

Deep Scale-spaces: Equivariance Over Scale

Daniel E. Worrall*
 AMLAB, Philips Lab
 University of Amsterdam
 d.e.worrall@uva.nl

Max Welling
 AMLAB, Philips Lab
 University of Amsterdam
 m.welling@uva.nl

Abstract

We introduce deep scale-spaces (DSS), a generalization of convolutional neural networks, exploiting the scale symmetry structure of conventional image recognition tasks. Put plainly, the class of an image is invariant to the scale at which it is viewed. We construct scale equivariant cross-correlations based on a principled extension of convolutions, grounded in the theory of scale-spaces and semigroups. As a very basic operation, these cross-correlations can be used in almost any modern deep learning architecture in a plug-and-play manner. We demonstrate our networks on the Patch Camelyon and Cityscapes datasets, to prove their utility and perform introspective studies to further understand their properties.

1 Introduction

Scale is inherent in the structure of the physical world around us and the measurements we make of it. Ideally, the machine learning models we run on this perceptual data should have a notion of scale, which is either learnt or built directly into them. However, the state-of-the-art models of our time, convolutional neural networks (CNNs) [Lecun et al., 1998], are predominantly local in nature due to small filter sizes. It is not thoroughly understood how they account for and reason about multiscale interactions in their deeper layers, and empirical evidence [Chen et al., 2018, Yu and Koltun, 2015, Yu et al., 2017] using dilated convolutions suggests that there is still work to be done in this arena.

In computer vision, typical methods to circumvent scale are: *scale averaging*, where multiple scaled versions of an image are fed through a network and then averaged [Kokkinos, 2015]; *scale selection*, where an object’s scale is found and then the region is resized to a canonical size before being passed to a network [Girshick et al., 2014]; and *scale augmentation*, where multiple scaled versions of an image are added to the training set [Barnard and Casasent, 1991]. While these methods help, they lack explicit mechanisms to fuse information from different scales into the same representation. In this work, we construct a generalized convolution taking, as input, information from different scales.

The utility of convolutions arises in scenarios where there is a *translational symmetry* (translation invariance) inherent in the task of interest [Cohen and Welling, 2016a]. Examples of such tasks are object classification [Krizhevsky et al., 2012], object detection [Girshick et al., 2014], or dense image labelling [Long et al., 2015]. By using translational weight-sharing [Lecun et al., 1998] for these tasks, we reduce the parameter count while preserving symmetry in the deeper layers. The overall effect is to improve sample complexity and thus reduce generalization error [Sokolic et al., 2017]. Furthermore, it has been shown that convolutions (and various reparameterizations of them) are the only linear operators that preserve symmetry [Kondor and Trivedi, 2018]. Attempts have been made to extend convolutions to scale, but they either suffer from breaking translation symmetry [Henriques and Vedaldi, 2017, Esteves et al., 2017], making the assumption that scalings can be modelled in the same way as rotations [Marcos et al., 2018], or ignoring symmetry constraints [Hilbert et al., 2018]. The problem with the aforementioned approaches is that they fail to account for the unidirectional nature of scalings. In data there exist many one-way transformations, which cannot be inverted.

*deworrall92.github.io



Figure 1: How to correctly downsample an image. Left to right: The original high-resolution image; an $\times 1/8$ subsampled image, notice how a lot of the image structure has been destroyed; a high-resolution, bandlimited (blurred) image; a bandlimited and $\times 1/8$ subsampled image. Compare the bandlimited and subsampled image with the naïvely subsampled image. Much of the low-frequency image structure is preserved in the bandlimited and subsampled image. Image source: ImageNet.

Examples are occlusions, causal translations, downscalings of discretized images, and pixel lighting normalization. In each example the transformation deletes information from the original signal, which cannot be regained, and thus it is non-invertible. We extend convolutions to these classes of symmetry under noninvertible transformations via the theory of semigroups. Our contributions are the introduction of a semigroup equivariant correlation and a scale-equivariant CNN.

2 Background

This section introduces some key concepts such as groups, semigroups, actions, equivariance, group convolution, and scale-spaces. These concepts are presented for the benefit of the reader, who is not expected to have a deep knowledge of any of these topics *a priori*.

