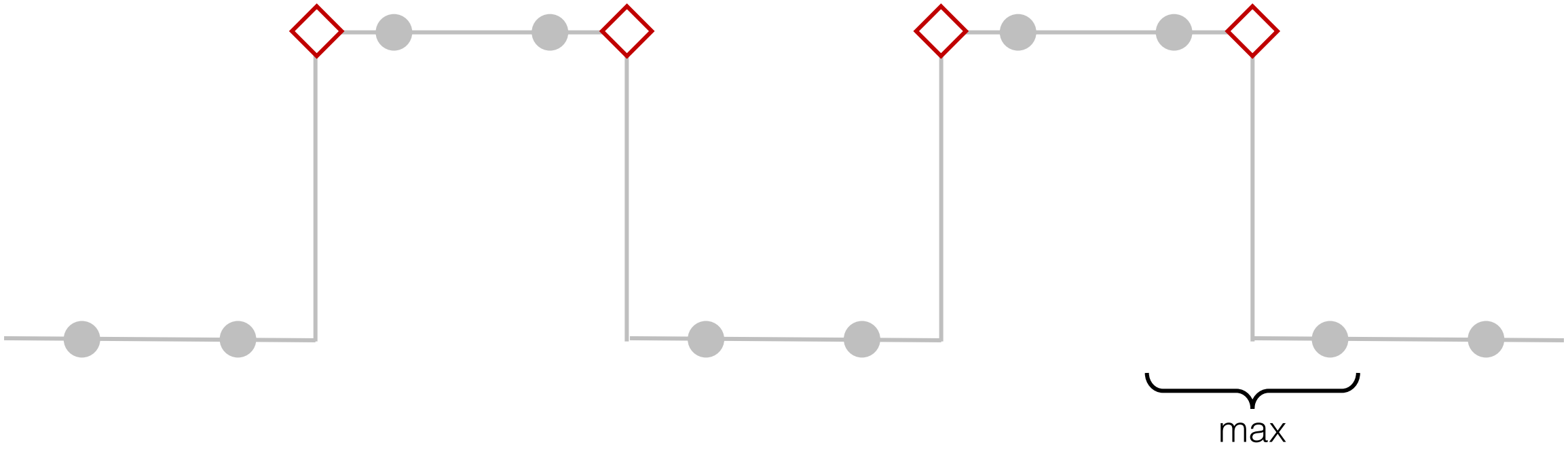
Re-examining Max-Pooling



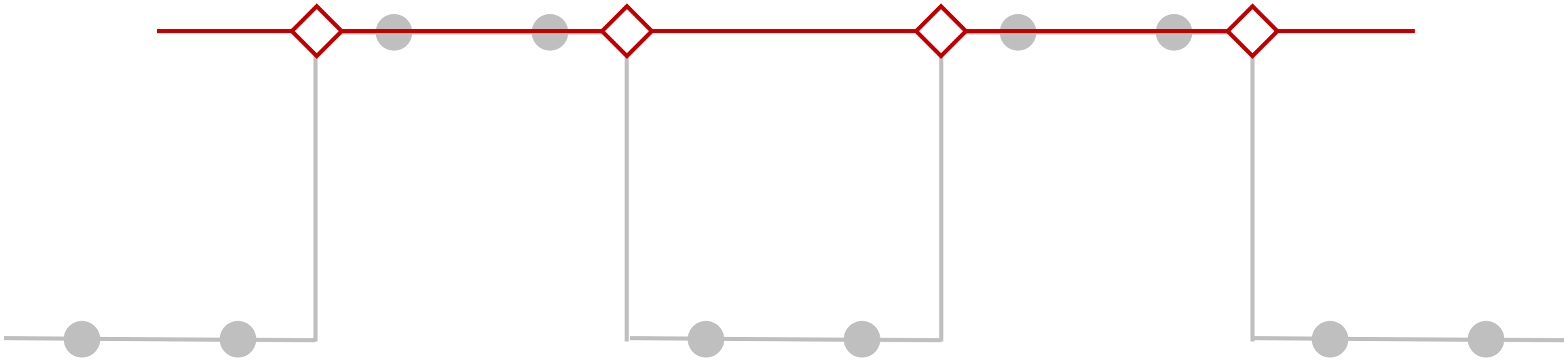
Re-examining Max-Pooling



Re-examining Max-Pooling

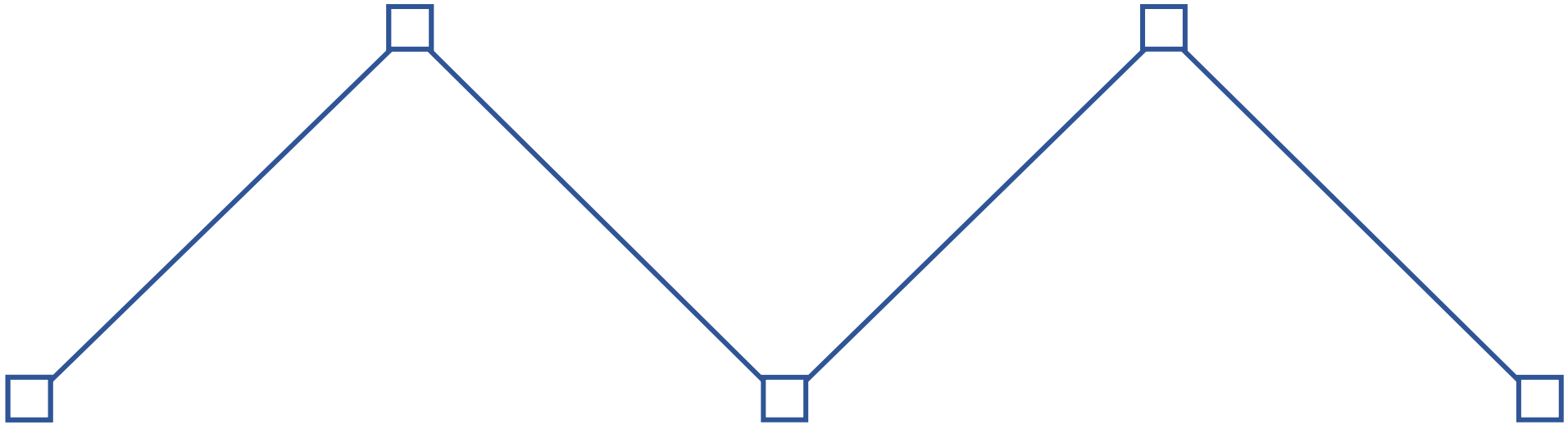


Re-examining Max-Pooling

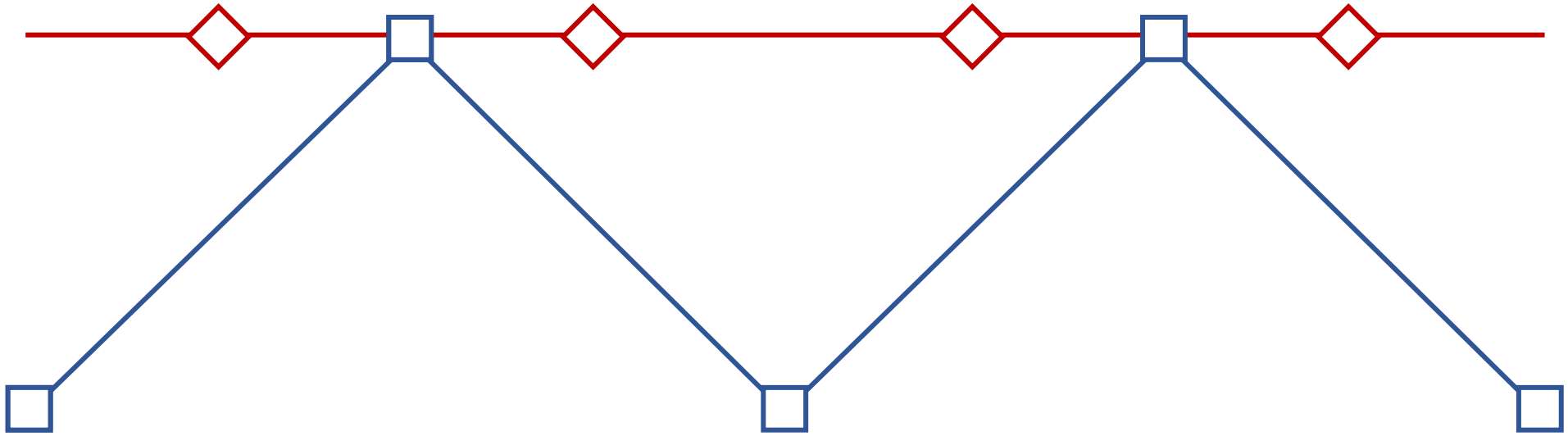


Re-examining Max-Pooling

Re-examining Max-Pooling



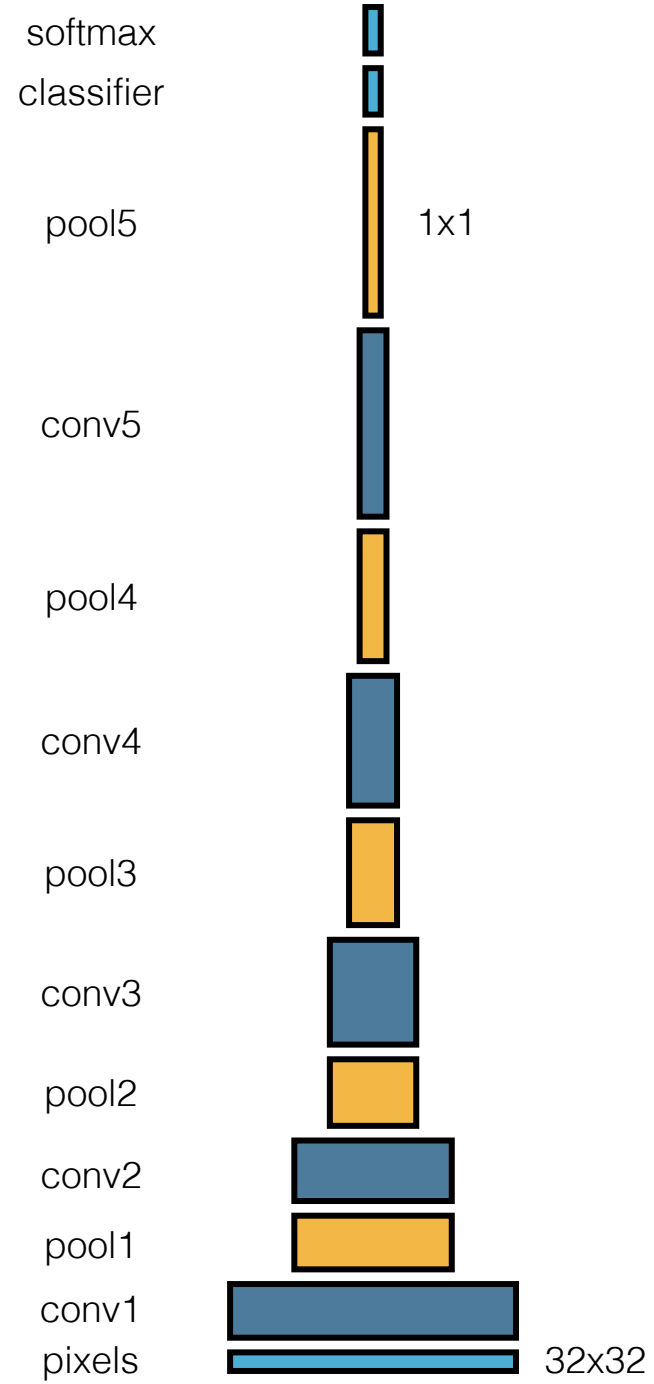
Re-examining Max-Pooling



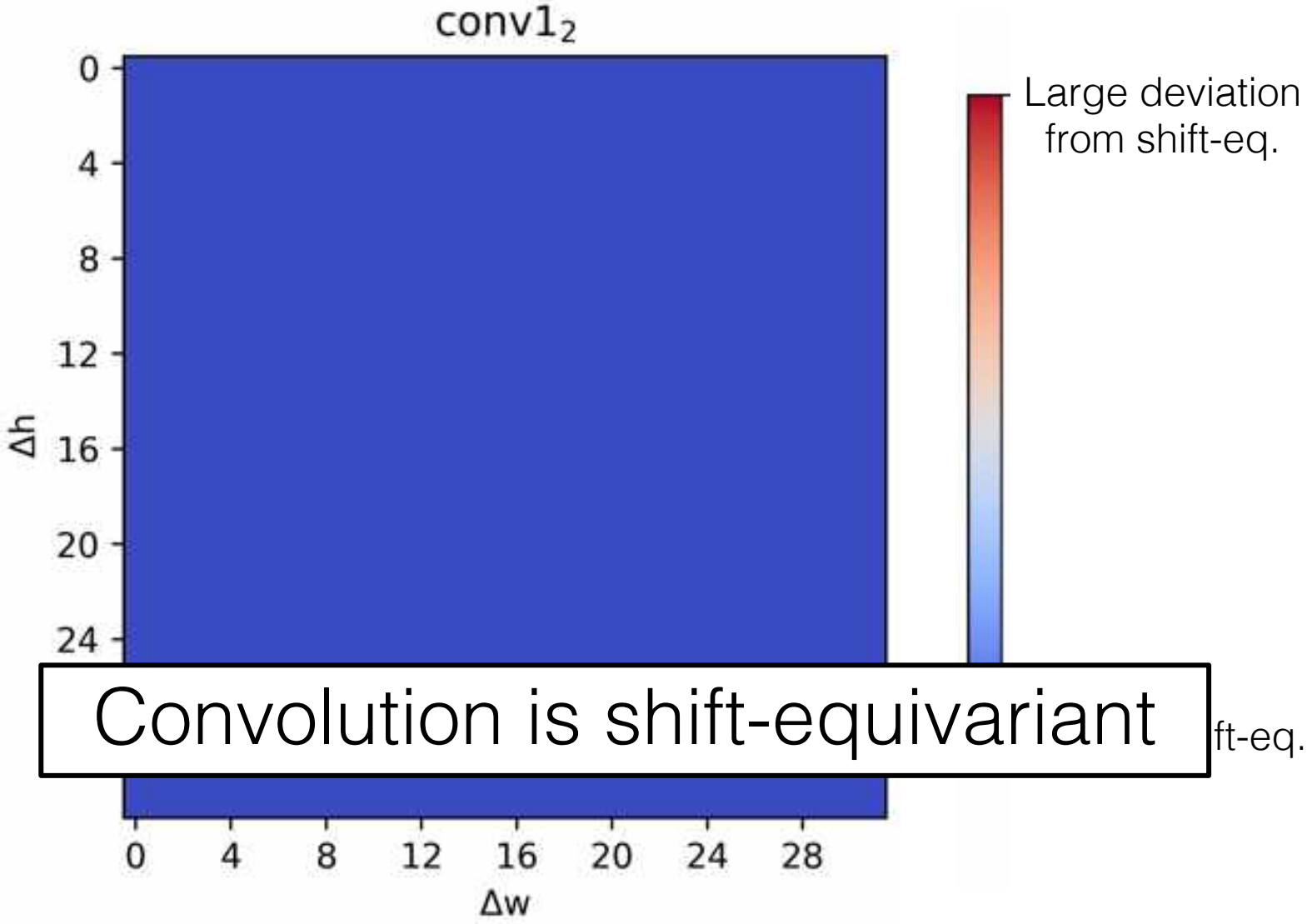
Max-pooling **breaks** shift-equivariance

Shift-equivariance Testbed

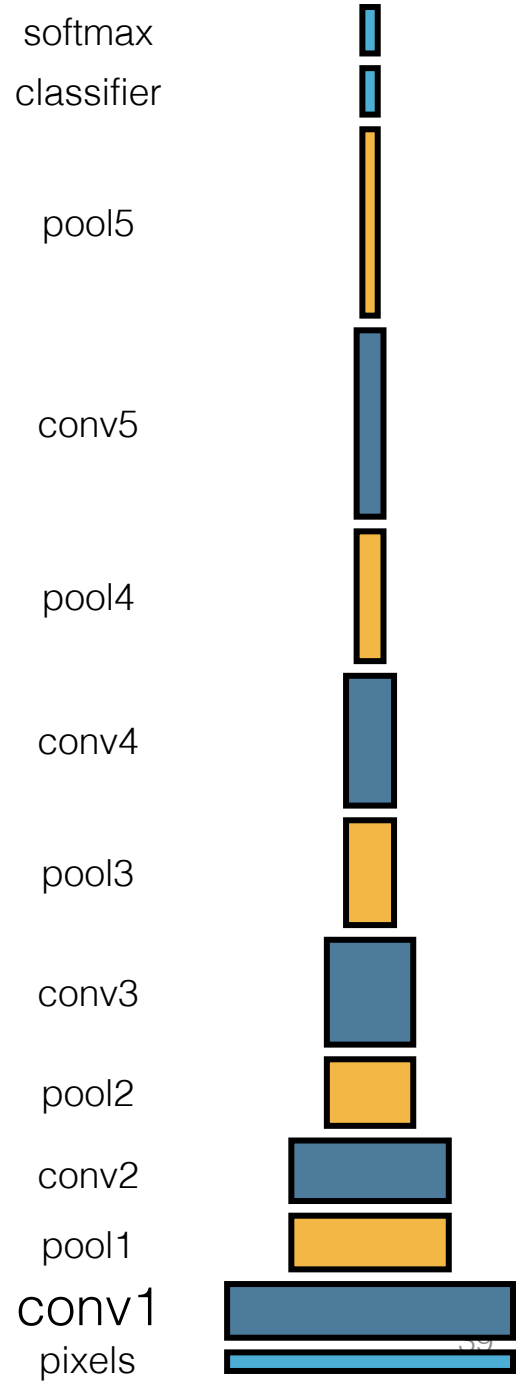
- CIFAR
- VGG network
  - 5 max-pools
- Test shift-equivariance condition
  - $F(\text{Shift}_{\Delta h,\Delta w}(X)) == \text{Shift}_{\Delta h,\Delta w}(F(X))$
- Circular convolution/shift