Downsizing images We consider sampled images $f \in L^2(\mathbb{Z}^2)$ such as in Figure 1. For image f , x is pixel position and $f(x)$ is pixel intensity. If we wish to downsize by a factor of 8, a naïve approach would be to subsample every 8th pixel: $f_{\text{down}}(x) = f(8x)$. This leads to an artifact, *aliasing* [Mallat, 2009, p.43], where the subsampled image contains information at a higher-frequency than can be represented by its resolution. The fix is to *bandlimit* pre-subsampling, suppressing high-frequencies with a blur. Thus a better model for downsampling is $f_{\text{down}}(x) = [G *_{\mathbb{Z}^d} f](8x)$, where $*_{\mathbb{Z}^d}$ denotes convolution over \mathbb{Z}^d , and G is an appropriate blur kernel (discussed later). Downsizing involves necessary information loss and cannot be inverted [Lindeberg, 1997]. Thus upsampling of images is not well-defined, since it involves imputing high-frequency information, not present in the low resolution image. As such, *in this paper we only consider image downscaling*.

Scale-spaces Scale-spaces have a long history, dating back to the late fifties with the work of Iijima [1959]. They consist of an image $f_0 \in L^2(\mathbb{R}^d)$ and multiple blurred versions of it. Although sampled images live on \mathbb{Z}^d , scale-space analysis tends to be over \mathbb{R}^d , but many of the results we present are valid on both domains. Among all variants, the Gaussian scale-space (GSS) is the commonest [Witkin, 1983]. Given an initial image f_0 , we construct a GSS by convolving f_0 with an isotropic (rotation invariant) Gauss-Weierstrass kernel $G(x, t) = (4\pi t)^{-d/2} \exp\{\|x\|^2/4t\}$ of variable width \sqrt{t} and spatial positions x . The GSS is the complete set of responses $f(t, x)$:

$$f(t, x) = [G(\cdot, t) *_{\mathbb{R}^d} f_0](x), \quad t > 0 \quad (1)$$

$$f(0, x) = f_0(x), \quad (2)$$

where $*_{\mathbb{R}^d}$ denotes convolution over \mathbb{R}^d . The higher the level t (larger blur) in the image, the more high frequency details are removed. An example of a scale-space $f(t, x)$ can be seen in Figure 2. An interesting property of scale-spaces is the *semigroup property* [Florack et al., 1992], sometimes referred to as the *recursivity principle* [Pauwels et al., 1995], which is

$$f(s + t, \cdot) = G(\cdot, s) *_{\mathbb{R}^d} f(t, \cdot) \quad (3)$$

for $s, t > 0$. It says that we can generate a scale-space from other levels of the scale-space, not just from f_0 . Furthermore, since $s, t > 0$ it also says that we cannot generate sharper images from blurry ones, using just a Gaussian convolution. Thus moving to blurrier levels encodes a degree of information loss. This property emanates from the closure of Gaussians under convolution, namely for multidimensional Gaussians with covariance matrices Σ and T

$$G(\cdot, \Sigma + T) = G(\cdot, \Sigma) *_{\mathbb{R}^d} G(\cdot, T). \quad (4)$$

We assume the initial image f_0 has a maximum spatial frequency content—dictated by pixel pitch in discretized images—which we model by assuming the image has already been convolved with a width s_0 Gaussian, which we call the *zero-scale*. Thus an image of bandlimit s in the scale-space is found at GSS slice $f(s - s_0, \cdot)$, which we see from Equation 3. There are many varieties of scale-spaces: the α -scale-spaces [Pauwels et al., 1995], the discrete Gaussian scale-spaces [Lindeberg, 1990], the binomial scale-spaces [Burt, 1981], etc. These all have specific kernels, analogous to the Gaussian, which are closed under convolution (details in supplement).

Slices at level t in the GSS correspond to images downsized by a *dilation factor* $0 \leq a \leq 1$ (1 = no downsizing, 0 = shrinkage to a point). An a -dilated and appropriately bandlimited image $p(a, x)$ is found as (details in supplement)

$$p(a, x) = f(t(a, s_0), a^{-1}x), \quad t(a, s_0) := \frac{s_0}{a^2} - s_0. \quad (5)$$

For clarity, we refer to decreases in the spatial dimensions of an image as *dilation* and increases in the blurriness of an image as *scaling*. For a generalization to anisotropic scaling we replace scalar scale t with matrix T , zero-scale s_0 with covariance matrix Σ_0 , and dilation parameter a with matrix A , so

$$T(A, \Sigma_0) = A^{-1}\Sigma_0A^{-T} - \Sigma_0. \quad (6)$$

Semigroups The semigroup property of Equation 3 is the gateway between classical scale-spaces and the group convolution [Cohen and Welling, 2016a] (see end of section). Semigroups (S, \circ) consist of a non-empty set S and a (binary) composition operator $\circ : S \times S \rightarrow S$. Typically, the composition $s \circ t$ of elements $s, t \in S$ is abbreviated st . For our purposes, these individual elements will represent dilation parameters. For S to be a semigroup, it must satisfy the following two properties

- Closure: $st \in S$ for all $s, t \in S$
- Associativity: $(st)r = s(tr) = str$ for all $s, t, r \in S$

Note that commutativity is not a given, so $st \neq ts$. The family of Gaussian densities under spatial convolution is a semigroup². Semigroups are a generalization of groups, which are used in Cohen and Welling [2016a] to model invertible transformations. For a semigroup to be a group, it must also satisfy the following conditions

- Identity element: there exists an $e \in S$ such that $es = se = s$ for all $s \in S$
- Inverses: for each $s \in S$ there exists a s^{-1} such that $s^{-1}s = ss^{-1} = e$.

Actions Semigroups are useful because we can use them to model transformations, also known as (semigroup) *actions*. Given a semigroup S , with elements $s \in S$ and a domain X , the action $L_s : X \rightarrow X$ is a map, also written $x \mapsto L_s[x]$ for $x \in X$. The defining property of the semigroup action (Equation 7) is that it is associative and closed under composition, essentially inheriting its compositional structure from the semigroup S . This action is in fact called a *left action*³. Notice that for a composed action L_{st} , we first apply L_t then L_s .

$$L_{st}[x] = L_s[L_t[x]]. \quad (7)$$

Actions can also be applied to functions $f : X \rightarrow Y$ by viewing f as a point in a function space. The result of the action is a new function, denoted $L_s[f]$. Since the domain of f is X , we commonly write $L_s[f](x)$. Say the domain is a semigroup S then an example left action⁴ is $L_s[f](x) = f(xs)$. Another example, which we shall use later, is the action $S_{A,z}^{\Sigma_0}$ used to form scale-spaces, namely

$$S_{A,z}^{\Sigma_0}[f_0](x) = [G_A^{\Sigma_0} *_{\mathbb{Z}^d} f_0](A^{-1}x + z), \quad G_A^{\Sigma_0} := G(\cdot, A^{-1}\Sigma_0A^{-T} - \Sigma_0). \quad (8)$$

²For the Gaussians we find the identity element as the limit $\lim_{t \downarrow 0} G(\cdot, t)$, which is a Dirac delta. Note that this element is not strictly in the set of Gaussians, thus the Gaussian family has no identity element.

³A *right action* is $R_{st}[x] = R_t[R_s[x]]$. Notice R_{st} applies R_s then R_t , opposite in order to the left action.

⁴This is indeed a left action because $L_s[L_t[f]](x) = L_t[f](xs) = f(xst) = L_{st}[f](x)$.

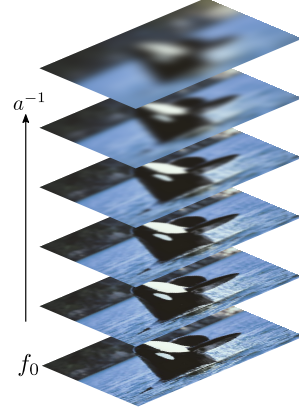


Figure 2: A Scale-space: For implementations we logarithmically discretize the scale-axis in a^{-1} .

The elements of the semigroup are the tuples (A, z) (dilation A , shift z) and $G_A^{\Sigma_0}$ is an anisotropic discrete Gaussian. The action first bandlimits by $G_A^{\Sigma_0} *_{\mathbb{Z}^d} f_0$ then dilates and shifts the domain by $A^{-1}x + z$. Note for fixed (A, z) this maps functions on \mathbb{Z}^d to functions on \mathbb{Z}^d .