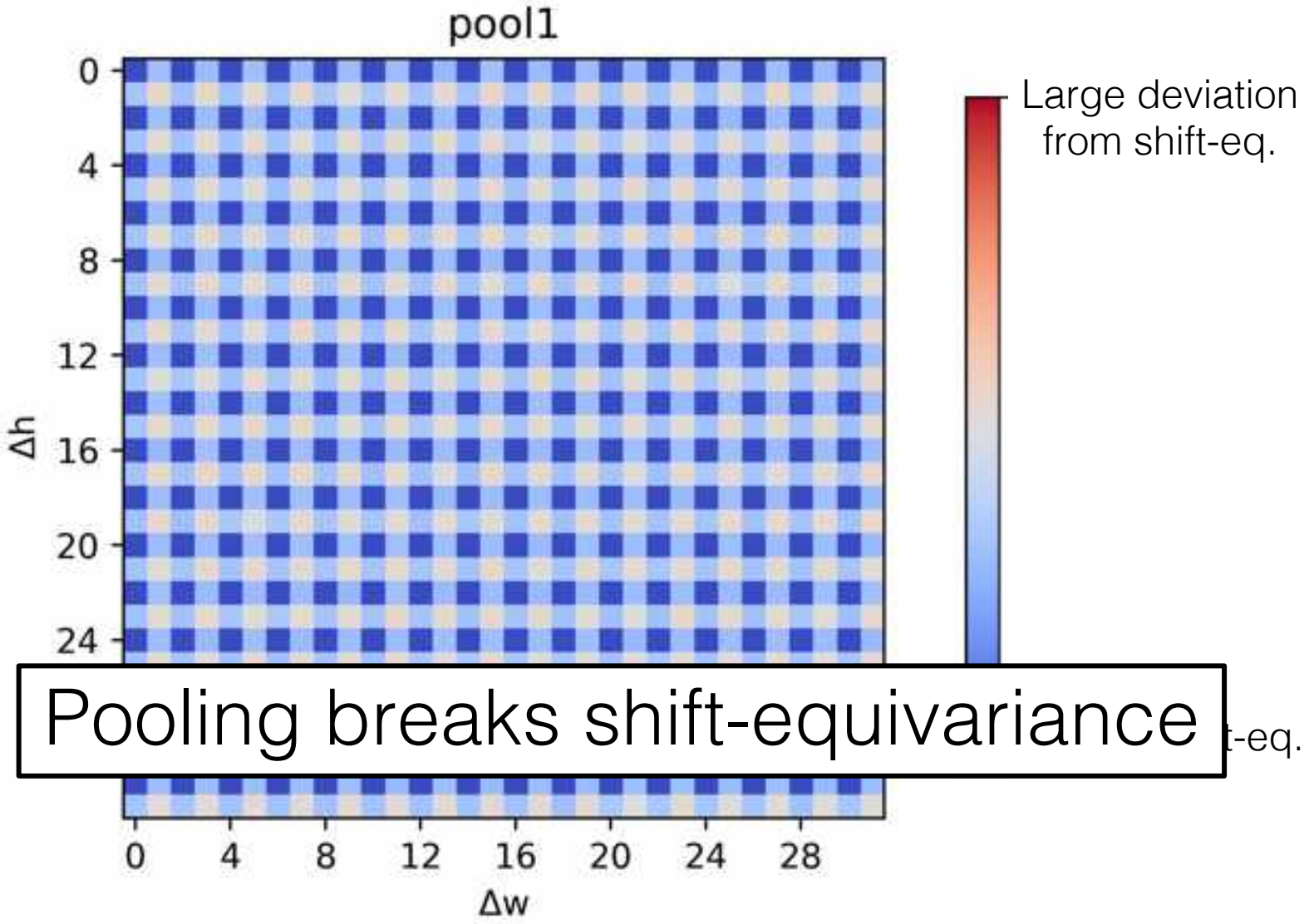
Shift-equivariance, per layer



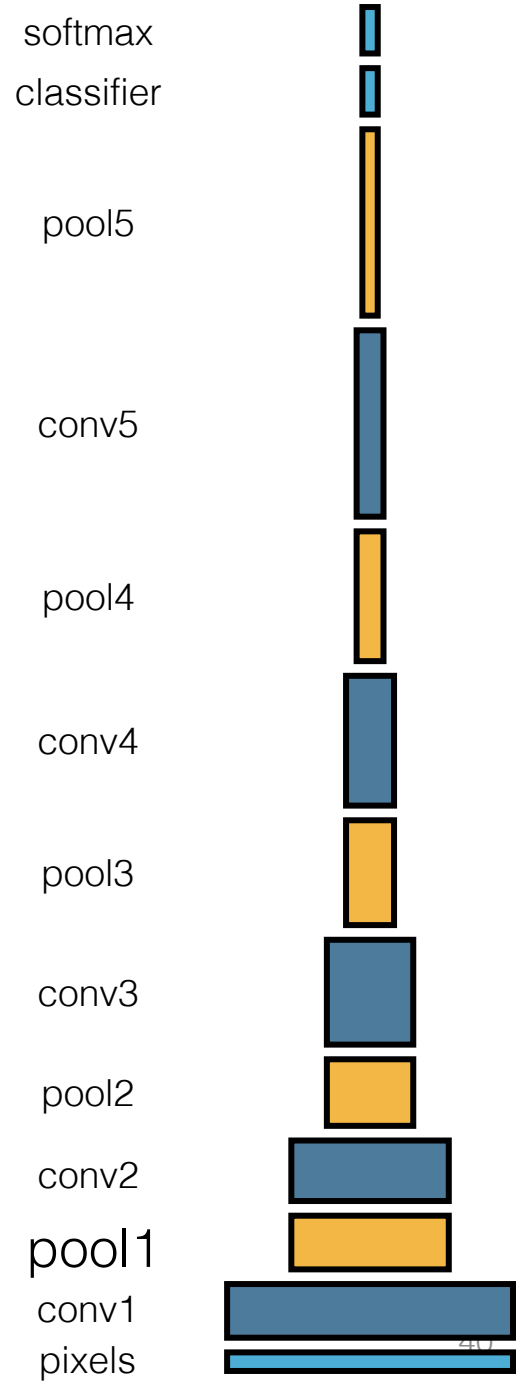
Convolution is shift-equivariant



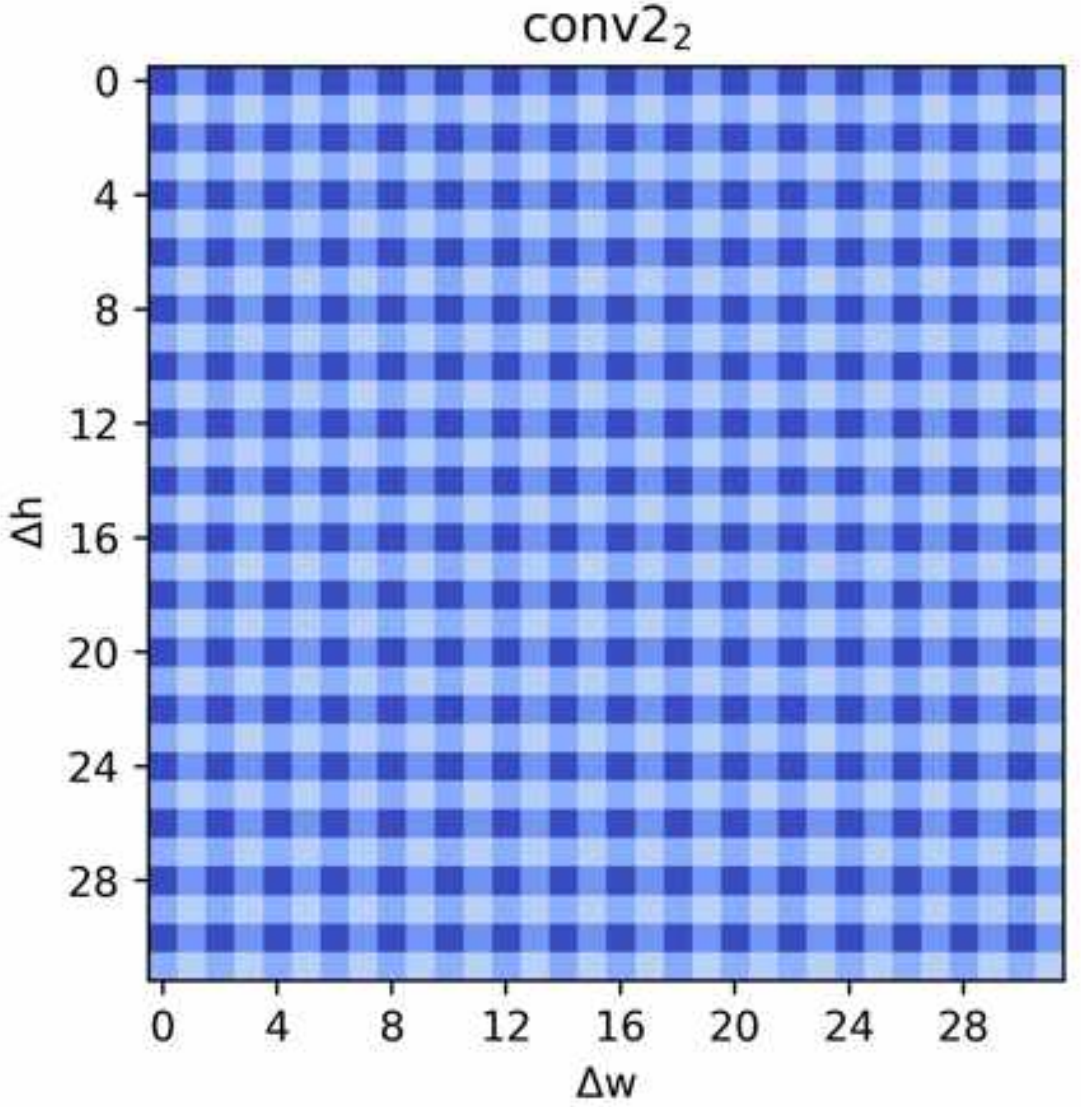
Shift-equivariance, per layer

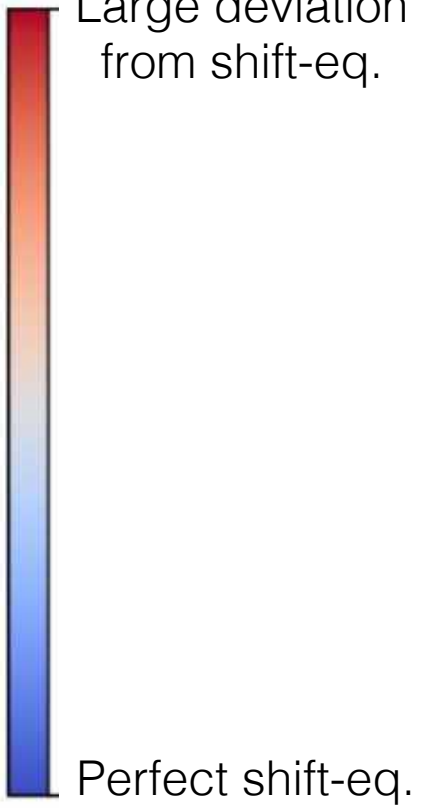


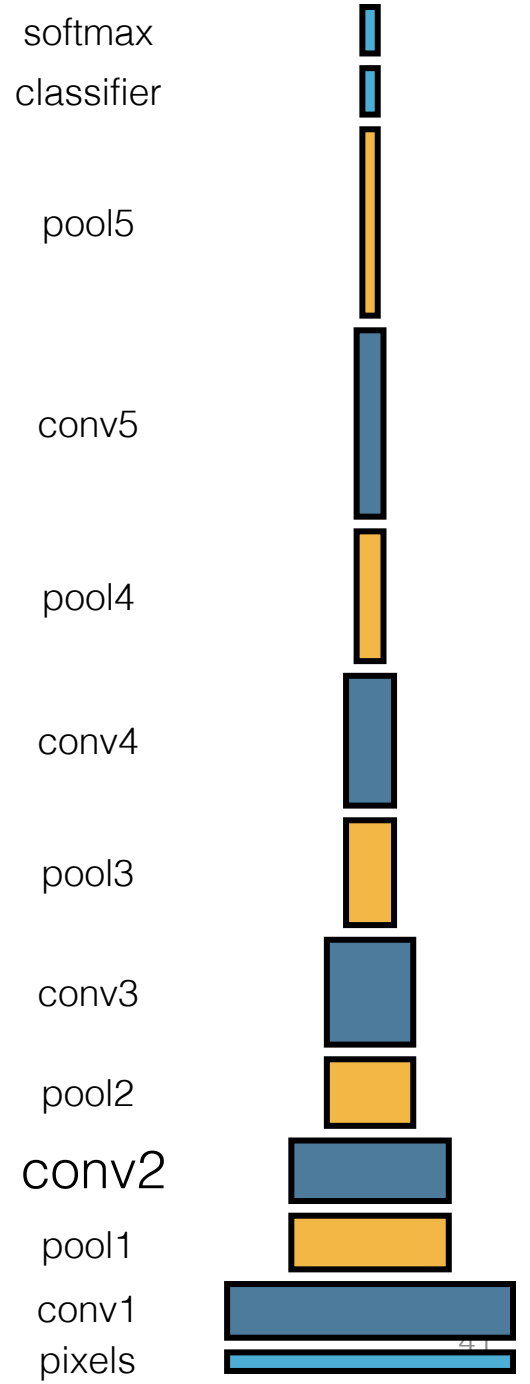
Pooling breaks shift-equivariance



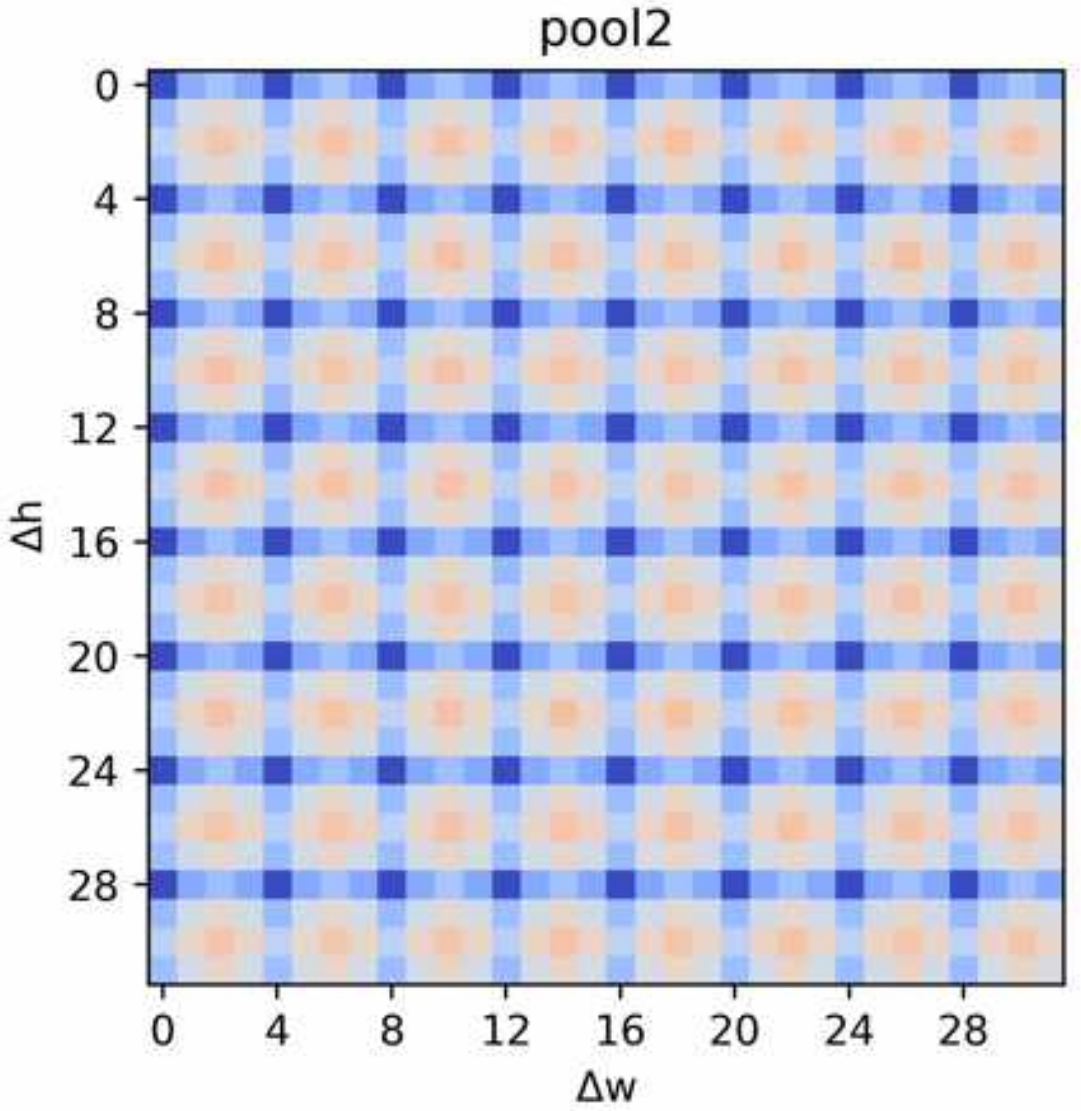
Shift-equivariance, per layer

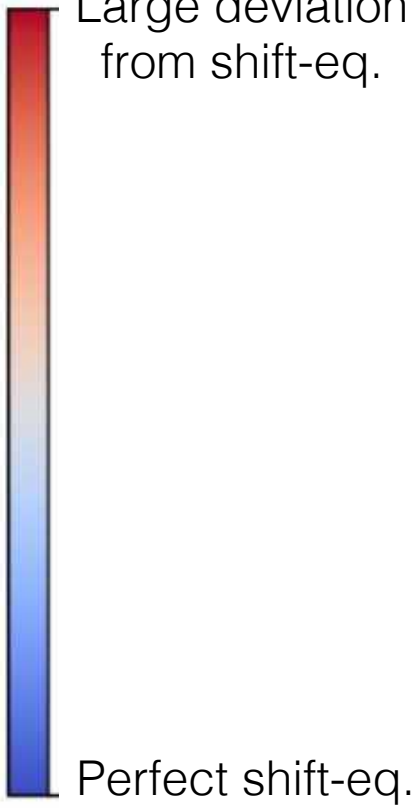


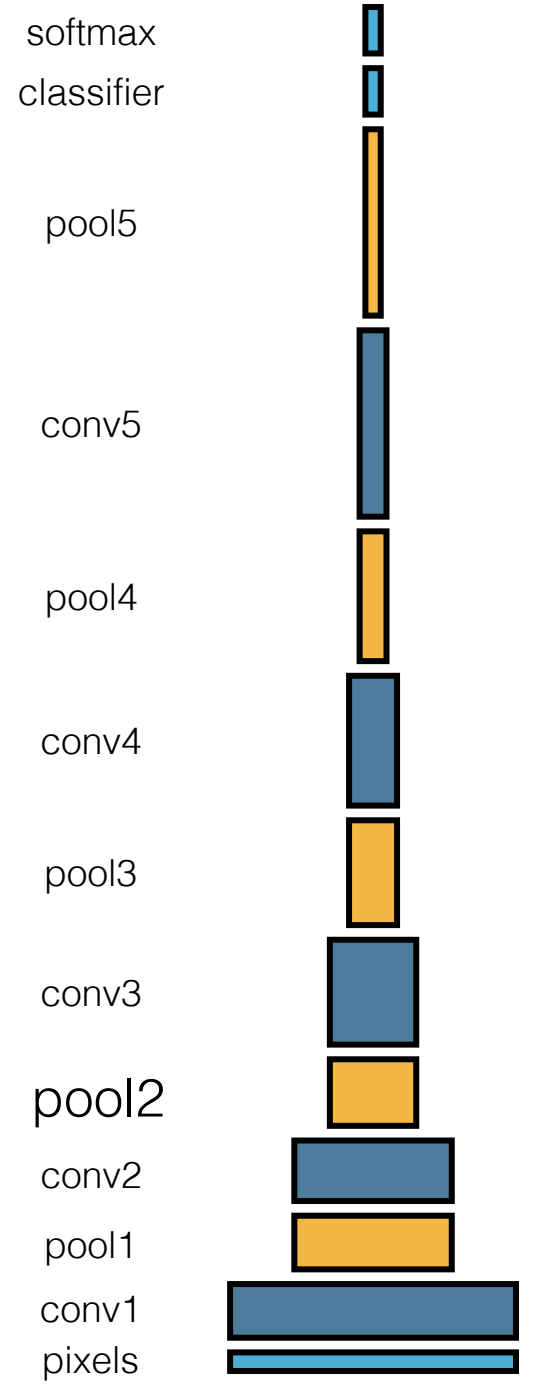




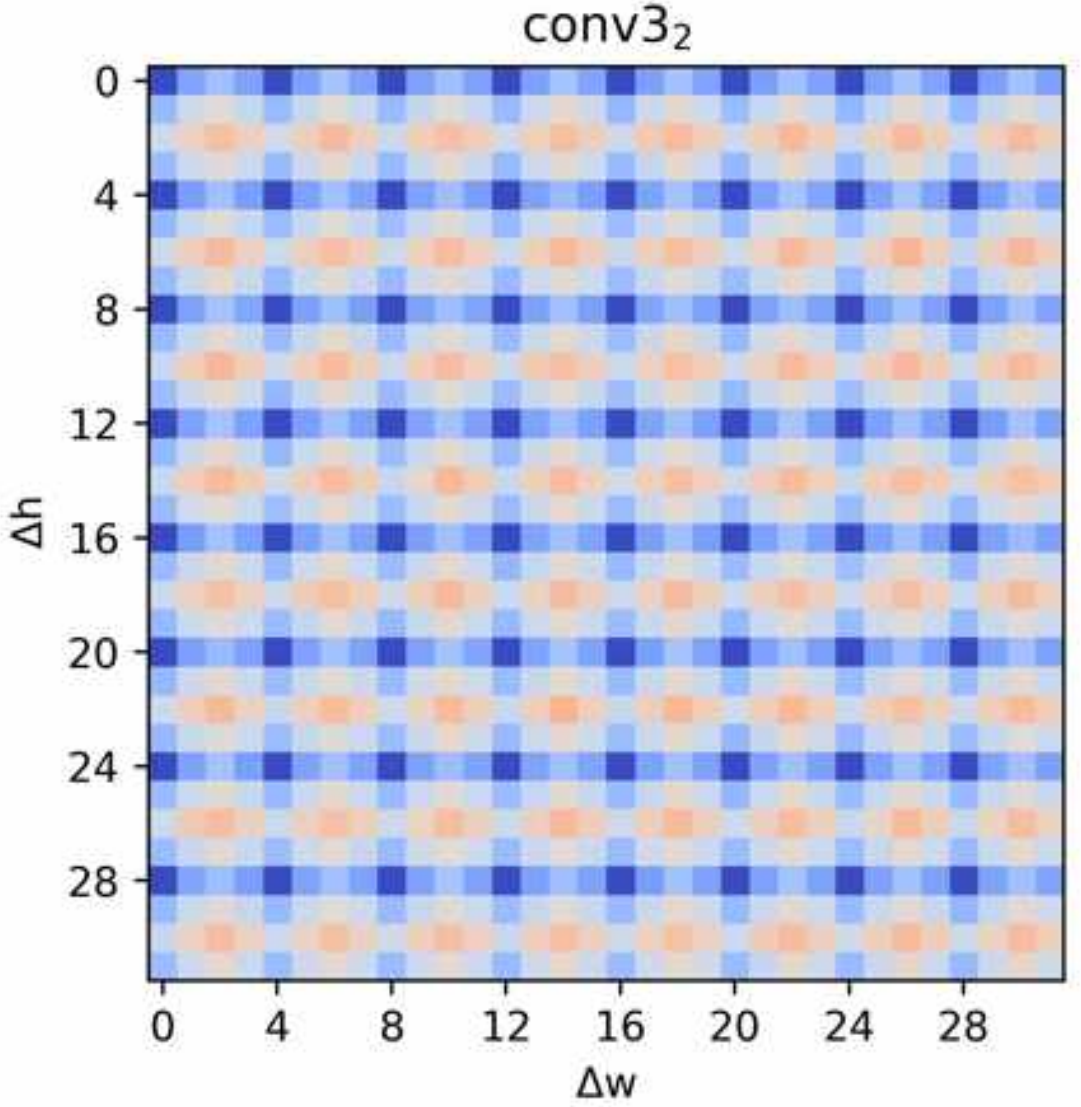
Shift-equivariance, per layer

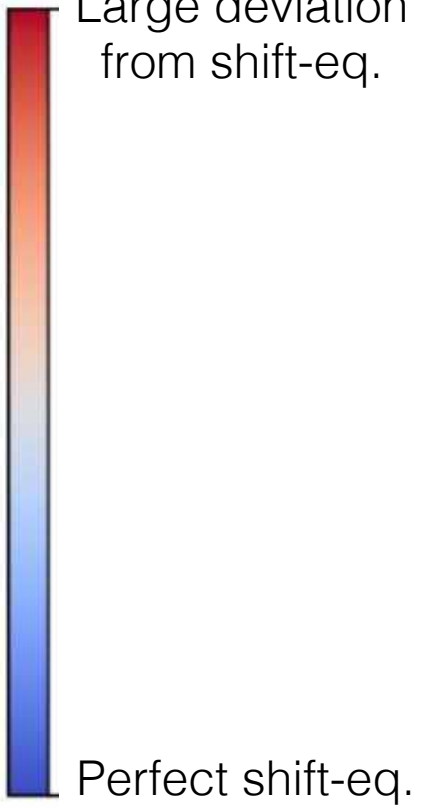


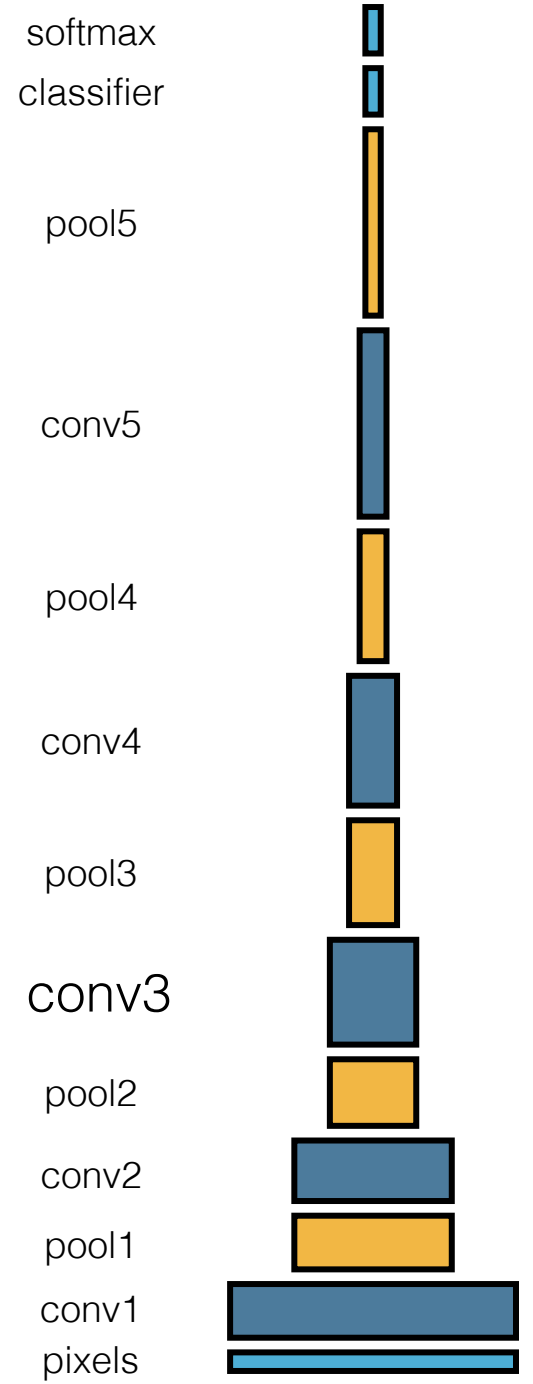




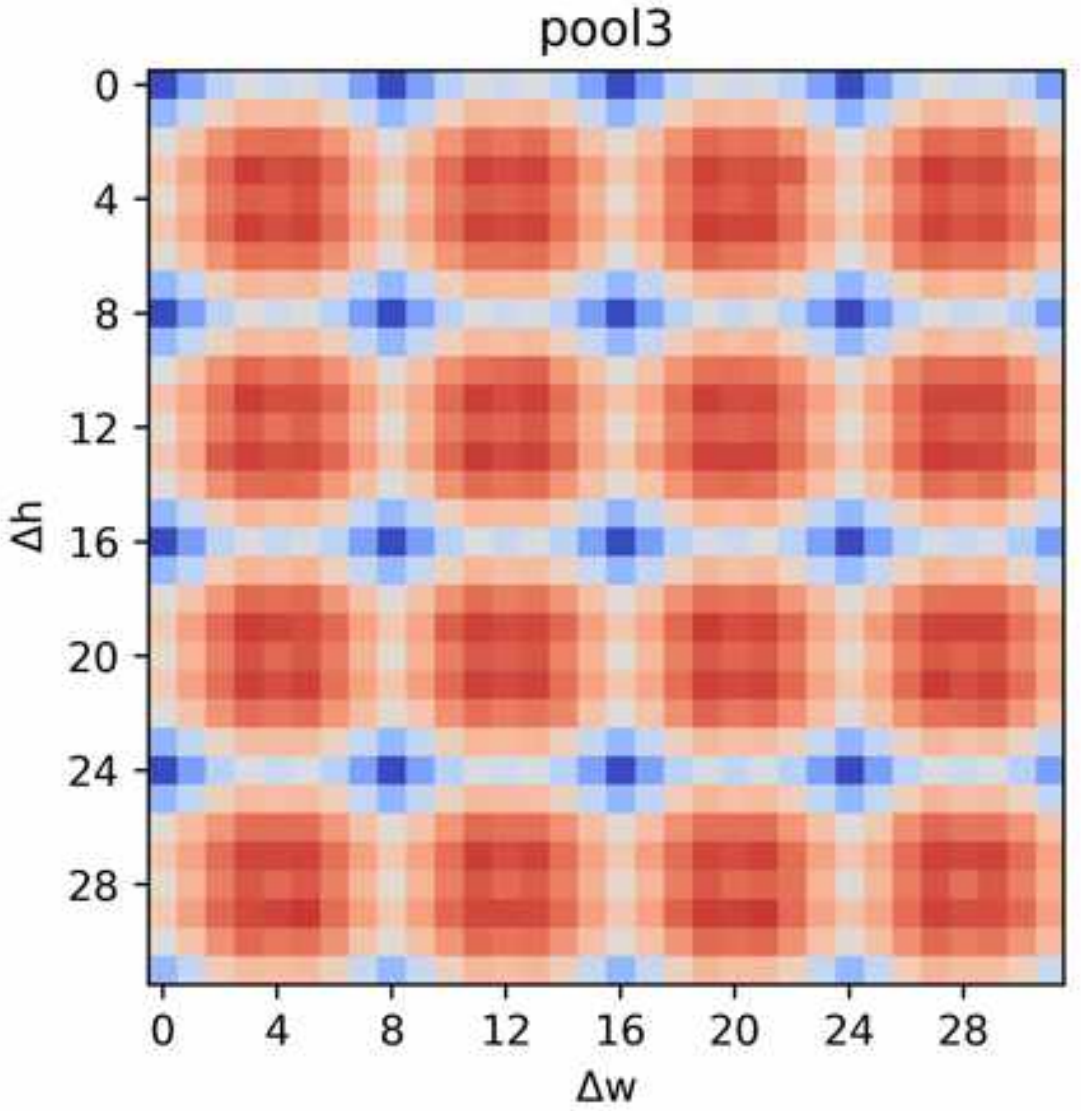
Shift-equivariance, per layer





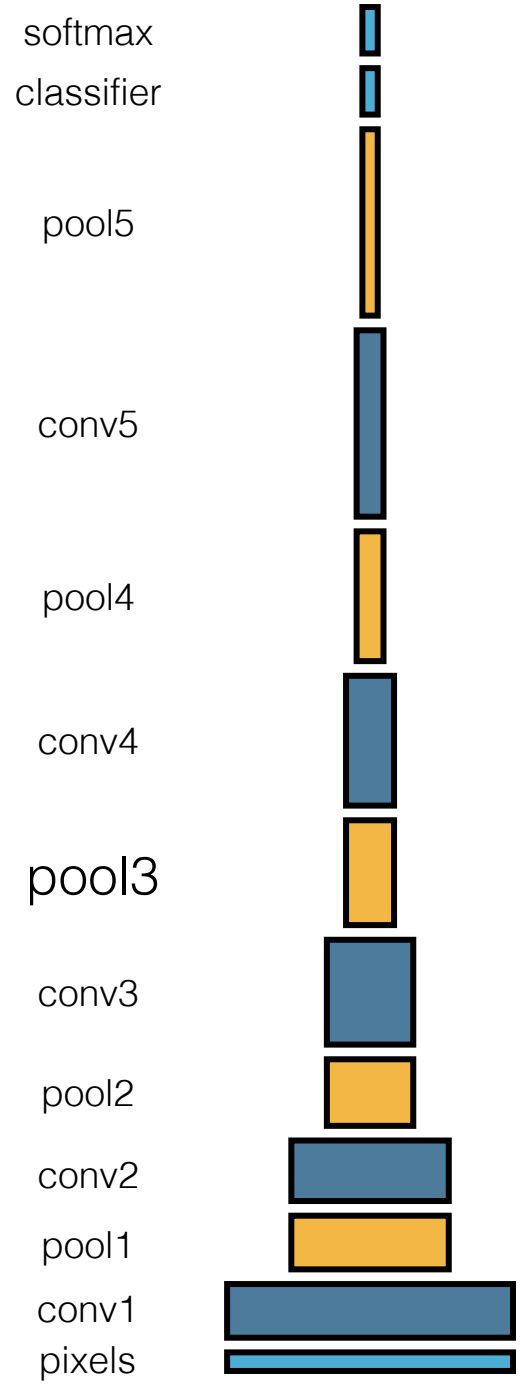


Shift-equivariance, per layer

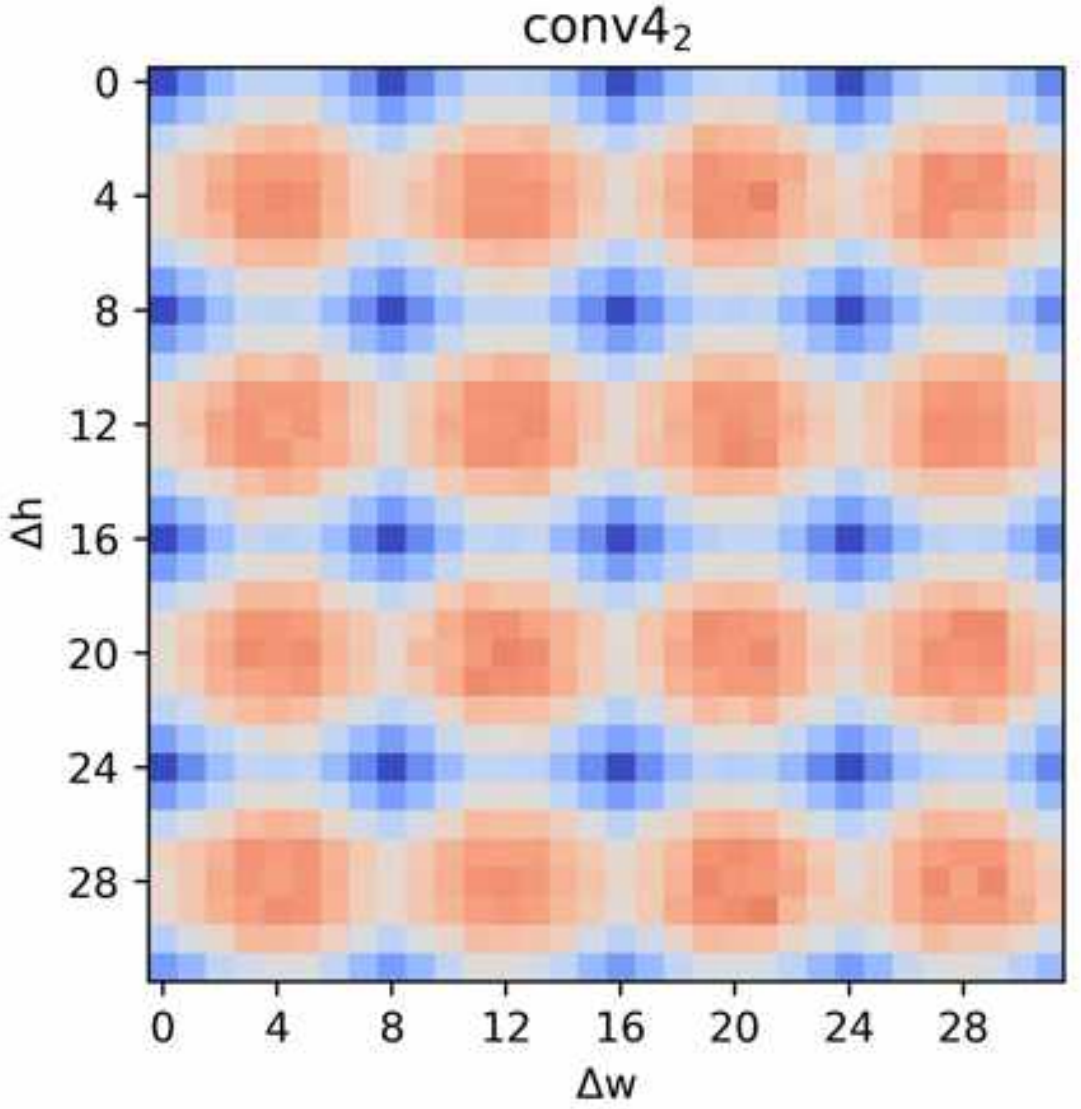


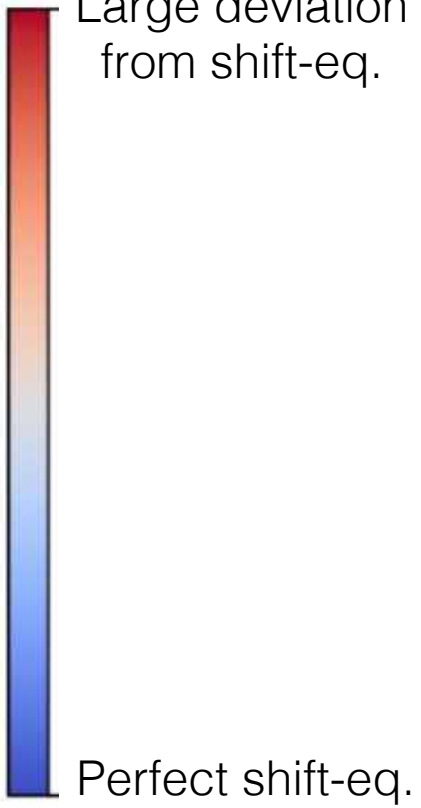
Large deviation from shift-eq.

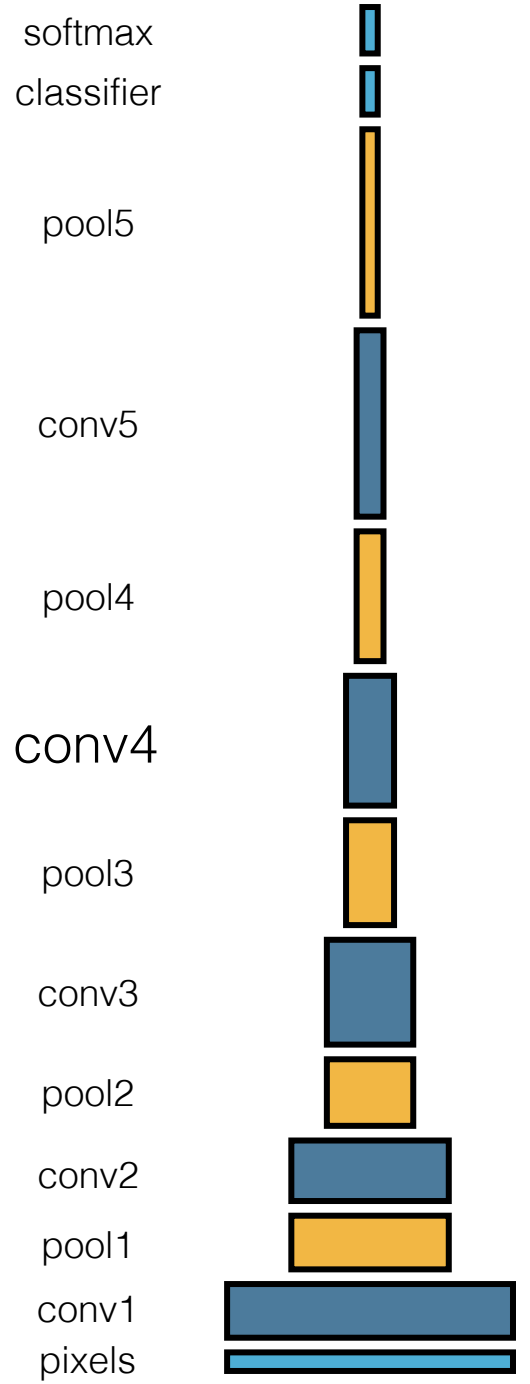
Perfect shift-eq.



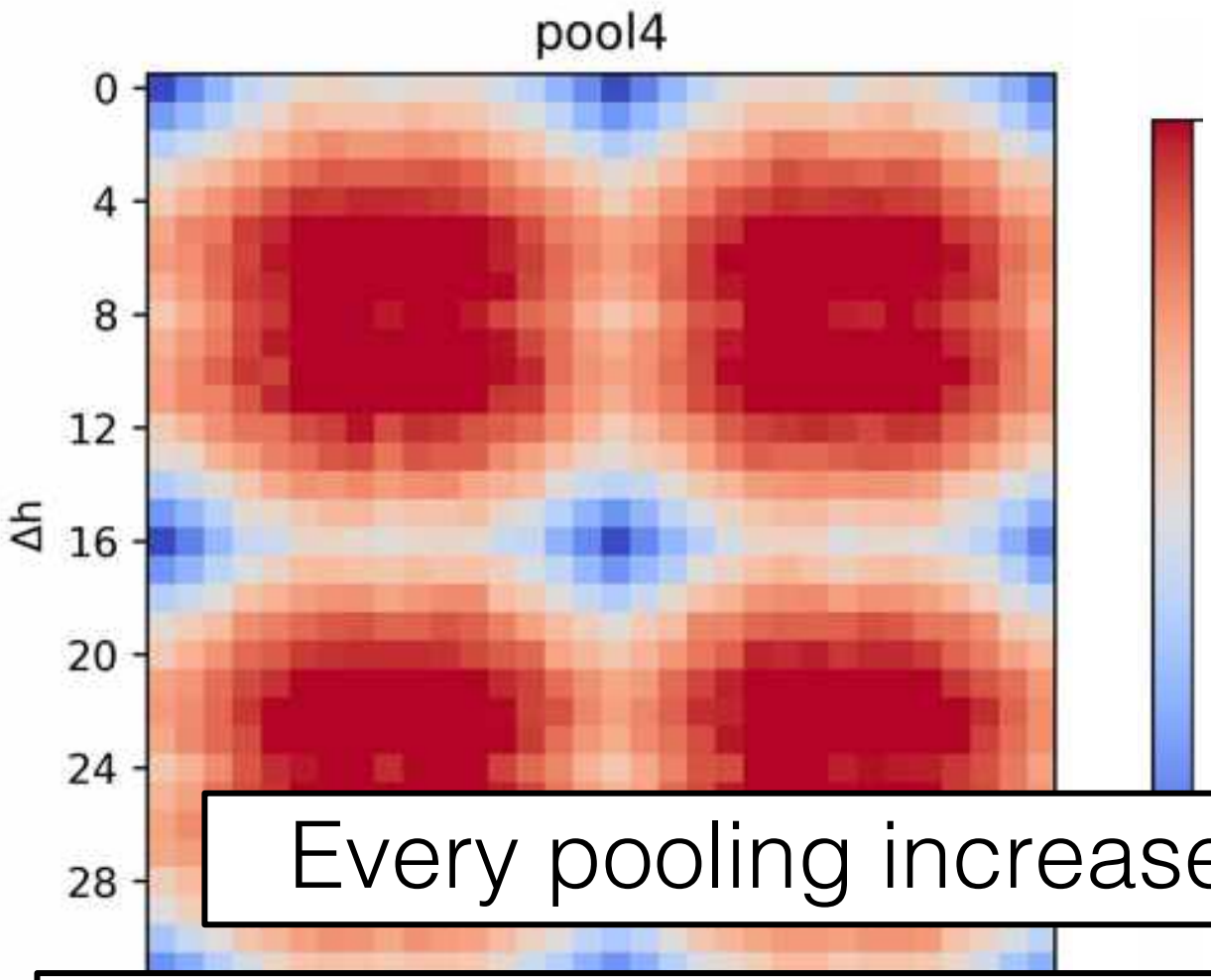
Shift-equivariance, per layer



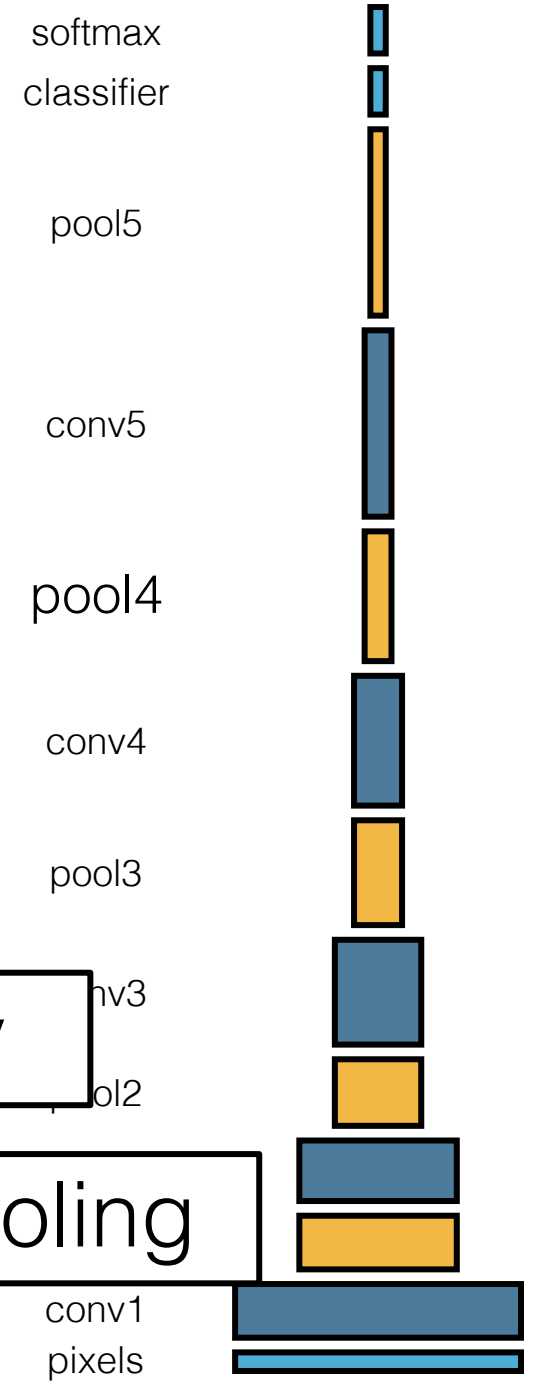




Shift-equivariance, per layer

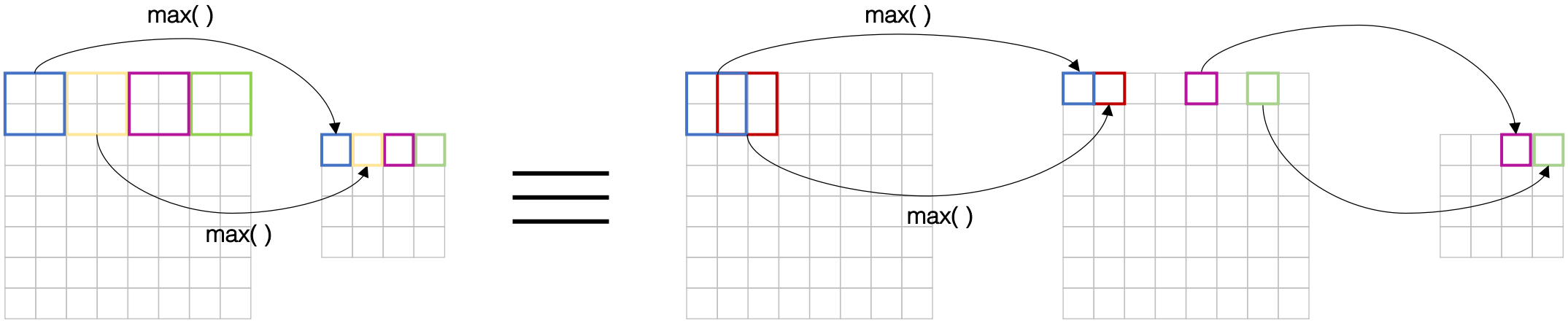


Large deviation from shift-eq.