Lifting We can also view actions from functions on X to functions X as maps from functions on X to functions on the semigroup S . We call this a *lift* [Kondor and Trivedi, 2018], denoting lifted functions as f^\uparrow . One example is the scale-space action (Equation 8). If we set $x = 0$, then

$$f^\uparrow(A, z) = S_{A,z}^{\Sigma_0}[f_0](0) = [G_A^{\Sigma_0} *_{\mathbb{R}^d} f_0](z), \quad (9)$$

which is the expression for an anisotropic scale-space (parameterized by the dilation A rather than the scale $T(A, \Sigma_0) = A^{-1}\Sigma_0 A^{-T} - \Sigma_0$). To lift a function on to a semigroup, we do not necessarily have to set x equal to a constant, but we could also integrate it out. An important property of *lifted* functions is that actions become simpler. For instance, if we define $f^\uparrow(s) = L_s[f](0)$ then

$$(L_t[f])^\uparrow(s) = L_s[L_t[f]](0) = L_{st}[f](0) = f^\uparrow(st). \quad (10)$$

The action on f could be complicated, like a Gaussian blur, but the action on f^\uparrow is simply a ‘shift’ on $f^\uparrow(s) \mapsto f^\uparrow(st)$. We can then define the action L_t on lifted functions as $L_t[f^\uparrow](s) = f^\uparrow(st)$.

Equivariance and Group Correlations CNNs rely heavily on the (cross-)correlation⁵. Correlations $\star_{\mathbb{Z}^d}$ are a special class of linear map of the form

$$[f \star_{\mathbb{Z}^d} \psi](s) = \sum_{x \in \mathbb{Z}^d} f(x) \psi(x - s). \quad (11)$$

Given a signal f and filter ψ , we interpret the correlation as the collection of inner products of f with all s -translated versions of ψ . This basic correlation has been extended to transformations other than translation via the *group correlation*, which as presented in Cohen and Welling [2016a] are

$$[f \star_H \psi](s) = \sum_{x \in H} f(x) \psi(L_{s^{-1}}[x]), \quad (12)$$

where H is the relevant group and L_s is a *group* action e.g. for rotation $L_s[x] = R_s x$, where R_s is a rotation matrix. Note how the summation domain is H , as are the domains of the signal and filter. Most importantly, the domain of the output is H . It truly highlights how this is an inner product of f and ψ under all s -transformations of the filter. The correlation exhibits a very special property. It is *equivariant* under actions of the group H . This means

$$L_s[f \star_H \psi] = L_s[f] \star_H \psi. \quad (13)$$

That is L_s and \star_H commute—group correlation followed by the action is equivalent to the action followed by the group correlation. Note, the group action may ‘look’ different depending on whether it was applied before or after the group correlation, but it *represents* the exact same transformation.

3 Method

We aim to construct a scale-equivariant convolution. We shall achieve this here by introducing an extension of the correlation to semigroups, which we then tailor to scalings.

Semigroup Correlation There are multiple candidates for a semigroup correlation \star_S . The basic ingredients of such a correlation will be the inner product, the semigroup action L_s , and the functions ψ and f . Furthermore, it must be equivariant to (left) actions on f . For a semigroup S , domain X , and action $L_s : X \rightarrow X$, we define:

$$[\psi \star_S f](s) = \sum_{x \in X} \psi(x) L_s[f](x). \quad (14)$$

It is the set of responses formed from taking the inner product between a filter ψ and a signal f under all transformations of the signal. Notice that we transform the signal and not the filter and that we write $\psi \star_S f$, not $f \star_S \psi$ —it turns out that a similar expression where we apply the action to the filter is not equivariant to actions on the signal. Furthermore this expression *lifts* a function from X to S ,

⁵In the deep learning literature these are inconveniently referred to as convolutions, but we stick to correlation.

so we expect actions on f to look like a ‘shift’ on the semigroup. A proof of equivariance to (left) actions on f is as follows

$$[\psi \star L_t[f]](s) = \sum_{x \in X} \psi(x) L_s[L_t[f]](x) = \sum_{x \in X} \psi(x) L_{st}[f](x) = [\psi \star f](st) = L_t[\psi \star f](s) \quad (15)$$

We have used the definition of the left action $L_s L_t = L_{st}$, the semigroup correlation, and our definition of the action for lifted functions. We can recover the standard ‘convolution’, by substituting $S = \mathbb{Z}^d$, $X = \mathbb{Z}^d$, and the translation action $L_s[f](x) = f(x + s)$:

$$[\psi \star_{\mathbb{Z}^d} f](s) = \sum_{x \in \mathbb{Z}^d} \psi(x) f(x + s) = \sum_{x' \in \mathbb{Z}^d} \psi(x' - s) f(x'). \quad (16)$$

where $x' = x + s$. We can also recover the group correlation by setting $S = H$, $X = H$, where H is a discrete group, and $L_s[f](x) = f(L'_s x)$, where L'_s is the form of L_s acting on the domain $X = H$:

$$[\psi \star_H f](s) = \sum_{x \in H} \psi(x) f(L'_s[x]) = \sum_{x' \in H} \psi(L'_{s^{-1}}[x']) f(x'). \quad (17)$$

where $x' = L'_s[x]$ and since L' is a *group* action inverses exist, so $x = L'_{s^{-1}}[x']$. The semigroup correlation has a two notable differences from the group correlation: i) In the semigroup correlation we transform the signal and not the filter. When we restrict to the group and standard convolution, transforming the signal or the filter are equivalent operations since we can apply a change of variables. This is not possible in the semigroup case, since this change of variables requires an inverse, which we do not necessarily have. ii) In the semigroup correlation, we apply an action to the *whole* signal as $L_s[f]$, as opposed to just the domain ($f(L'_s[x])$). This allows for more general transformations than allowed by the group correlation of Cohen and Welling [2016a], since transformations of the form $f(L'_s[x])$ can only move pixel locations, but transformations of the form $L_s[f]$ can alter the values of the pixels as well, and can incorporate neighbourhood information into the transformation.

The Scale-space Correlation We now have the tools to create a scale-equivariant correlation. All we have to choose is an appropriate action for L_s . We choose the scale-space action of Equation 8. The scale-space action $S_{z,A}^{\Sigma_0}$ for functions on \mathbb{Z}^d is given by

$$S_{A,z}^{\Sigma_0}[f](x) = [G_A^{\Sigma_0} *_{\mathbb{Z}^d} f](A^{-1}x + z), \quad S_{A,z}^{\Sigma_0} S_{B,y}^{\Sigma_0} = S_{AB,Ay+z}^{\Sigma_0}. \quad (18)$$

Since our scale-space correlation only works for discrete semigroups, we have to find a suitable discretization of the dilation parameter A . Later on we will choose a discretization of the form $A_k = 2^{-k}I$ for $k \geq 0$, but for now we will just assume that there exists some countable set \mathcal{A} , such that $S = \{(A, z)_{A \in \mathcal{A}, z \in \mathbb{Z}^d}\}$ is a valid discrete semigroup. We begin by assuming we have lifted an input image f on to the scale-space via Equation 9. The lifted signal is indexed by coordinates A, z and so the filters share this domain and are of the form $\psi(A, z)$. The scale-space correlation is then

$$[\psi \star_S f](A, z) = \sum_{(B,y) \in S} \psi(B, y) S_{B,y}^{\Sigma_0}[f](A, y) = \sum_{B \in \mathcal{A}} \sum_{y \in \mathbb{Z}^d} \psi(B, y) f(BA, A^{-1}y + z). \quad (19)$$

For the second equality we recall the action on a lifted signal is governed by Equation 10. The appealing aspect of this correlation is that we do not need to convolve with a bandlimiting filter—a potentially expensive operation to perform at every layer of a CNN—since we use signals that have been lifted on to the semigroup. Instead, the action of scaling by A is accomplished by ‘fetching’ a slice $f(BA, \cdot)$ from a blurrier level of the scale-space. Let’s restrict the scale correlation to the scale-space where $A_k = 2^{-k}I$ for $k \geq 0$, with zero-scale $\Sigma_0 = \frac{1}{4}I$. Denoting $f(2^{-k}I, \cdot)$ as $f_k(\cdot)$, this can be seen as a dilated convolution [Yu and Koltun, 2015] between ψ_ℓ and slice $f_{\ell+k}$. This form of the scale-space correlation (shown below) we use in our experiments. A diagram of this can be seen in Figure 3:

$$[\psi \star_S f]_k(z) = \sum_{\ell \geq 0} \sum_{y \in \mathbb{Z}^d} \psi_\ell(y) f_{\ell+k}(2^k y + z). \quad (20)$$