Every pooling increases periodicity

Goal: Reconcile antialiasing with max-pooling



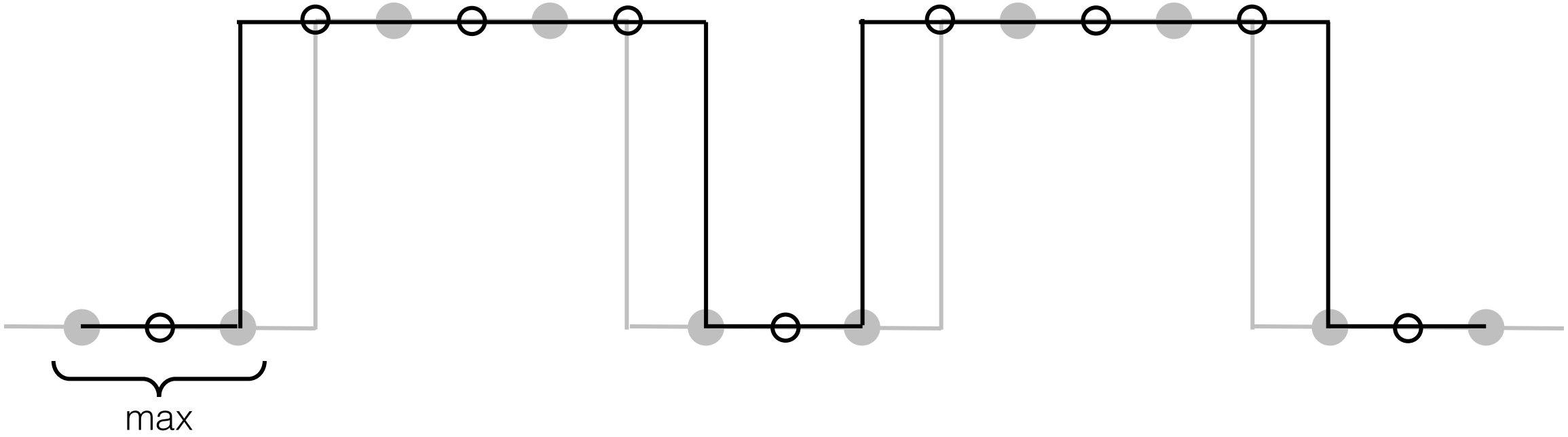
Strided-MaxPool

Equivalent Interpretation

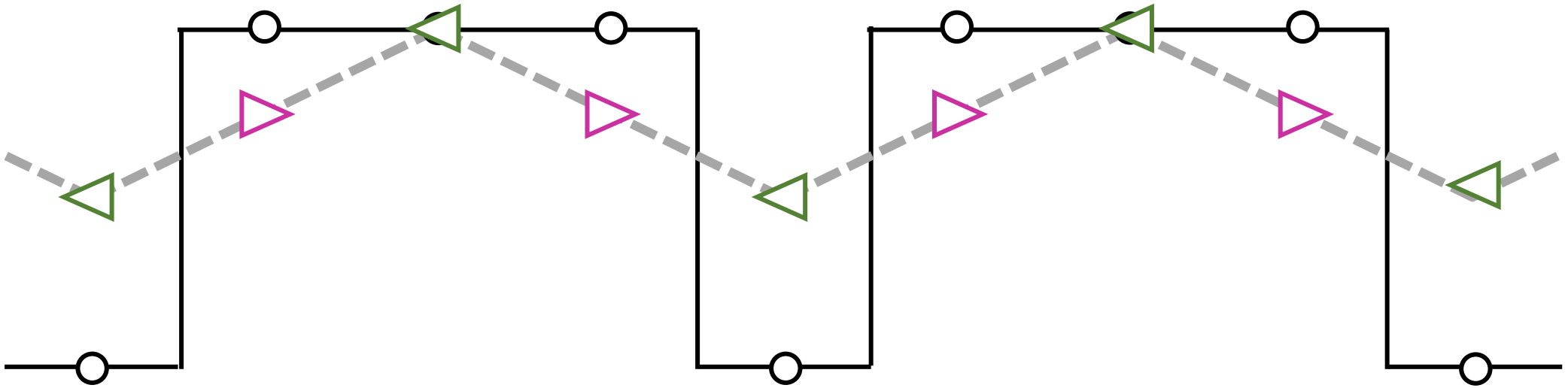
\+

Anti-aliased pooling (MaxBlurPool)

(1) Max (densely)

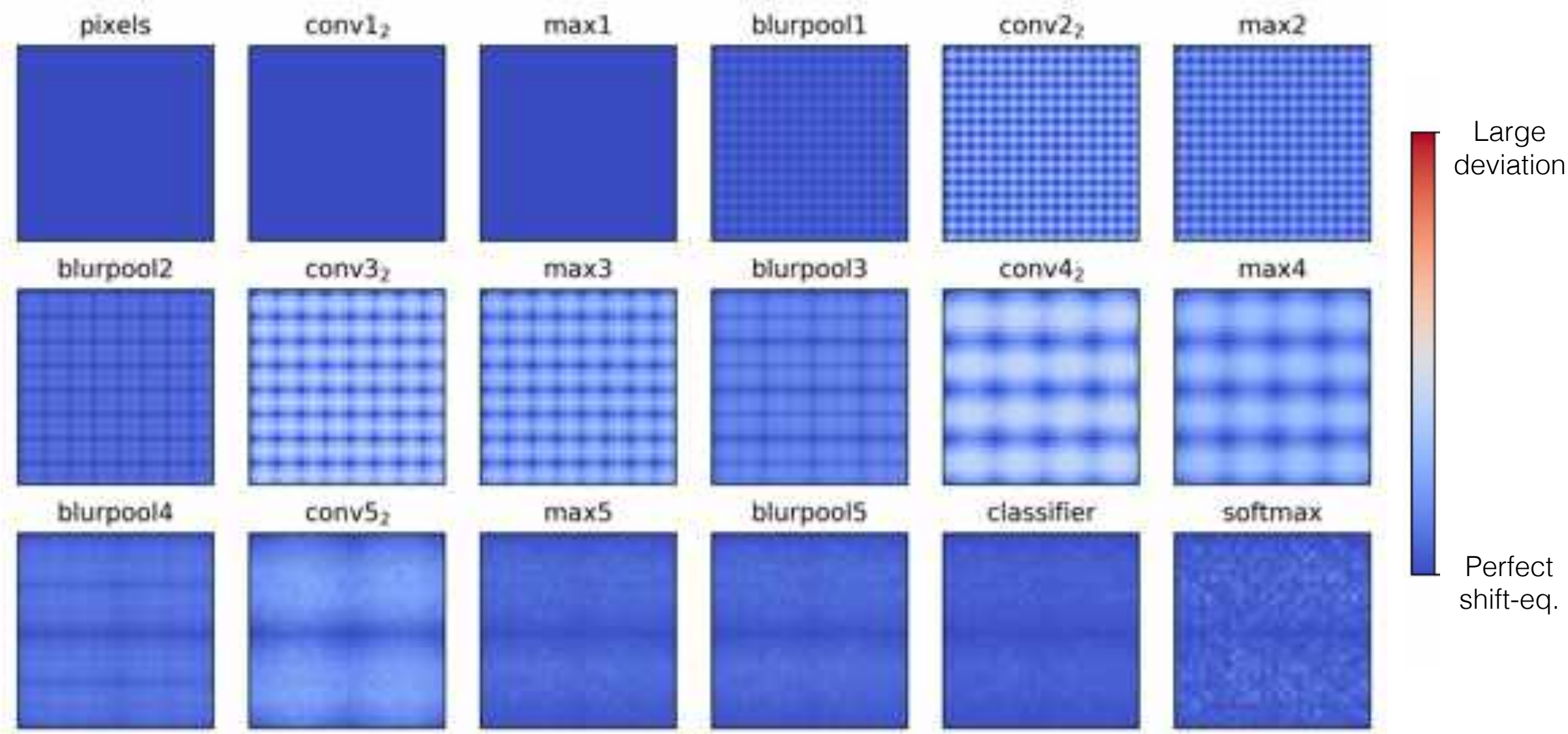


(2) Blur+Subsample

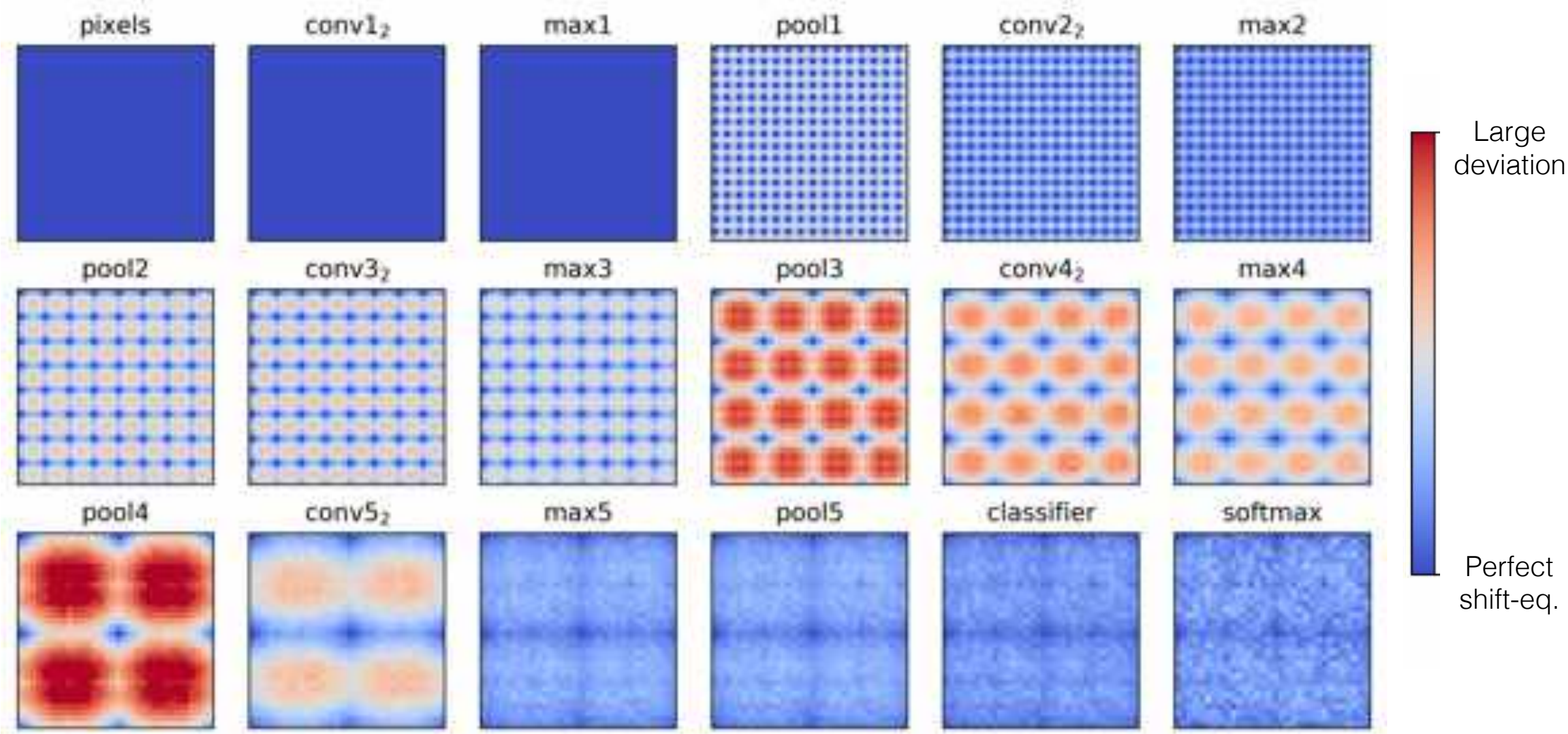


Shift-equivariance better preserved

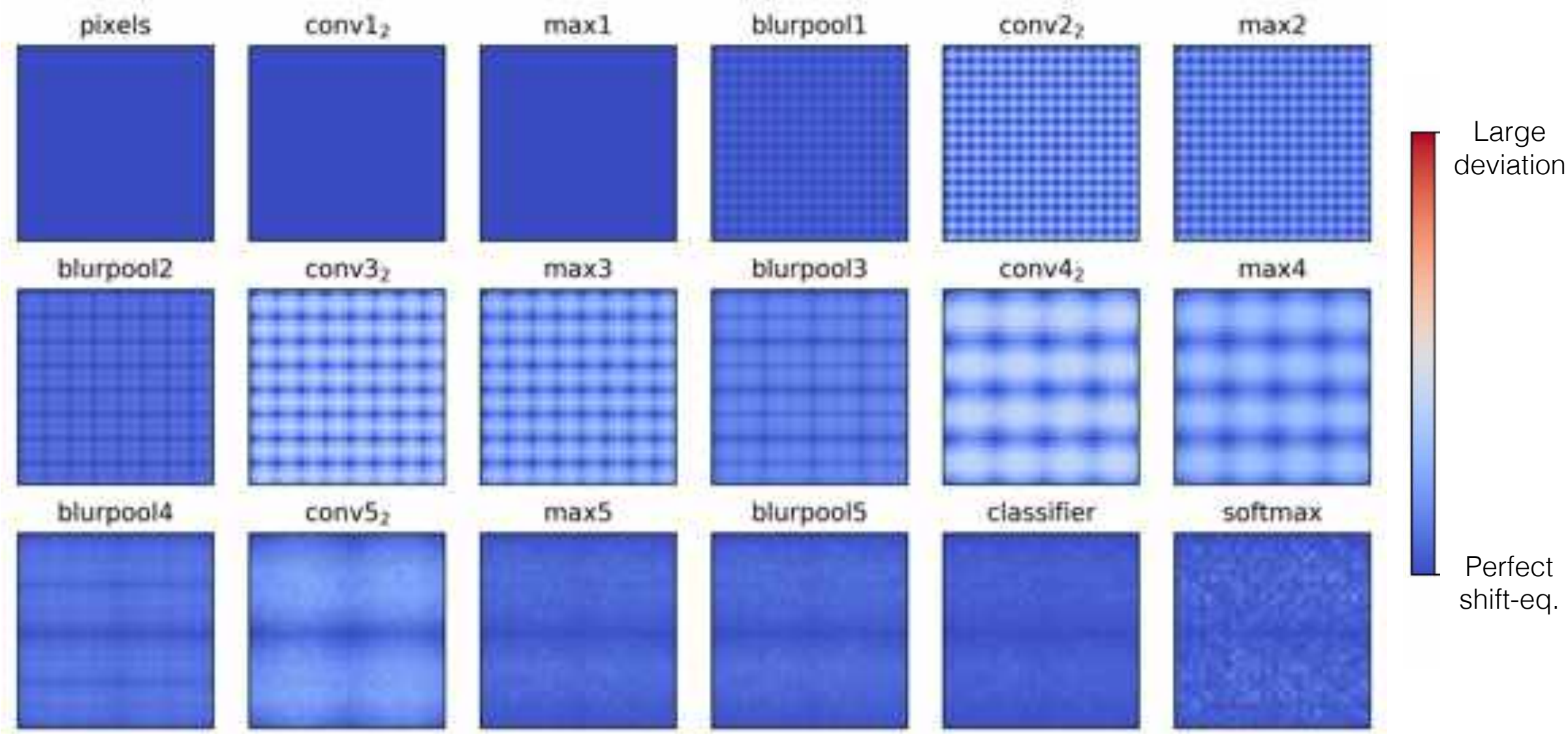
Anti-aliased



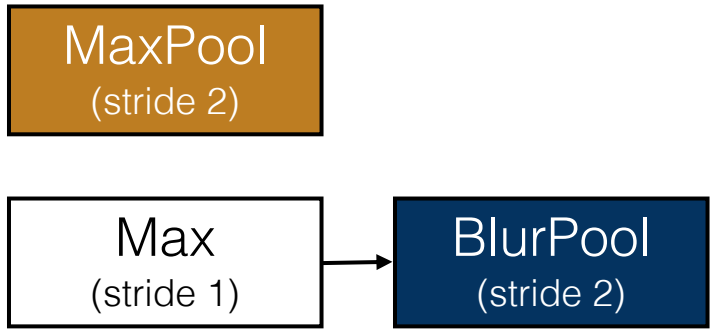
Baseline


pixels
conv1$_2$
max1
pool1
conv2$_2$
max2
pool2
conv3$_2$
max3
pool3
conv4$_2$
max4
pool4
conv5$_2$
max5
pool5
classifier
softmax
Large deviation
Perfect shift-eq.

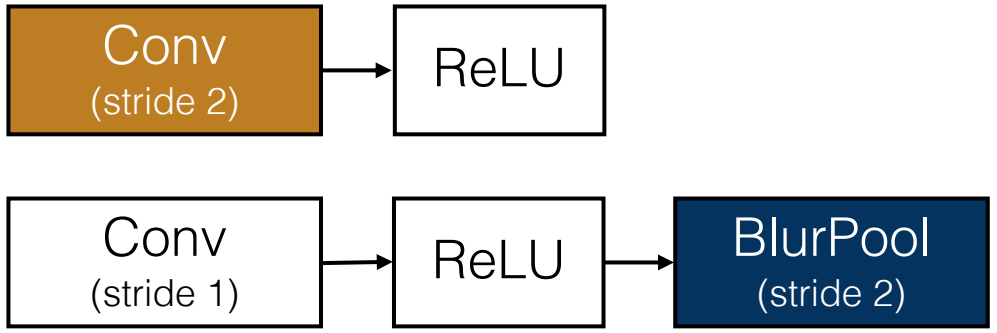
Anti-aliased



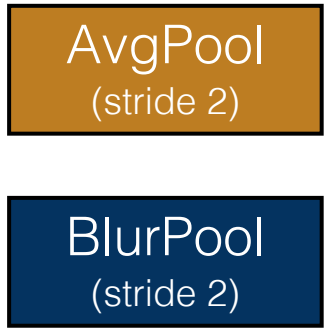
Other downsampling layers



Max Pooling
(VGG, AlexNet)



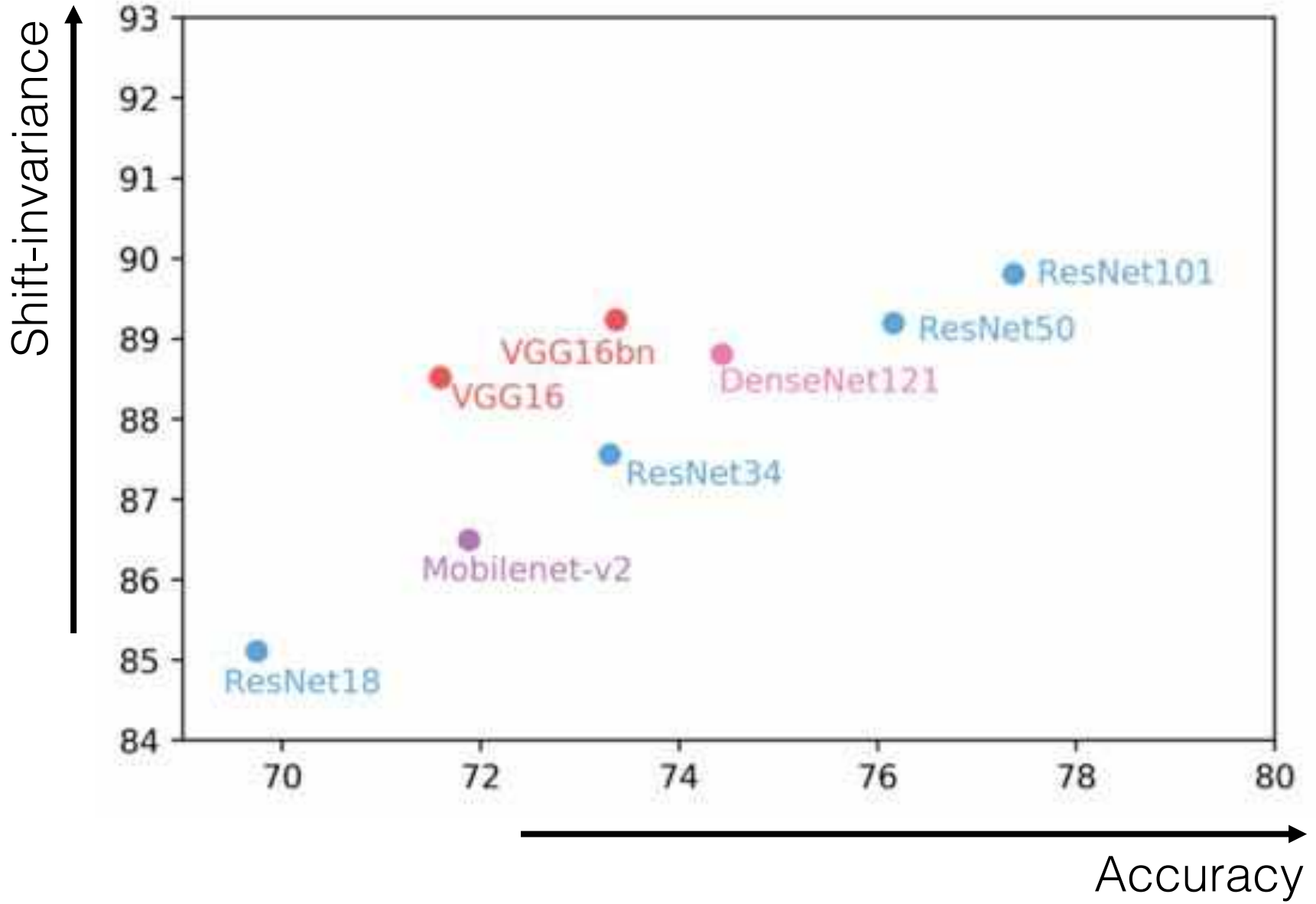
Strided-Convolution
(ResNet, MobileNetv2)



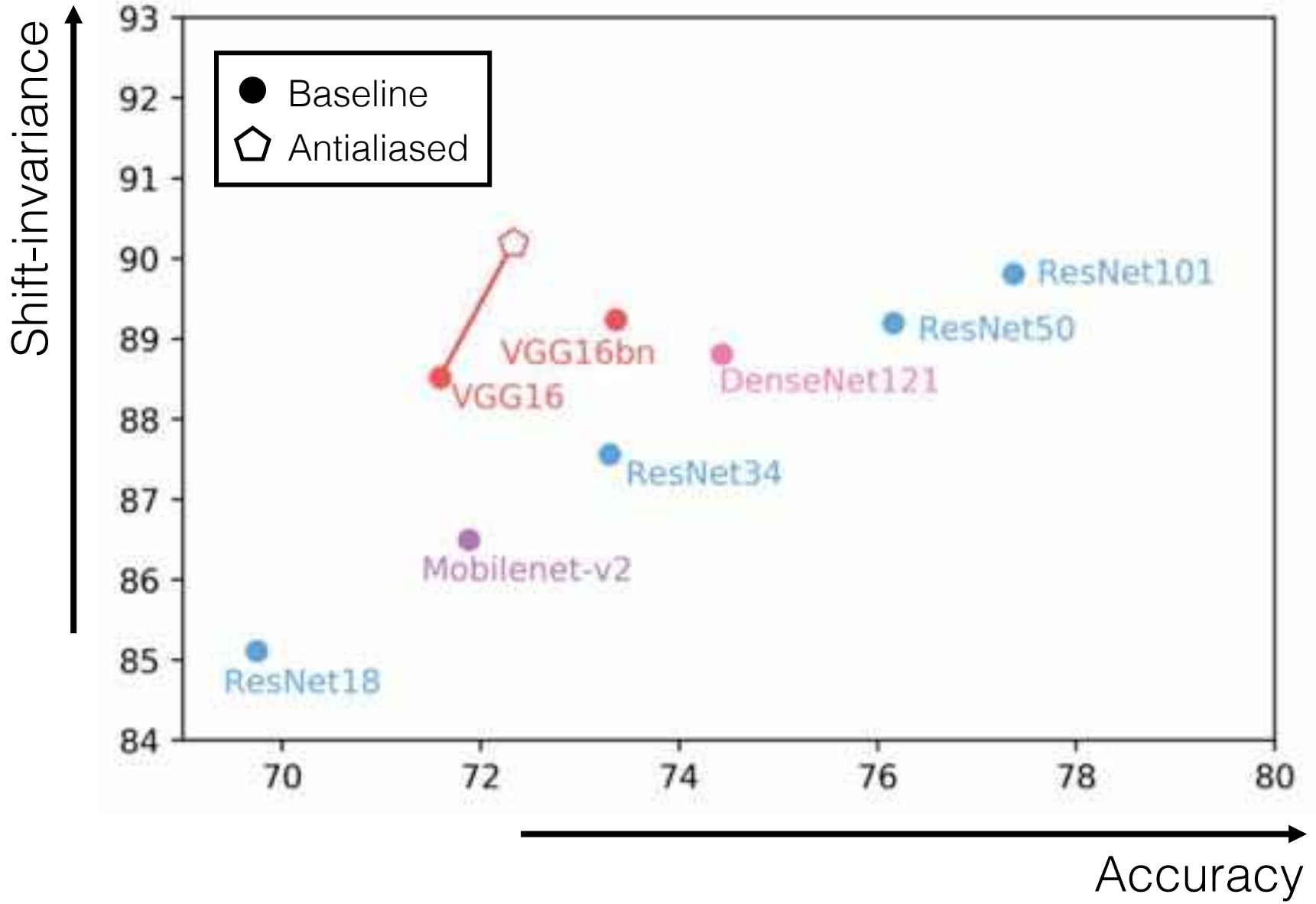
Average Pooling
(DenseNet)

Antialias any off-the-shelf network

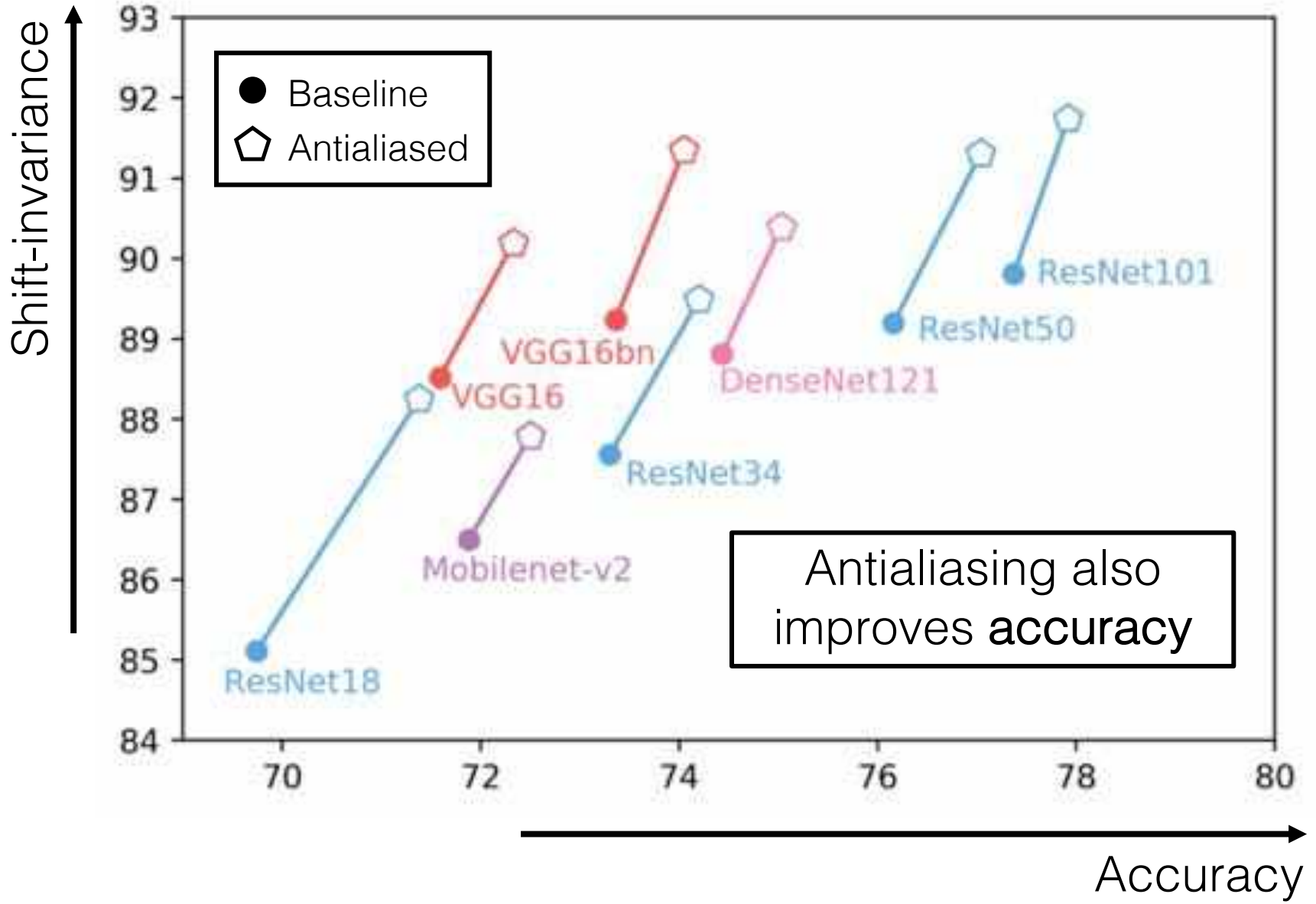
ImageNet


Shift-invariance
93
92
91
90
89
88
87
86
85
84
70
72
74
76
78
80
ResNet101
ResNet50
VGG16bn
DenseNet121
VGG16
ResNet34
Mobilenet-v2
ResNet18
Accuracy

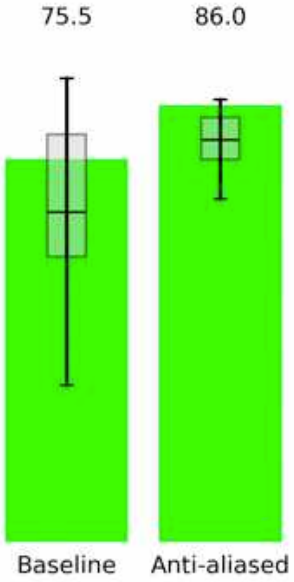
ImageNet

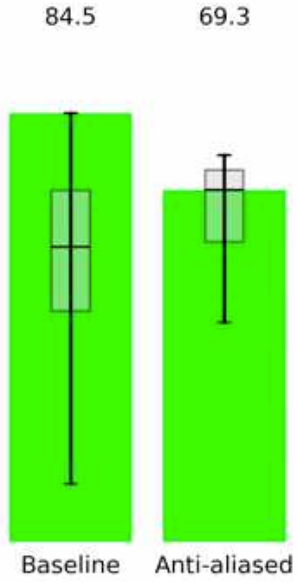

Shift-invariance
93
92
91
90
89
88
87
86
85
84
Baseline
Antialiased
ResNet101
ResNet50
VGG16bn
DenseNet121
VGG16
ResNet34
Mobilenet-v2
ResNet18
70
72
74
76
78
80
Accuracy

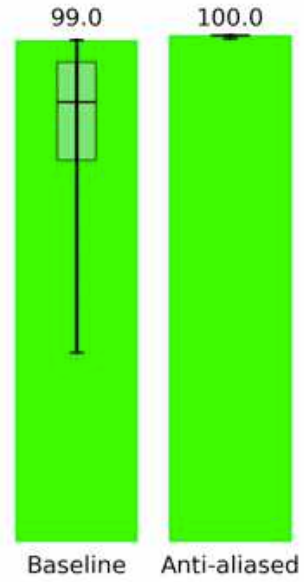
ImageNet

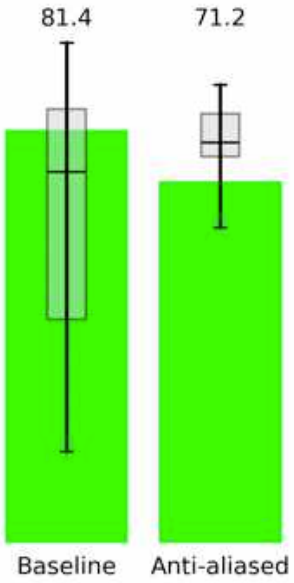

Shift-invariance
93
92
91
90
89
88
87
86
85
84
70
72
74
76
78
80
Baseline
Antialiased
ResNet18
VGG16
Mobilenet-v2
VGG16bn
ResNet34
DenseNet121
ResNet50
ResNet101
Antialiasing also improves **accuracy**
Accuracy

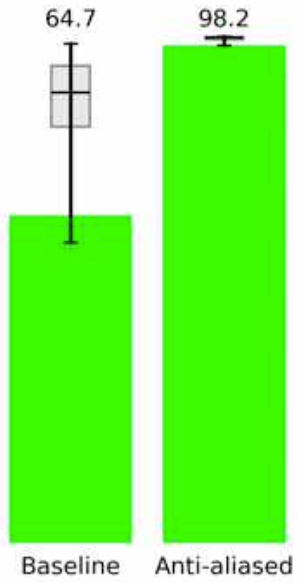
Qualitative examples

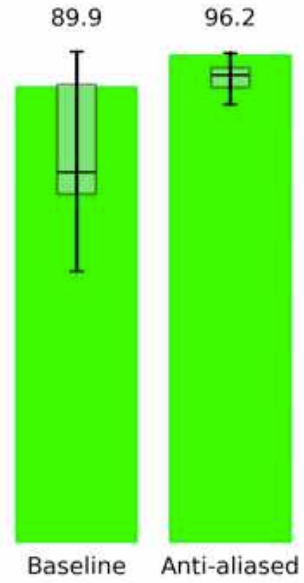












Discussion

Striding aliases

Antialiasing improves

+ shift-equivariance, accuracy
+ stability to perturbations
+ robustness to corruptions

Code + models: richzhang.github.io/antialiased-cnns/

Ex: `models.resnet50()` → `models_lpf.resnet50()`

→ Implications on image generation models?

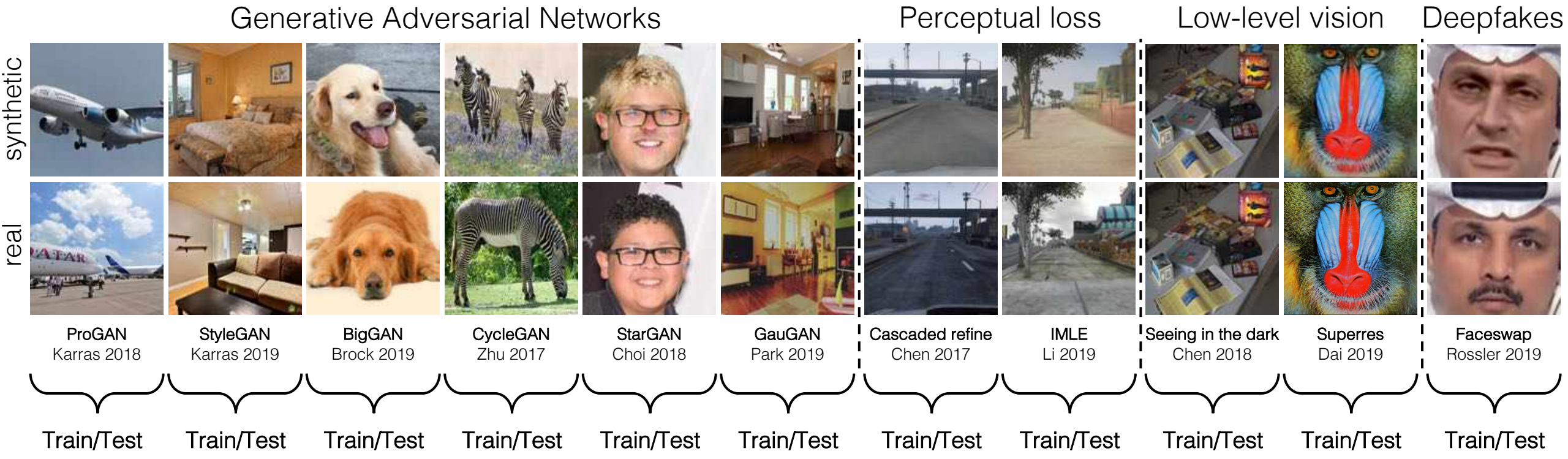
Making fake images is getting easier

“Deepfakes”

GANs

Can we create a “universal” detector?

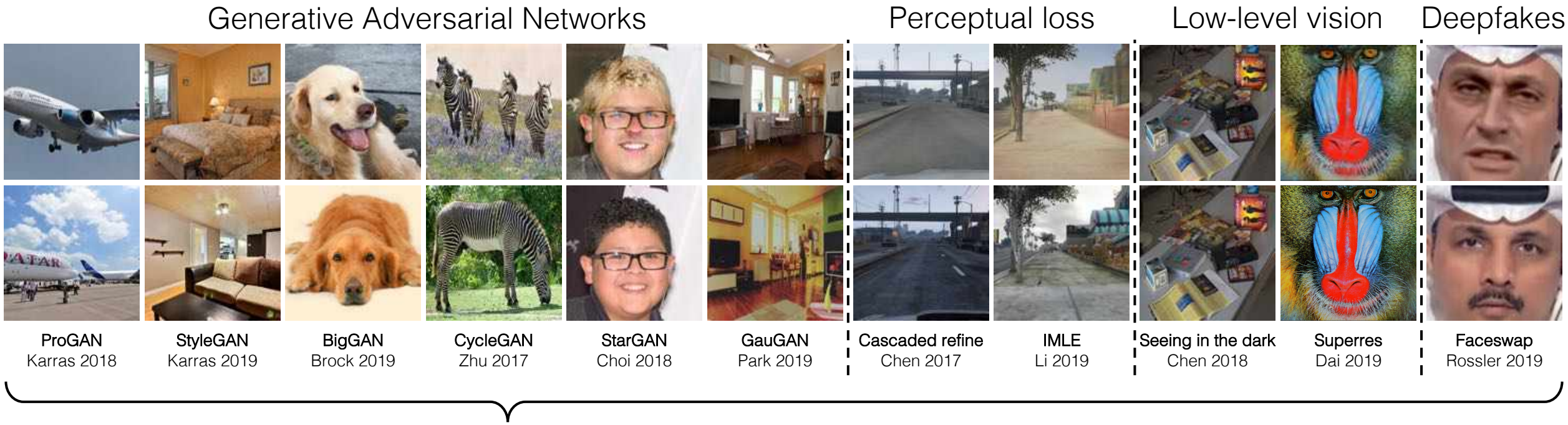
Dataset of CNN-generated fakes



Test: ??? et al. 20XX

Can we create a “universal” detector?

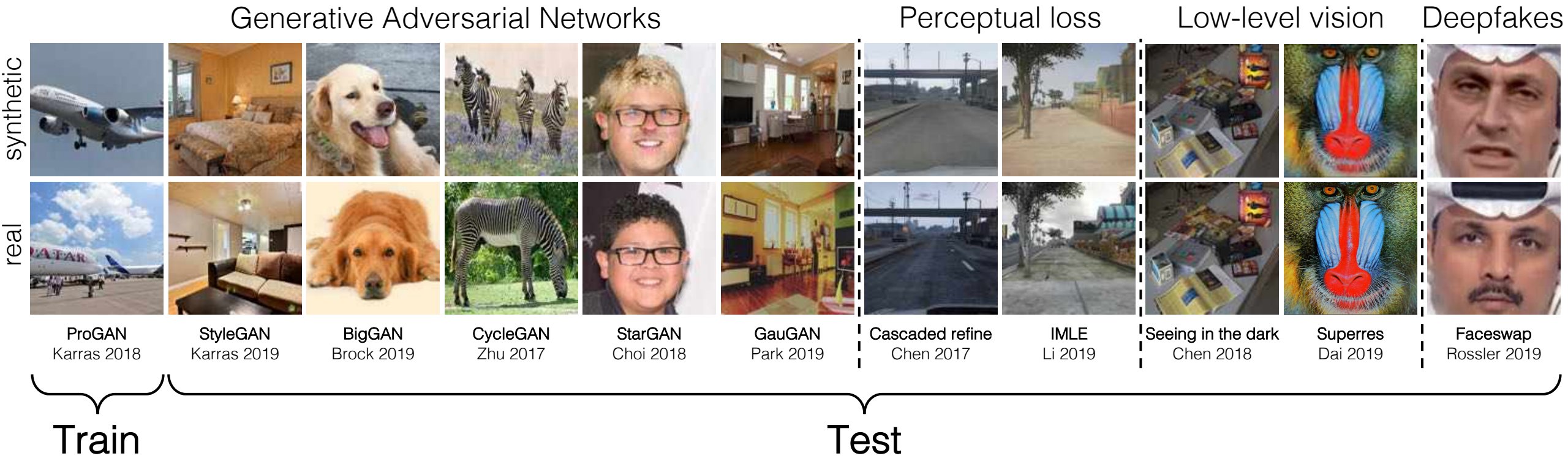
Dataset of CNN-generated fakes



Train ⟶ **Test:** ??? et al. 20XX

Can we create a “universal” detector?

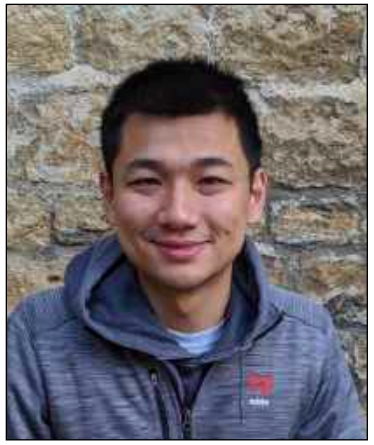
Dataset of CNN-generated fakes



Many **differences** (architecture, dataset, objective)

CNN-generated images are surprisingly easy to spot…for now





Sheng-Yu Wang Oliver Wang Richard Zhang Andrew Owens Alexei A. Efros

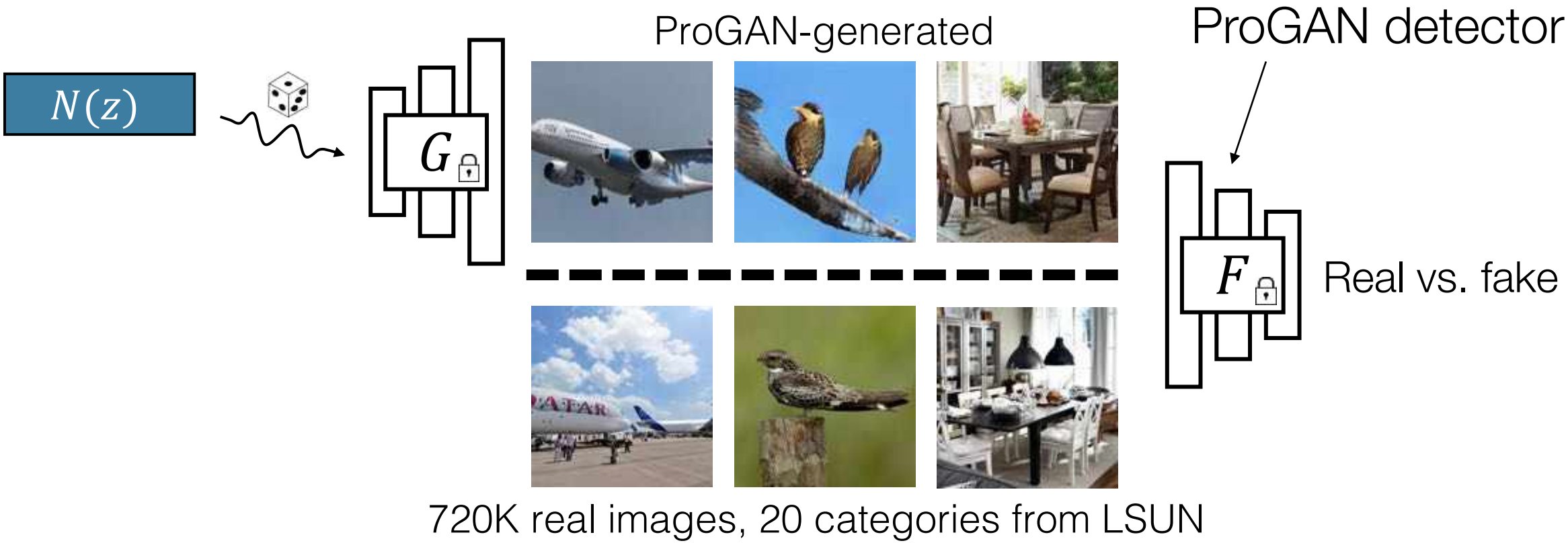
In **CVPR**, 2020 (oral).