Equivariant Nonlinearities Not all nonlinearities commute with semigroup actions, but it turns out that pointwise nonlinearities ν commute with a special subset of actions of the form,

$$L_s[f](x) = f(L'_s[x]) \quad (21)$$

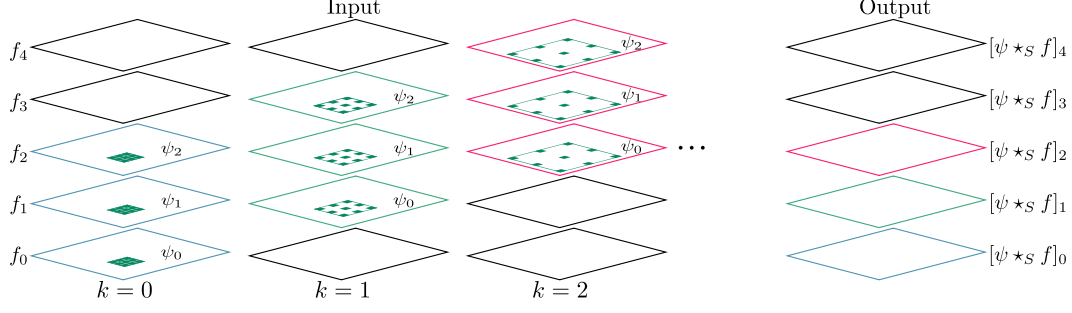


Figure 3: Scale correlation schematic: The left 3 stacks are the same input f , with levels $f_\ell(\cdot) = f(2^{-\ell}I, \cdot)$. Each stack shows the inner product between filter ψ (in green) at translation z for dilation $2^{-k}I$ corresponding to the output level $[\psi \star_S f]_k$ on the right with matching color. Notice that as we dilate the filter, we also shift it one level up in the scale-space, according to Equation 20.

where L'_s is a representation of L_s acting on the domain of f . For these sorts of actions, we cannot alter the values of the f , just the locations of the values. If we write function composition as $[\nu \bullet f](x) = \nu(f(x))$, then a proof of equivariance is as follows:

$$[\nu \bullet L_s[f]](x) = \nu(f(L'_s[x])) = [\nu \bullet f](L'_s[x]) = L_s[\nu \bullet f](x). \quad (22)$$

Equation 21 may at first glance seem overly restrictive, but it turns out that this is not the case. Recall that for functions lifted on to the semigroup, the action is $L_s[f](x) = f(xs)$. This satisfies Equation 21, and so we are free to use pointwise nonlinearities.

Batch normalization For batch normalization, we compute batch statistics over all dimensions of an activation tensor except its channels, as in Cohen and Welling [2016a].

Initialization Since our correlations are based on dilated convolutions, we use the initialization scheme presented in [Yu and Koltun, 2015]. For pairs of input and output channels the center pixel of each filter is set to one and the rest are filled with random noise of standard deviation 10^{-2} .

Boundary conditions In our semigroup correlation, the scale dimension is infinite—this is a problem for practical implementations. Our solution is to truncate the scale-space to finite scale. This breaks global equivariance to actions in S , but is locally correct. Boundary effects occur at activations with receptive fields covering the truncation boundary. To mitigate these effects we: i) use filters with a scale-dimension no larger than two, ii) interleave filters with scale-dimension 2 with filters of scale-dimension 1. The scale-dimension 2 filters enable multiscale interactions but they propagate boundary effects; whereas, the scale-dimension 1 kernels have no boundary overlap, but also no multiscale behavior. Interleaving trades off network expressiveness against boundary effects.

Scale-space implementation We use a 4 layer scale-space and zero-scale $1/4$ with dilations at integer powers of 2, the maximum dilation is 8 and kernel width 33 (4 std. of a discrete Gaussian). We use the discrete Gaussian of Lindeberg [1990]. In 1D for scale parameter t , this is

$$G(x, t) = e^{-t} I_{|x|}(t) \quad (23)$$

where $I_x(t)$ is the modified Bessel function of integer order. For speed, we make use of the separability of isotropic kernels. For instance, convolution with a 2D Gaussian can be written as a convolution with 2 identical 1D Gaussians sequentially along the x and then the y -axis. For an $N \times N$ input image and $M \times M$ blurring kernel, this reduces computational complexity of the convolution as $O(M^2 N^2) \rightarrow O(2MN^2)$. With GPU parallelization, this saving is $O(M^2) \rightarrow O(M)$, which is especially significant for us since the largest blur kernel we use has $M = 33$.