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

UNIVERSITY OF MICHIGAN

Adobe

Training on ProGAN

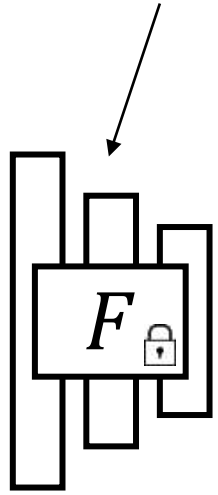


Testing across architectures

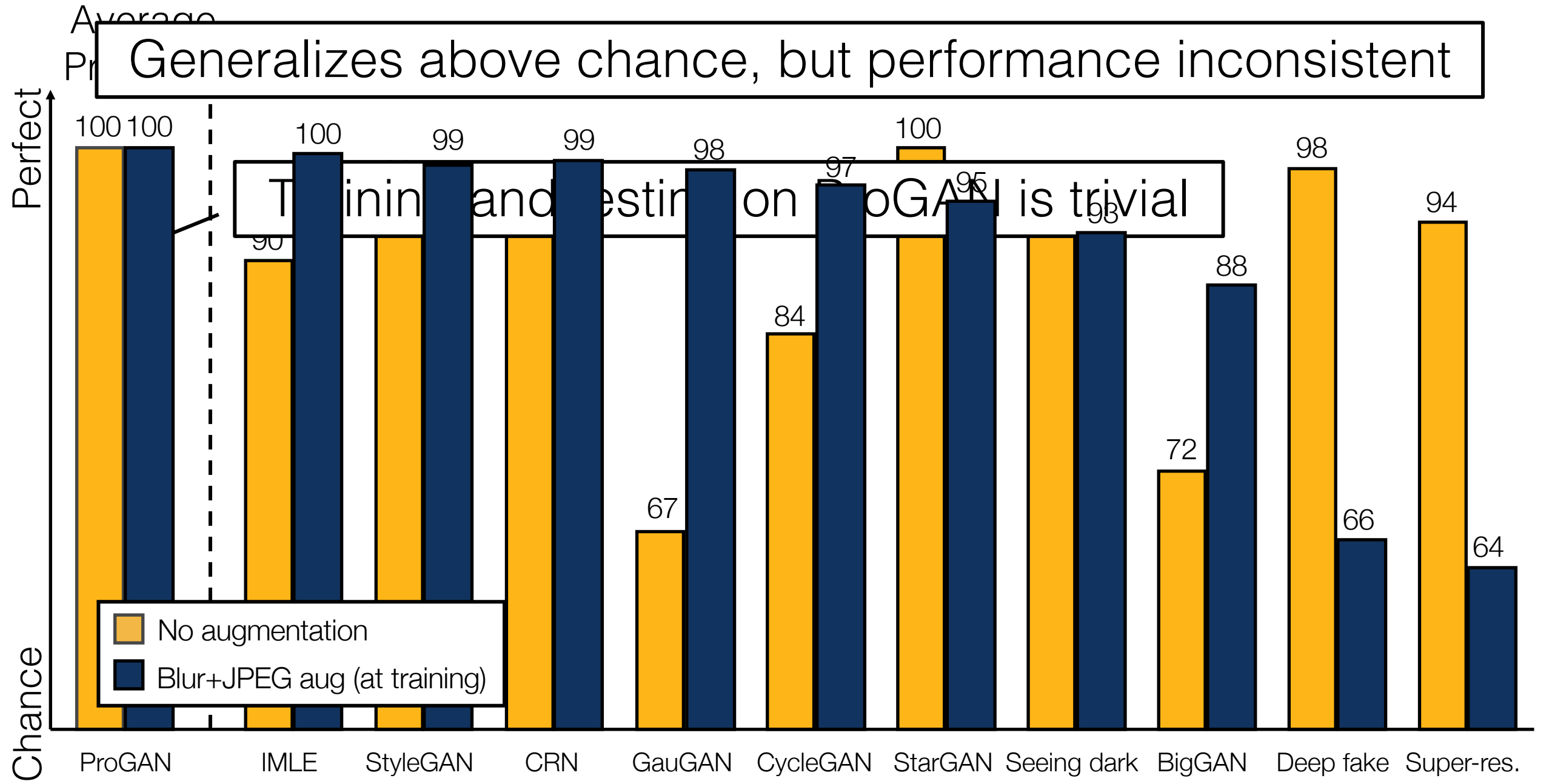
Synthesized images from **other** CNNs

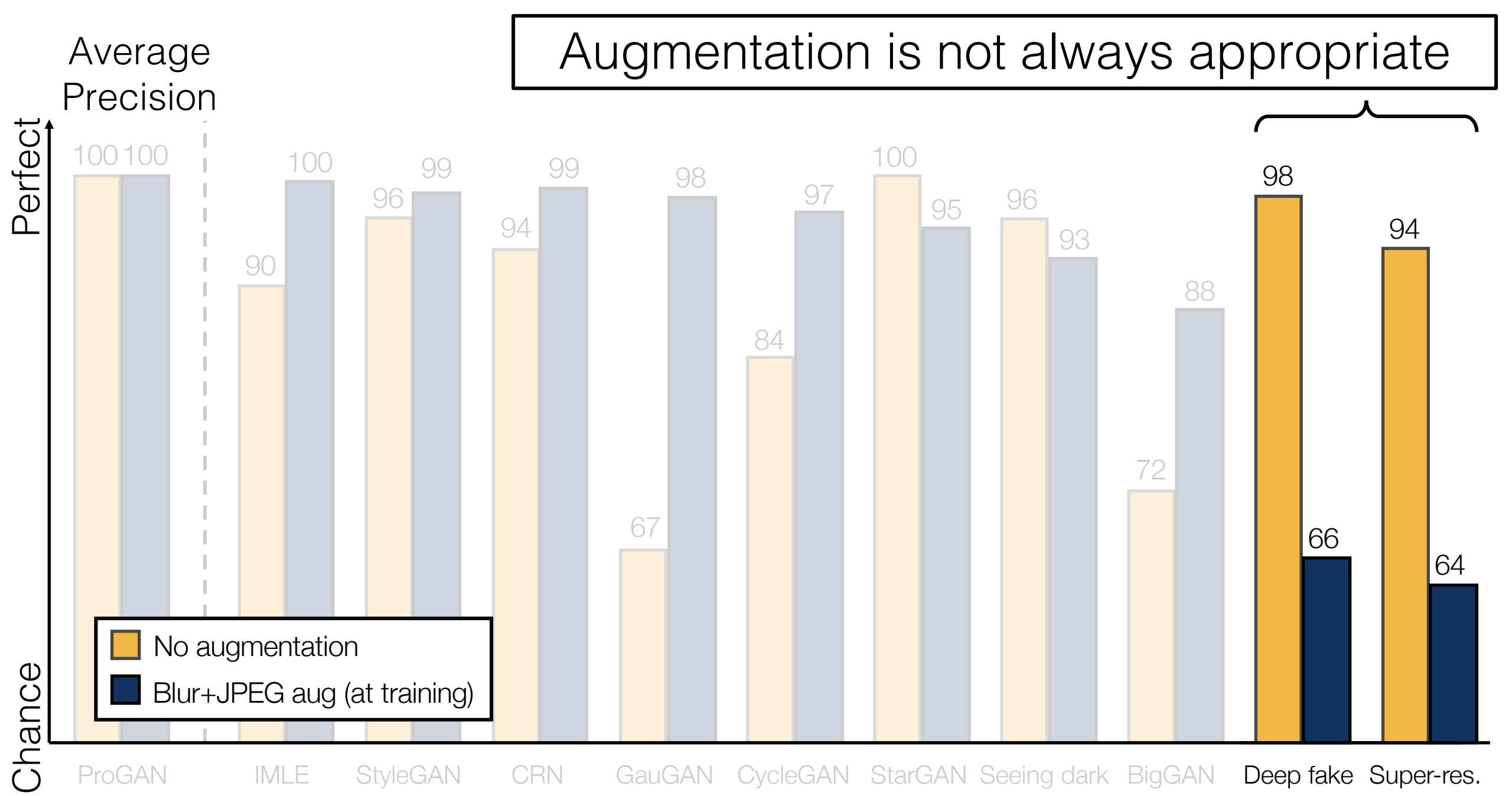
Real images

ProGAN detector

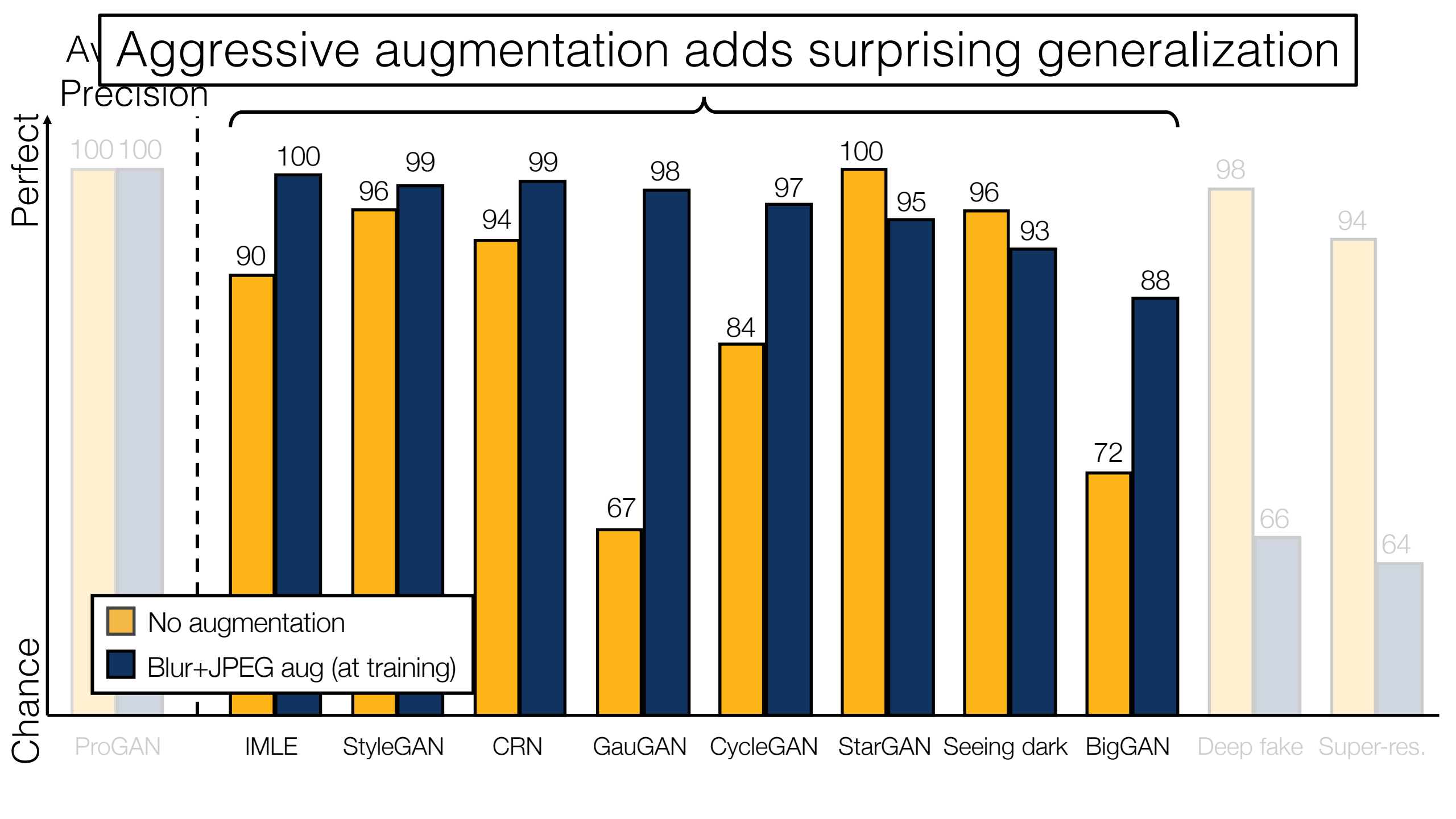


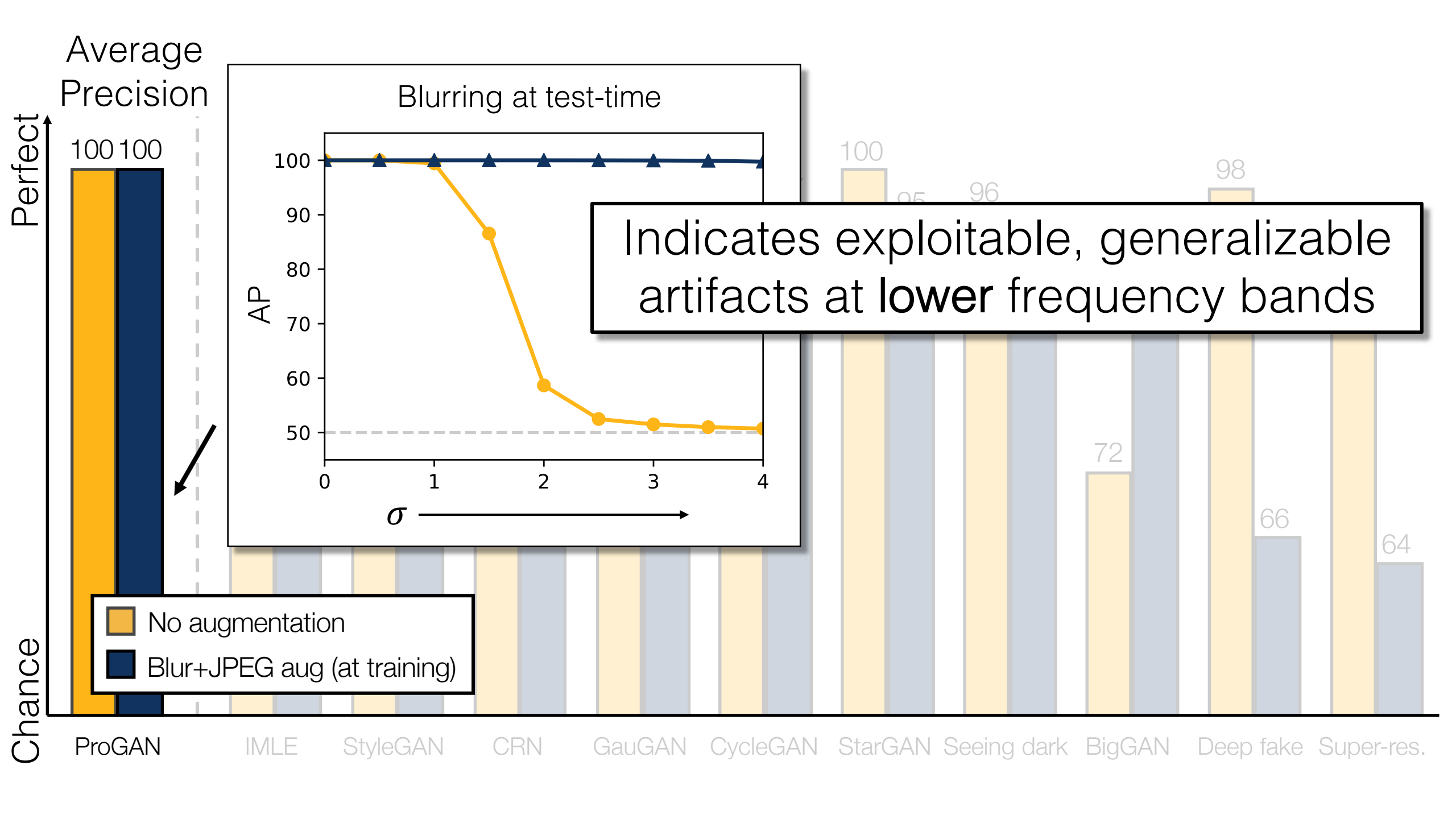
Real vs. fake


Average
Pr
Generalizes above chance, but performance inconsistent
Training and testing on ProGAN is trivial
Perfect
Chance
100 100
100
99
99
98
97
95
100
98
94
90
93
88
84
72
67
66
64
No augmentation
Blur+JPEG aug (at training)
ProGAN
IMLE
StyleGAN
CRN
GauGAN
CycleGAN
StarGAN
Seeing dark
BigGAN
Deep fake
Super-res.


Augmentation is not always appropriate
Average Precision
Perfect
Chance
100 100
90 100
96 99
94 99
67 98
84 97
100 95
96 93
72 88
98 66
94 64
No augmentation
Blur+JPEG aug (at training)
ProGAN
IMLE
StyleGAN
CRN
GauGAN
CycleGAN
StarGAN
Seeing dark
BigGAN
Deep fake
Super-res.

Aggressive augmentation adds surprising generalization




Average Precision
Perfect
Chance
100 100
Blurring at test-time
AP
100
90
80
70
60
50
0
1
2
3
4
σ
Indicates exploitable, generalizable artifacts at lower frequency bands
No augmentation
Blur+JPEG aug (at training)
100
95
96
98
72
66
64
ProGAN
IMLE
StyleGAN
CRN
GauGAN
CycleGAN
StarGAN
Seeing dark
BigGAN
Deep fake
Super-res.

Discussion

- Suggests CNN-generated images have common artifacts

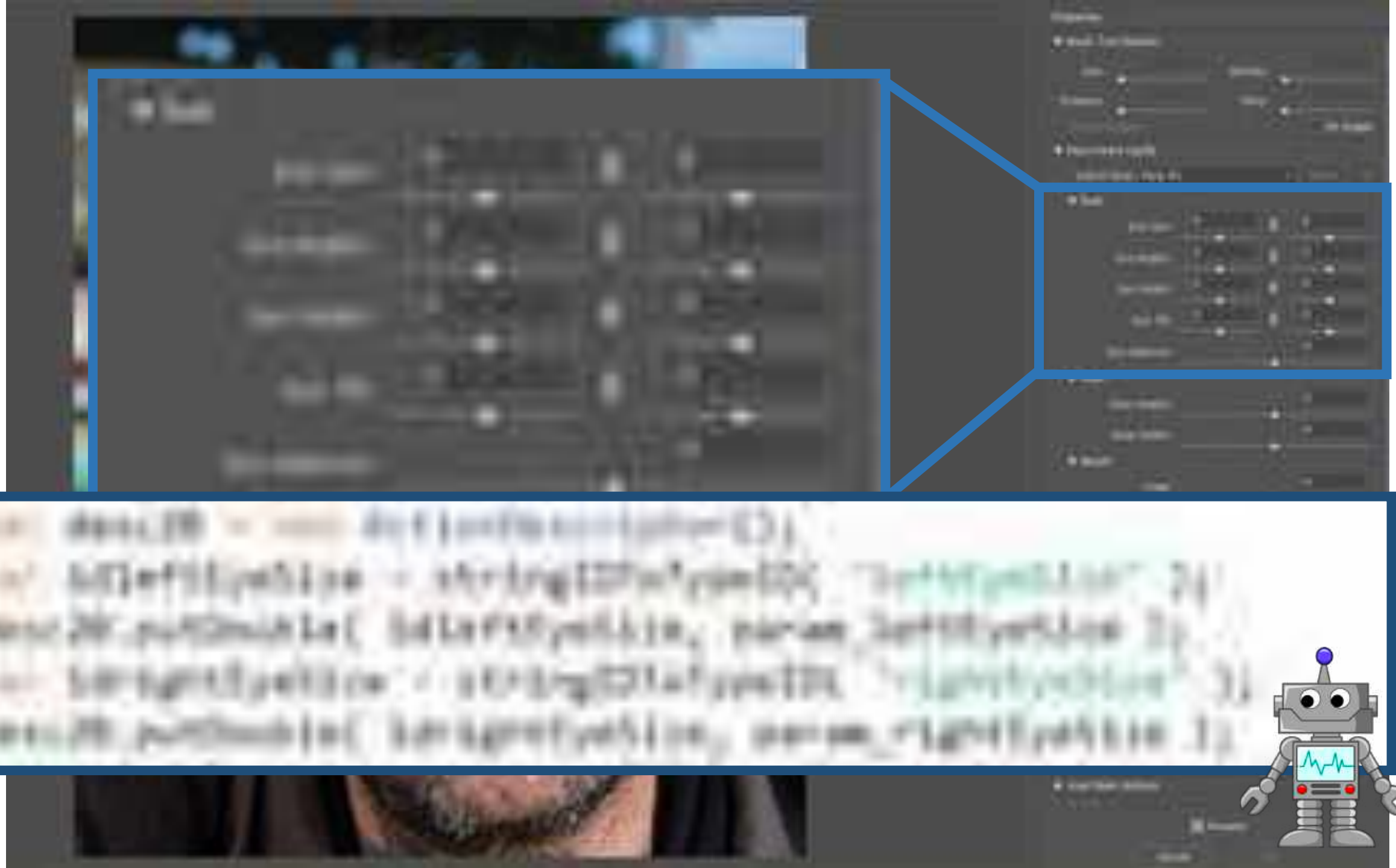
- Artifacts can be detected by a simple classifier!
  - StyleGAN2 (released **after** our submission): 100% AP on FFHQ
  - Swapping Autoencoder (Park et al.): 95% AP on FFHQ
  - **Note**: AP is computed on a collection of images;
    a real/fake decision on a per-image basis is more difficult

- Situation may not persist
  - GANs train with a discriminator
  - Future architecture changes
  - “Shallow” fakes, e.g., Photoshop

Media manipulation example

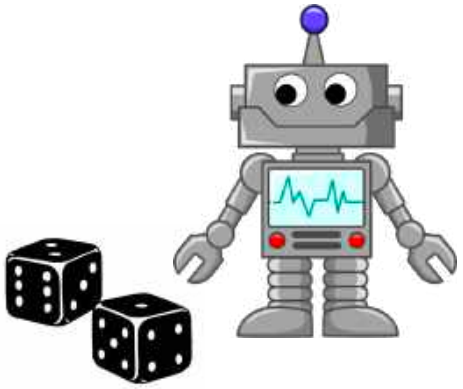
https://gizmodo.com/russian-state-tv-photoshops-an-awkward-smile-on-kim-jon-1826529277 June 2018

https://www.youtube.com/watch?v=5Qqv_C6iVvQ

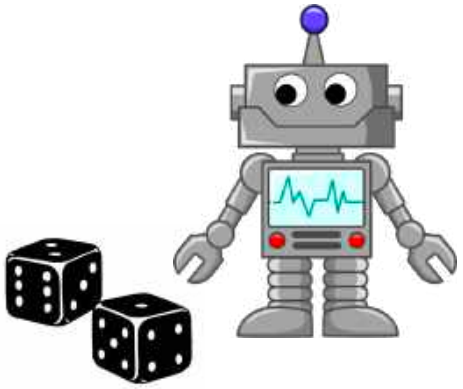


Original

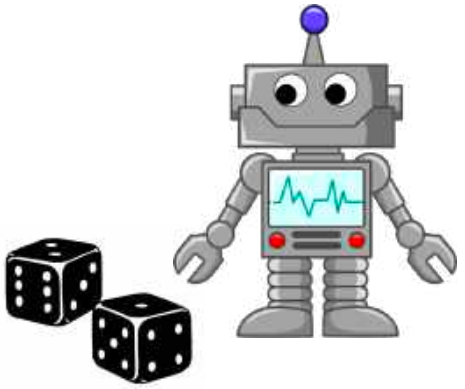
#1 modification



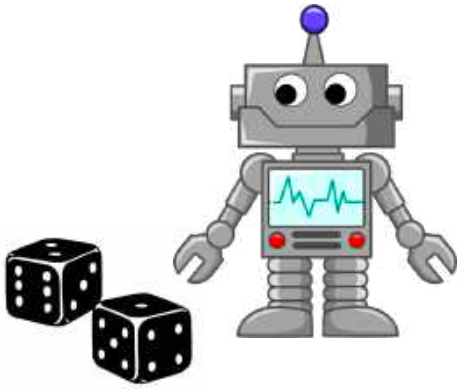
#2 modification



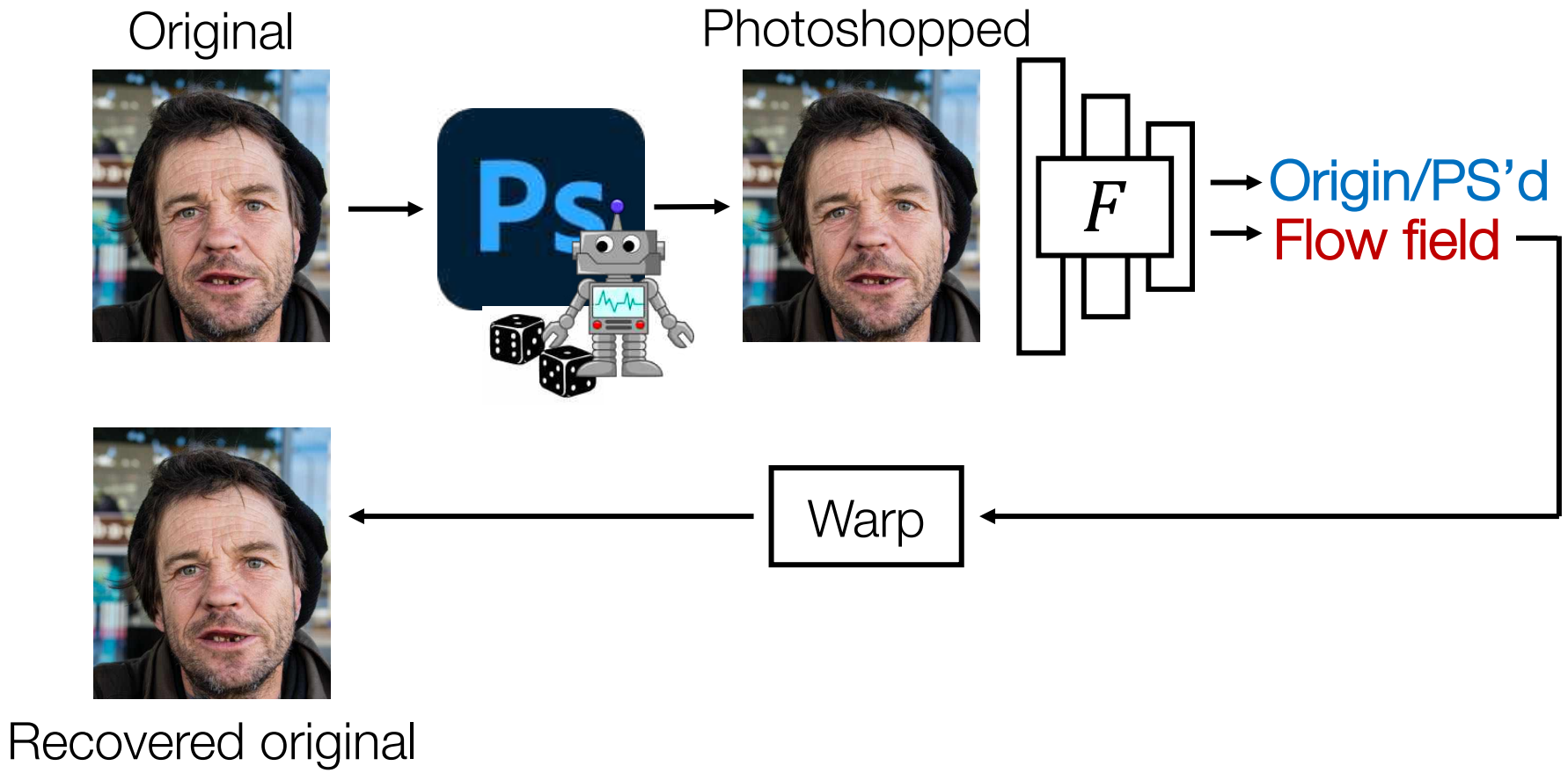
#3 modification



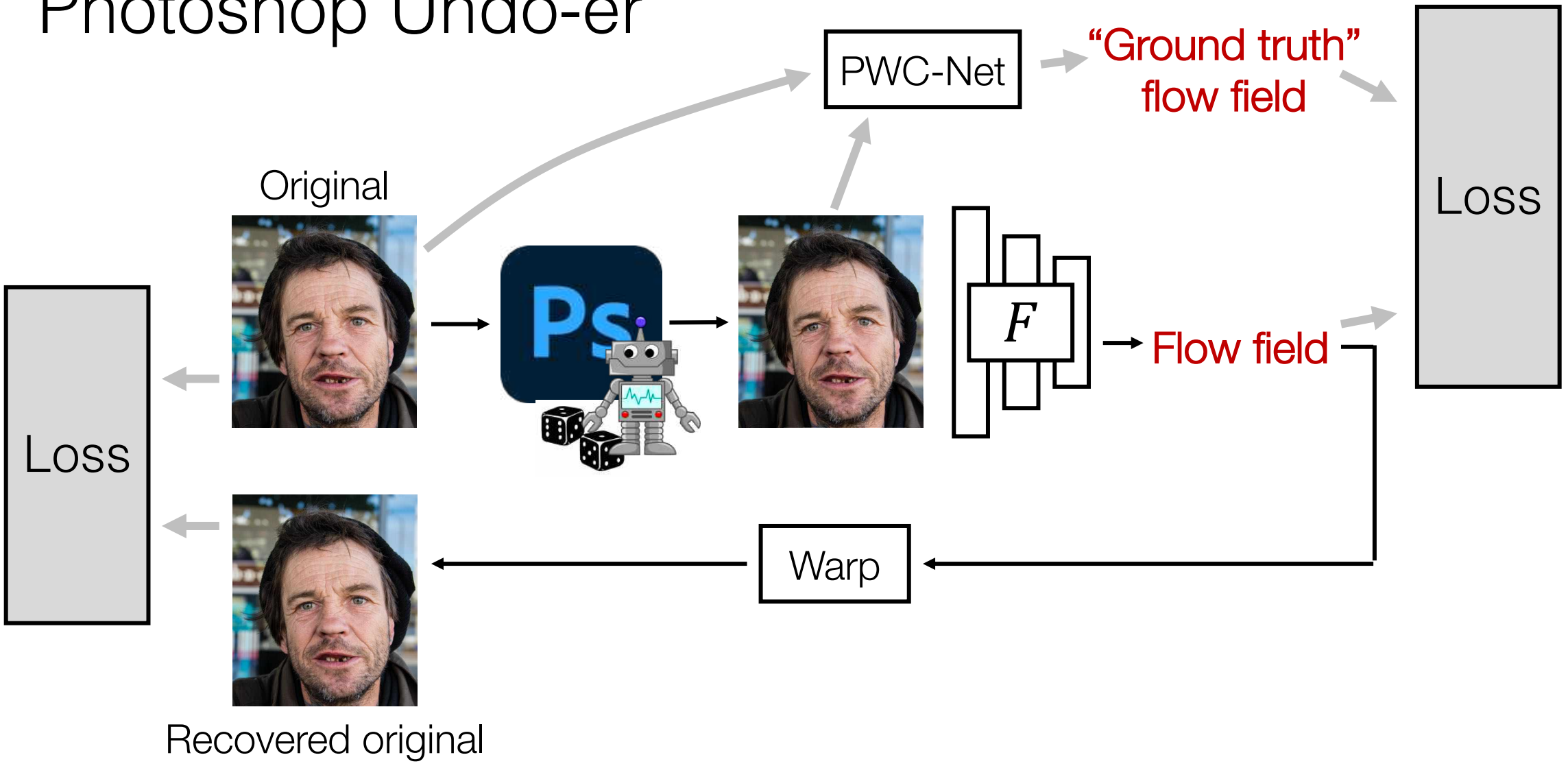
#4 modification



Photoshop Undo-er



Photoshop Undo-er



Manipulated

Flow prediction

Suggested “undo”

Original

Manipulated vs. Original

Undo vs. Original

Manipulated

Flow prediction

Suggested “undo”

Original

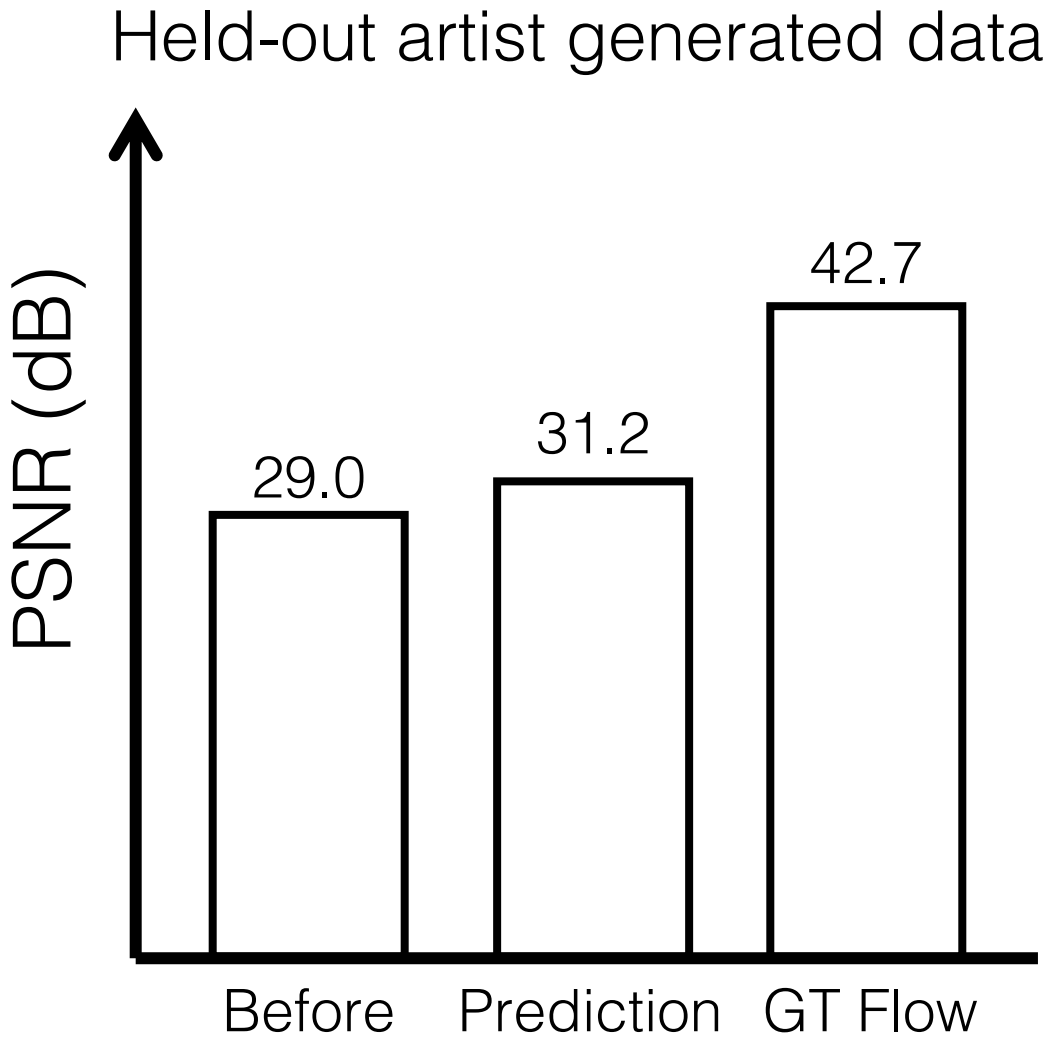
Manipulated vs. Original

Undo vs. Original

Reversal evaluation

Senses of generalization

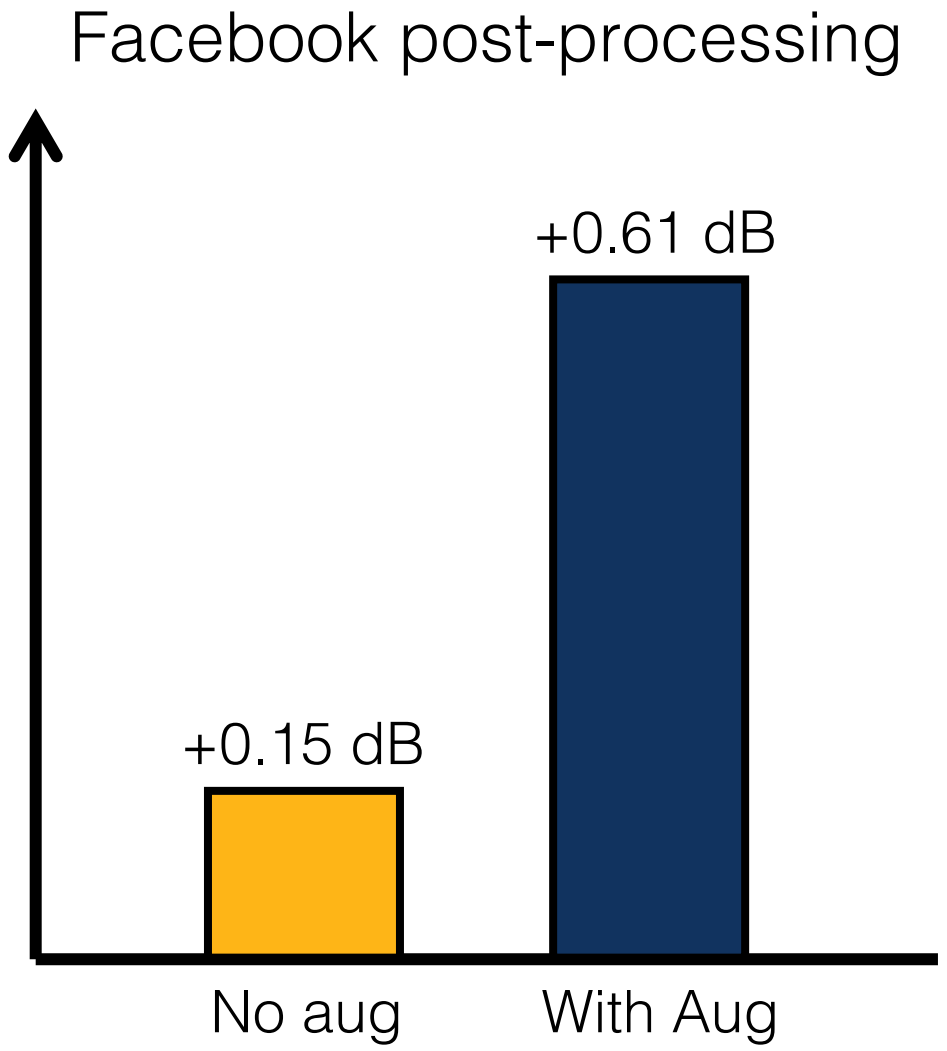
- Heldout artist data



Reversal evaluation

Senses of generalization

- Heldout artist data
- Post-processing



Data augmentation important (again)

Snapchat warps

Original Photo

Snapchat warps

Manipulated Photo

Snapchat warps

Flow Prediction

Snapchat warps

Suggested “Undo”

Snapchat warps

Some generalization across warp methods

Original Photo

Different image domain

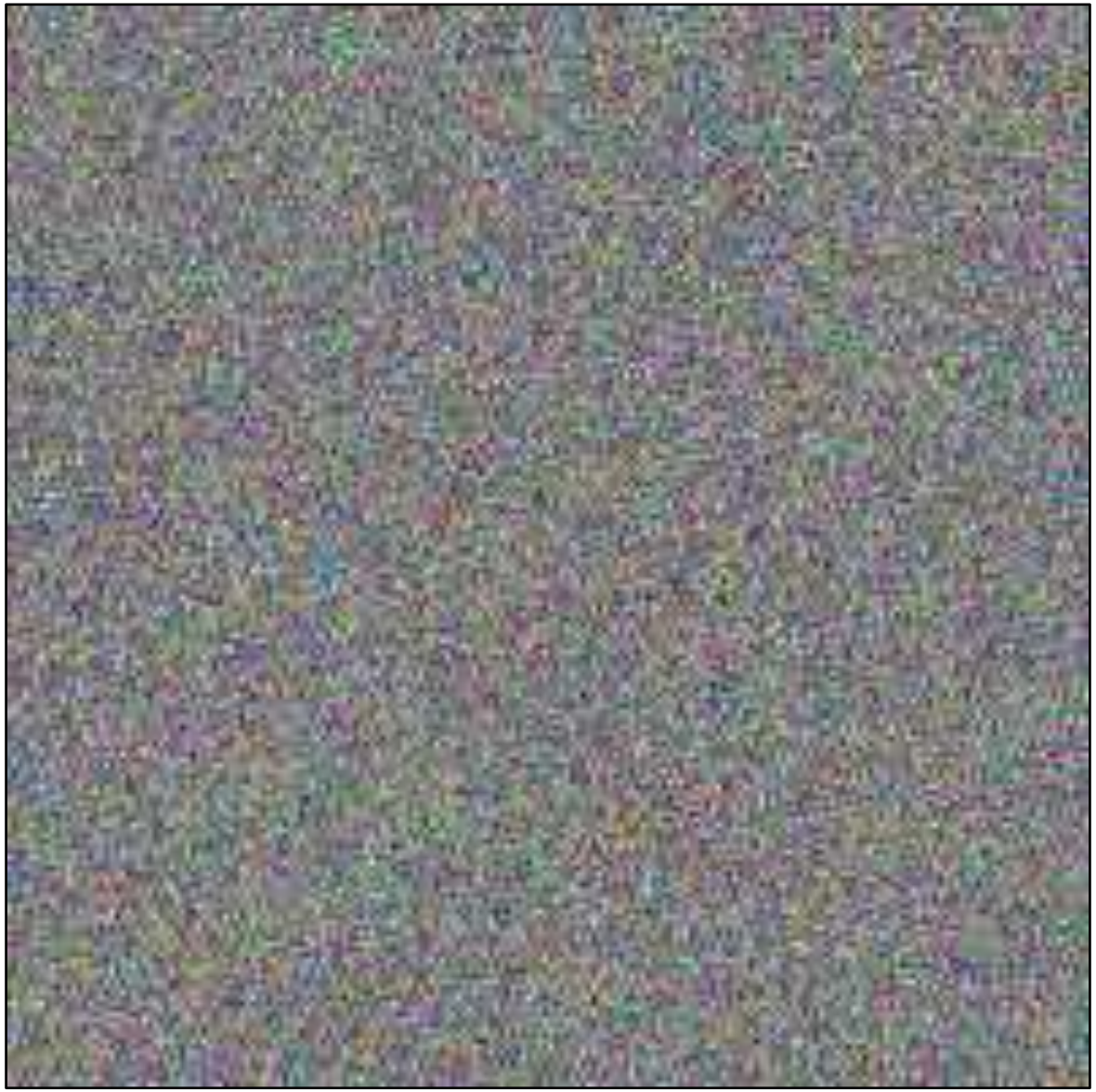
Different image domain

Different image domain

Predicted warp (not successful)

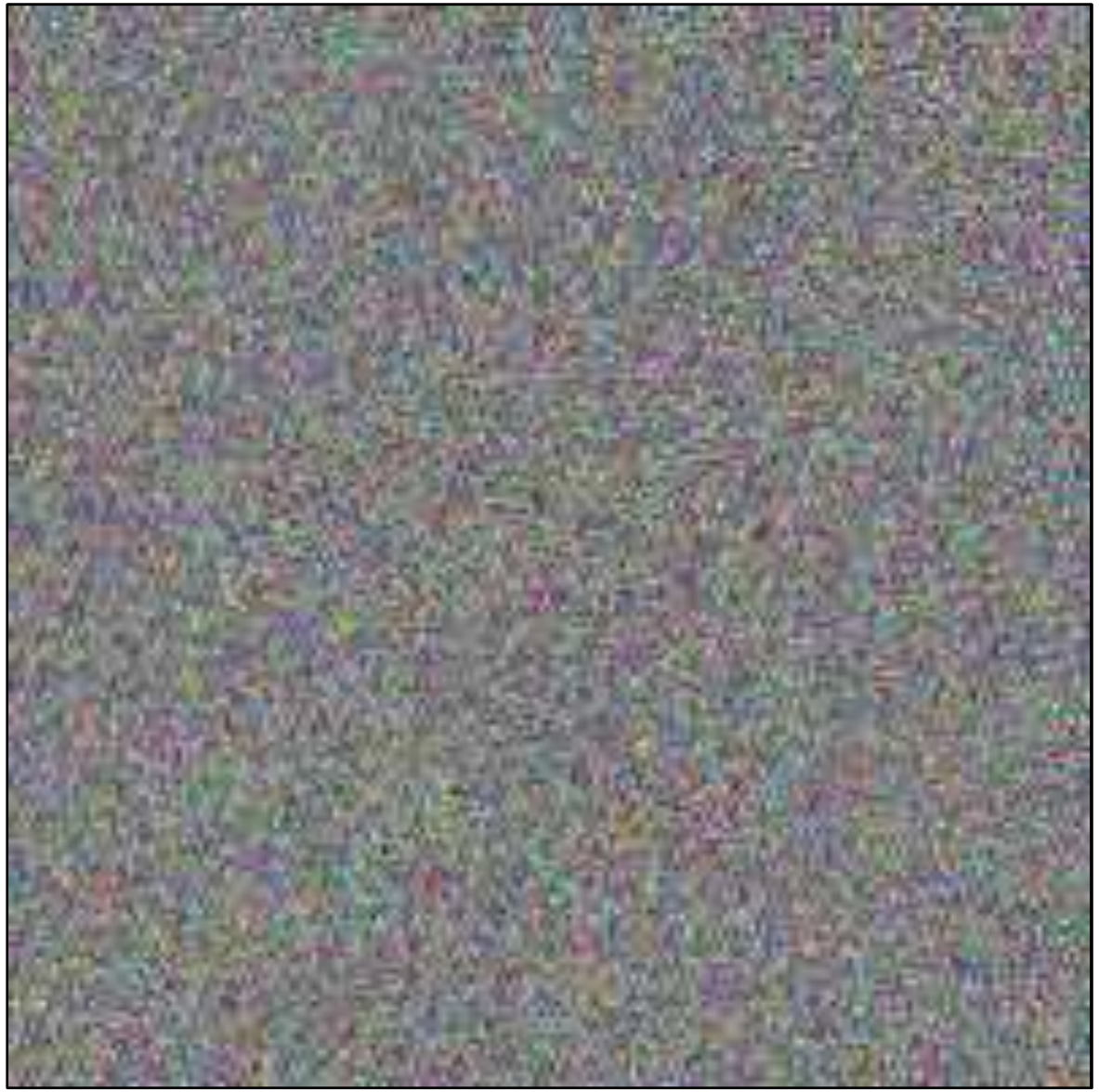
Does not generalize well to arbitrary image

Warped noise

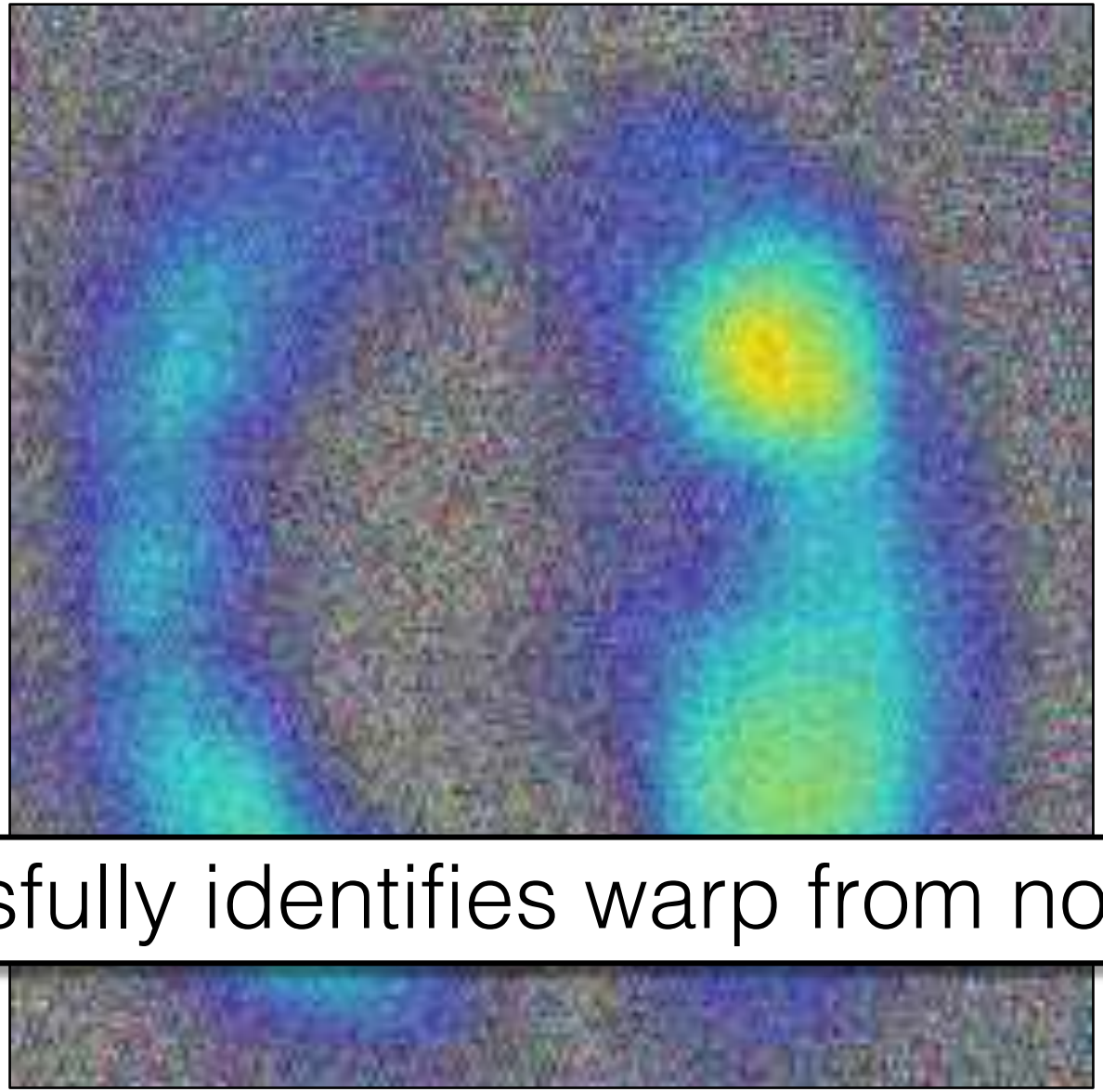


Warped noise

Warped noise



Warped noise



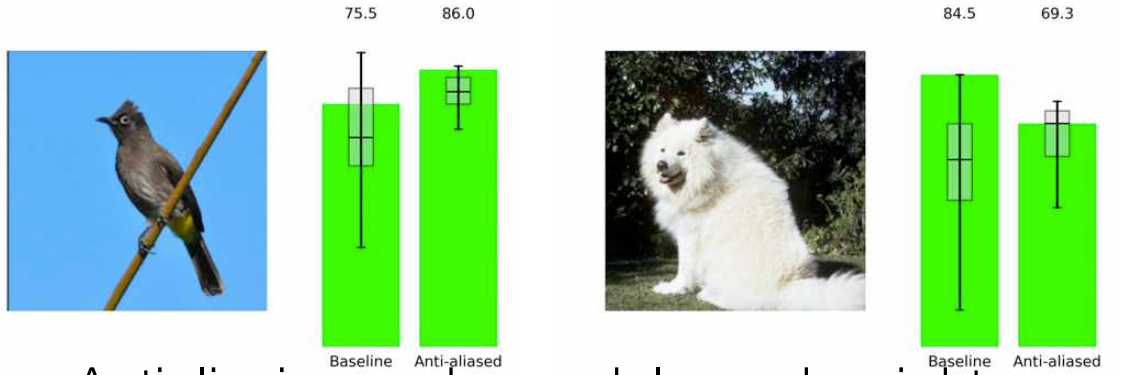
Successfully identifies warp from noise map

Discussion

- Suggests a multi-prong approach
- For rapidly evolving tools, continuous training and generalize
- For a relatively static tool, directly specialize
- Detection is only a piece of the puzzle
  - e.g., Content Authenticity Initiative: https://contentauthenticity.org/, collaboration between Adobe, New York Times, and Twitter

Making convolutional networks shift-invariant again.

R. Zhang. In **ICML**, 2019.



Antialiasing code, models, and weights
richzhang.github.io/antialiased-cnns/

CNN-generated images are surprisingly easy to spot…for now.

Wang, Wang, Zhang, Owens, Efros. In **CVPR**, 2020.

peterwang512.github.io/CNNDetection/

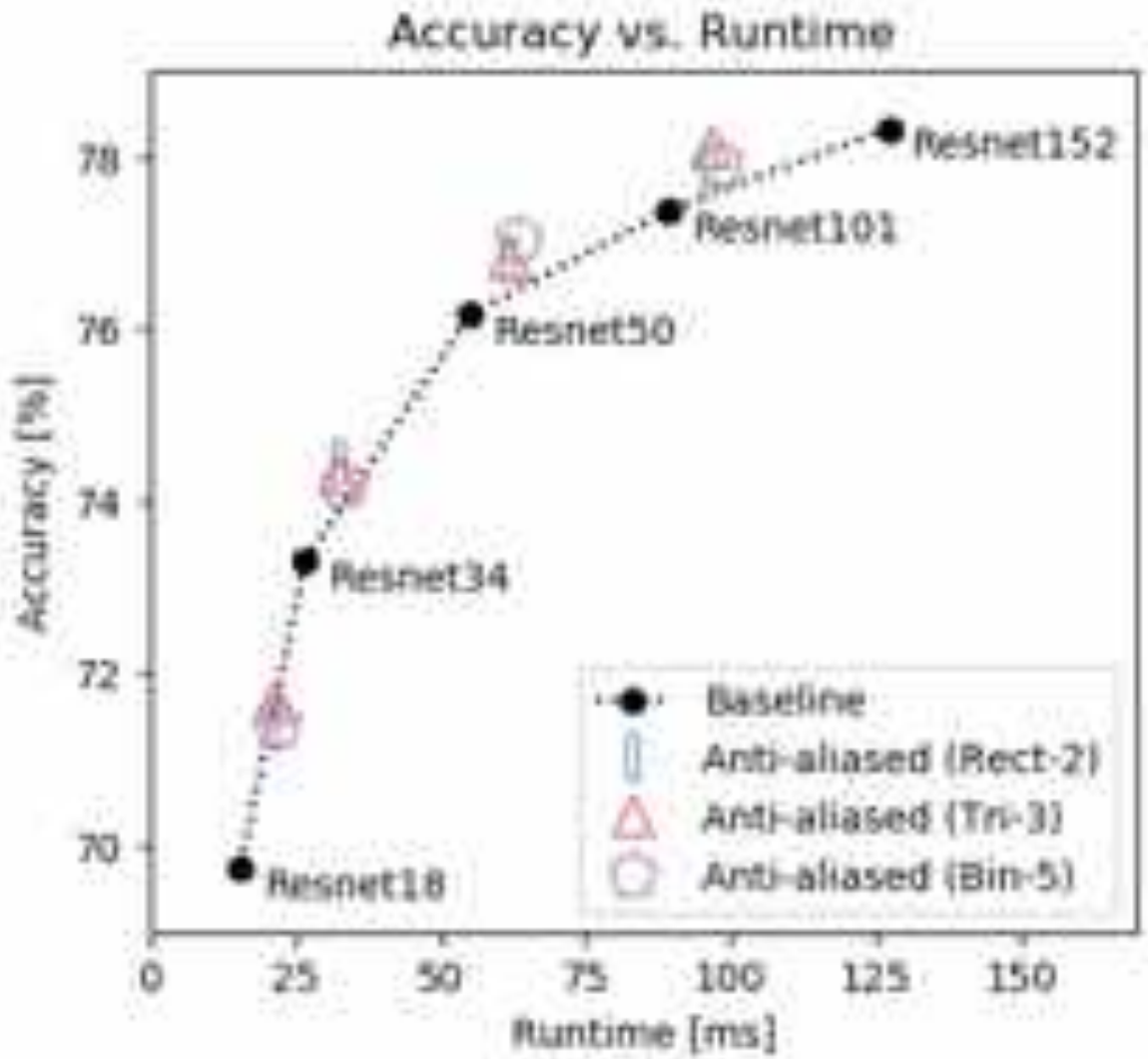
Detecting Photoshopped Images by Scripting Photoshop.

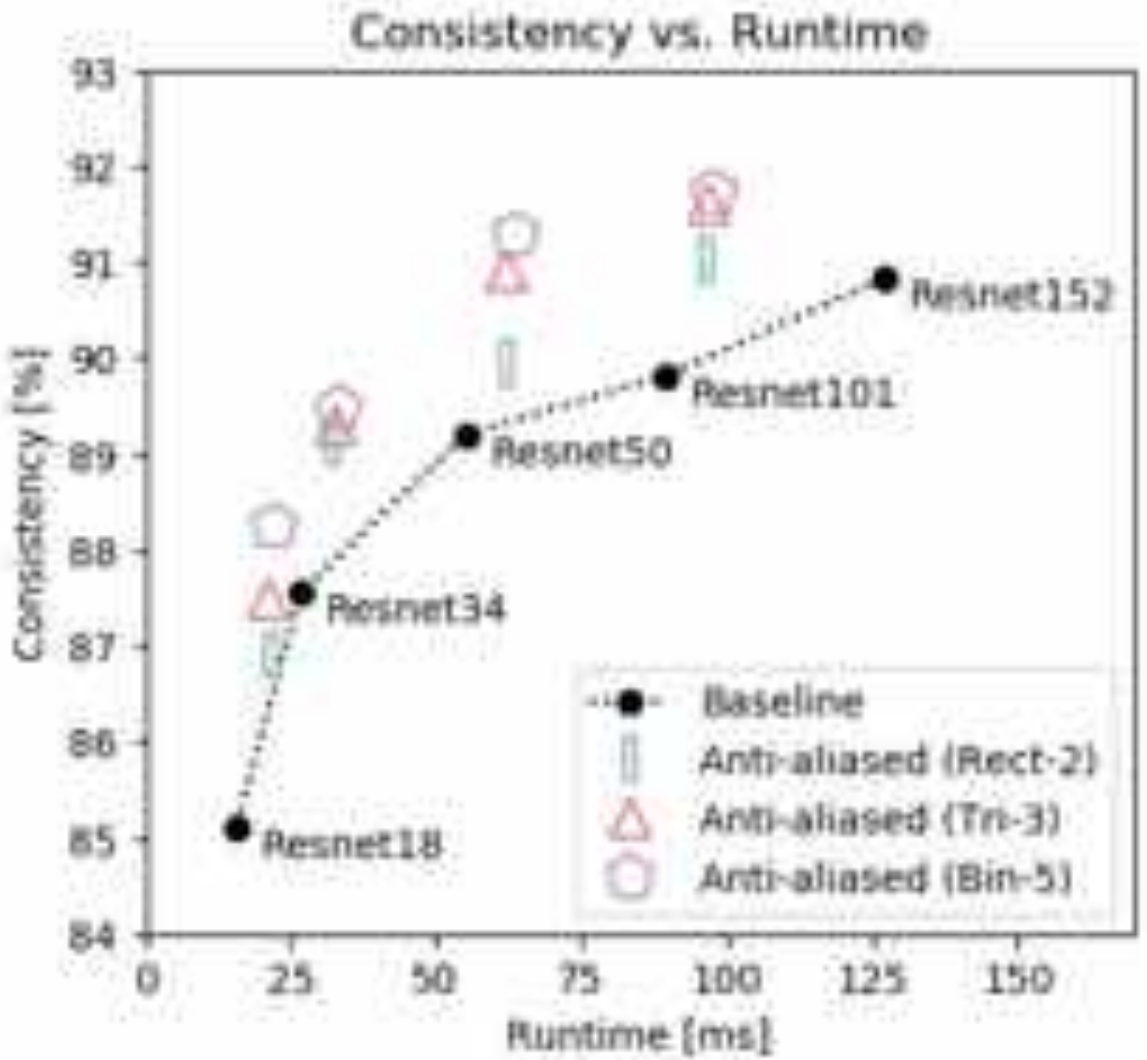
Wang, Wang, Owens, Zhang, Efros. In **ICCV**, 2019.

peterwang512.github.io/FALdetector/

Thank You!

Backup


Accuracy vs. Runtime
Accuracy [%]
Runtime [ms]
Resnet18
Resnet34
Resnet50
Resnet101
Resnet152
Baseline
Anti-aliased (Rect-2)
Anti-aliased (Tri-3)
Anti-aliased (Bin-5)


Consistency vs. Runtime
Consistency [%]
Runtime [ms]
Resnet18
Resnet34
Resnet50
Resnet101
Resnet152
Baseline
Anti-aliased (Rect-2)
Anti-aliased (Tri-3)
Anti-aliased (Bin-5)

ResNet50 on ImageNet-P (Hendrycks et al., 2019)

Flip Rate (FR) (lower is better)

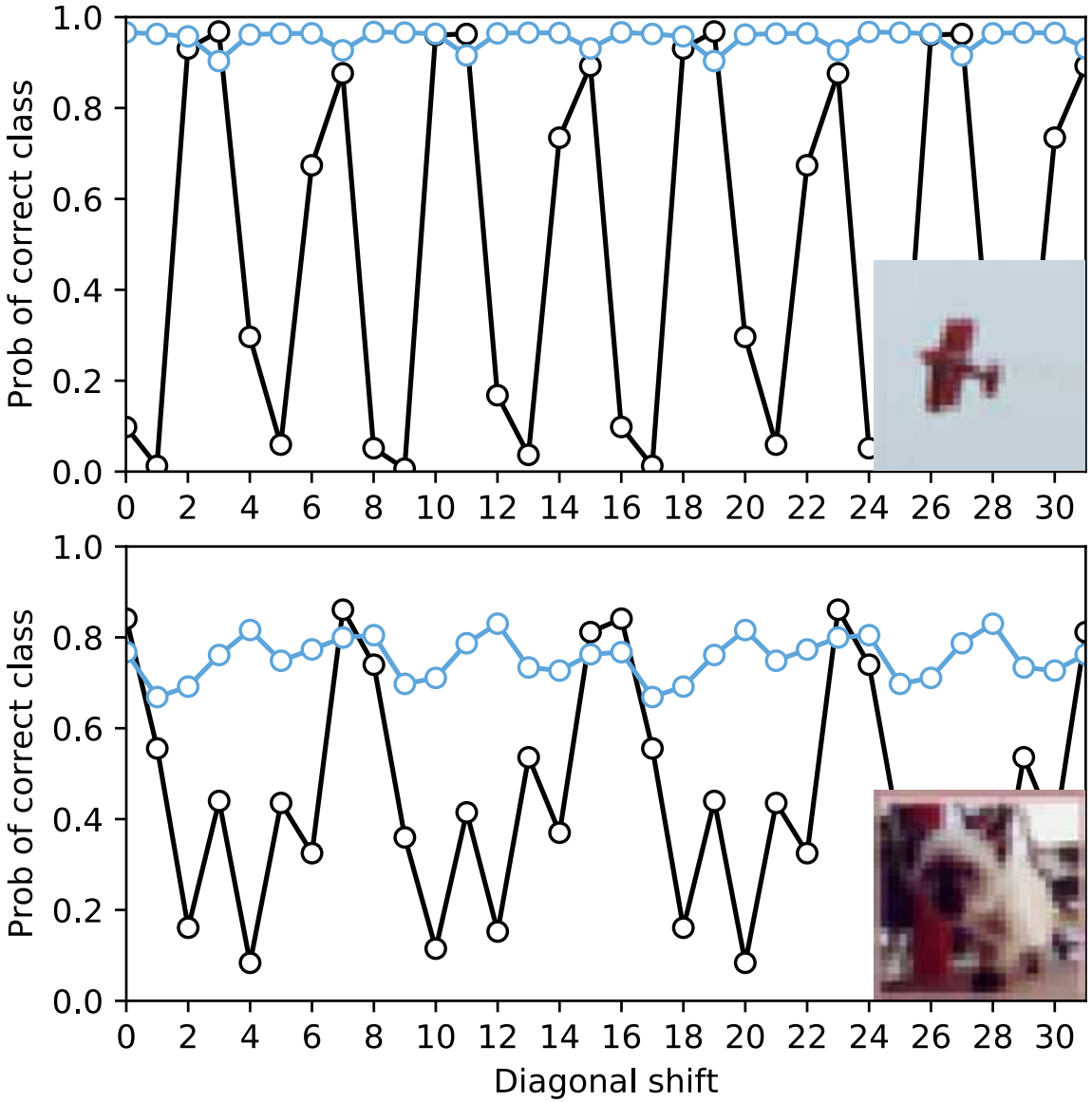
	Noise		Blur		Weather		Geometric				Mean	
	Gauss	Shot	Motion	Zoom	Snow	Bright	Translate	Rotate	Tilt	Scale	Unnorm	Norm
Baseline	14.04	17.38	6.00	4.29	7.54	3.03	4.86	6.79	4.01	11.32	7.92	57.99
Rect-2	14.08	17.16	5.98	4.21	7.34	3.20	4.42	6.43	3.80	10.61	7.72	56.70
Tri-3	12.59	15.57	**5.39**	3.79	6.98	**3.01**	3.95	5.80	3.53	9.90	7.05	51.91
Bin-5	**12.39**	**15.22**	5.44	**3.72**	**6.76**	3.15	**3.78**	**5.67**	**3.44**	**9.45**	**6.90**	**51.18**

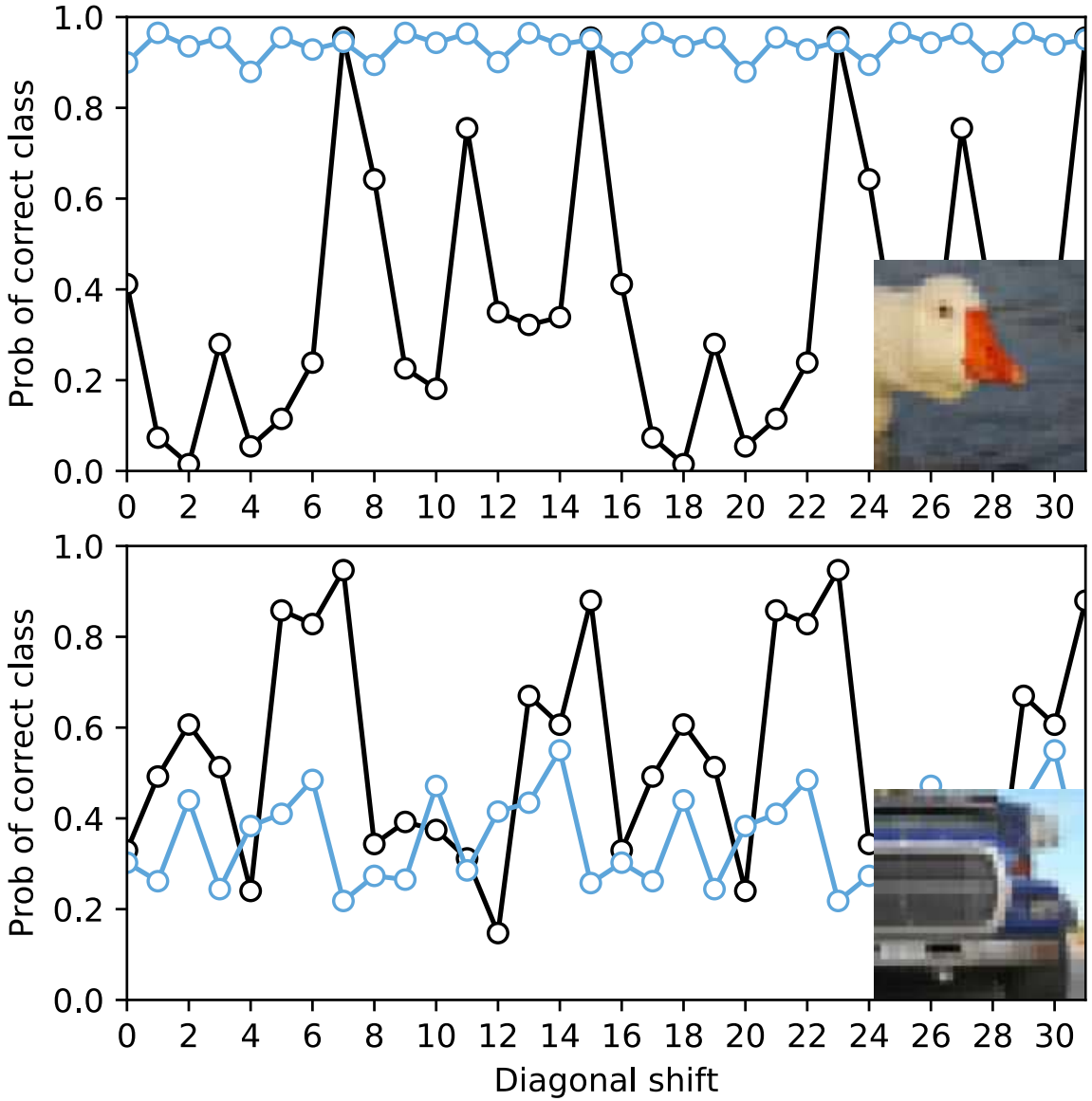
ResNet50 on ImageNet-C (Hendrycks et al., 2019)

Corruption Error (CE) (lower is better)

	Noise			Blur				Weather				Digital				Mean	
	Gauss	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	Pixel	Jpeg	Unnorm	Norm
Baseline	68.70	71.10	74.04	61.40	73.39	61.43	63.93	67.76	62.08	54.61	32.04	61.25	**55.24**	55.24	46.32	60.57	76.43
Rect-2	65.81	68.27	70.49	60.01	72.14	62.19	63.96	68.00	61.83	54.95	32.09	60.25	55.56	53.89	43.62	59.54	75.16
Tri-3	**63.86**	**66.07**	**69.15**	**58.36**	71.70	**60.74**	61.58	**66.78**	60.29	54.40	**31.48**	**58.09**	55.26	53.89	43.62	58.35	73.73
Bin-5	64.31	66.39	69.88	60.31	**71.37**	61.60	**61.25**	66.82	**59.82**	**51.84**	31.51	58.12	55.29	**50.81**	**42.84**	**58.14**	**73.41**

Baseline vs Anti-aliased





Loss Function

- End point error :

$$\Sigma_{p \in I} ||U(p) - \hat{U}(p)||_2$$

Loss Function

- Multi-scale gradient error:

$$\Sigma_{p \in I} \Sigma_s (||\nabla_x^s U(p) - \nabla_x^s \hat{U}(p)||_2 + ||\nabla_y^s U(p) - \nabla_y^s \hat{U}(p)||_2$$

$$\nabla_x^s A \equiv A[..., s:] - A[..., :-s]$$

Loss Function

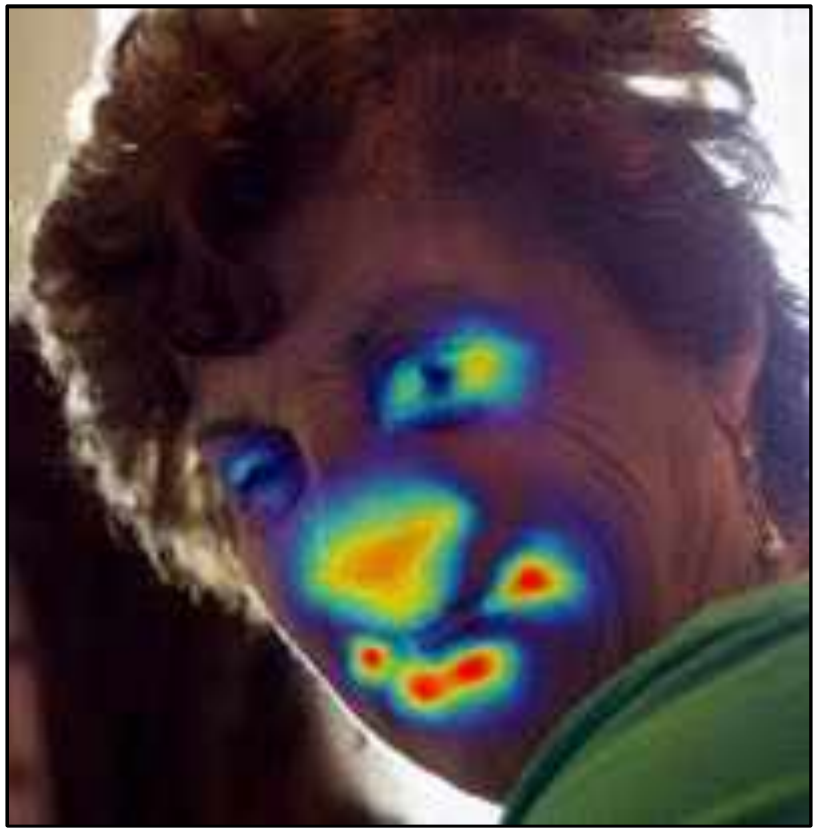
- Reconstruction error:

$$\Sigma_{p \in I} ||X_{orig}(p) - X_{reconst}(p)||_1$$

Loss Function

- Reconstruction error:

$$\Sigma_{p \in I} ||X_{orig}(p) - X_{reconst}(p)||_1$$



Loss Function

- Reconstruction error:

$$\Sigma_{p \in I} ||X_{orig}(p) - X_{reconst}(p)||_1$$