Multi-channel features Typically CNNs use multiple channels per activation tensor, which we have not included in our above treatment. In our experiment we include input channels i and output channels o , so a correlation layer is

$$[\psi \star_S f]_k^o(z) = \sum_i \sum_{\ell \geq 0} \sum_{x \in \mathbb{Z}^d} \psi_\ell^{i,o}(x) f_{\ell+k}^i(2^k x + z). \quad (24)$$

Table 1: Results on the Patch Camelyon and Cityscapes Dataset. Higher is better

PCam Model	Accuracy	Cityscapes Model	mAP
DenseNet Baseline	87.0	ResNet, matched parameters	45.66
S-DenseNet (Ours)	88.1	ResNet, matched channels	49.99
[Veeling et al., 2018]	89.8	S-ResNet, multiscale (Ours)	63.53
		S-ResNet, no interaction (Ours)	64.78

4 Experiments and Results

Here we present our results on the Patch Camelyon [Veeling et al., 2018] and Cityscapes [Cordts et al., 2016] datasets. We also visualize the quality of scale-equivariance achieved.

Patch Camelyon The Patch Camelyon or PCam dataset [Veeling et al., 2018] contains 327 680 tiles from two classes, metastatic (tumorous) and non-metastatic tissue. Each tile is a 96×96 px RGB-crop labelled as metastatic if there is at least one pixel of metastatic tissue in the central 32×32 px region of the tile. We test a 4-scale DenseNet model [Huang et al., 2017], ‘S-DenseNet’, on this task (architecture in supplement). We also train a scale non-equivariant DenseNet baseline and the rotation equivariant model of [Veeling et al., 2018]. Our training procedure is: 100 epochs SGD, learning rate 0.1 divided by 10 every 40 epochs, momentum 0.9, batch size of 512, split over 4 GPUs. For data augmentation, we follow the procedure of [Veeling et al., 2018, Liu et al., 2017], using random flips, rotation, and 8 px jitter. For color perturbations we use: brightness delta 64/255, saturation delta 0.25, hue delta 0.04, contrast delta 0.75. The evaluation metric we test on is accuracy. The results in Table 1 show that both the scale and rotation equivariant models outperform the baseline.

Cityscapes The Cityscapes dataset [Cordts et al., 2016] contains 2975 training images, 500 validation images, and 1525 test images of resolution 2048×1024 px. The task is semantic segmentation into 19 classes. We train a 4-scale ResNet He et al. [2016], ‘S-ResNet’, and baseline. We train an equivariant network with and without multiscale interaction layers. We also train two scale non-equivariant models, one with the same number of channels, one with the same number of parameters. Our training procedure is: 100 epochs Adam, learning rate 10^{-3} divided by 10 every 40 epochs, batch size 8, split over 4 GPUs. The results are in Table 1. The evaluation metric is mean average precision. We see that our scale-equivariant model outperforms the baselines. We must caution however, that better competing results can be found in the literature. The reason our baseline underperforms compared to the literature is because of the parameter/channel-matching, which have shrunk its size somewhat due to our own resource constraints. On a like-for-like comparison scale-equivariance appears to help.

Quality of Equivariance We validate the quality of equivariance empirically by comparing activations of a dilated image against the theoretical action on the activations. Using Φ to denote the deep scale-space (DSS) mapping, we compute the normalized L_2 -distance at each level k of a DSS. Mathematically this is

$$L(2^{-\ell}, k) = \frac{\|\Phi[f](k + \ell, 2^\ell \cdot) - \Phi[S_{2^{-\ell}, 0}^{1/4I}[f]](k, \cdot)\|_2}{\|\Phi[f](k + \ell, 2^\ell \cdot)\|_2}. \quad (25)$$

The equivariance errors are in Figure 4 for 3 DSSs with random weights and a scale-space truncated to 8 scales. We see that the average error is below 0.01, indicating that the network is mostly equivariant, with errors due to truncation of the discrete Gaussian kernels used to lift the input to scale-space. We also see that the equivariance errors blow up whenever for constant $\ell + k$ in each graph. This is the point where the receptive field of an activation overlaps with the scale-space truncation boundary.

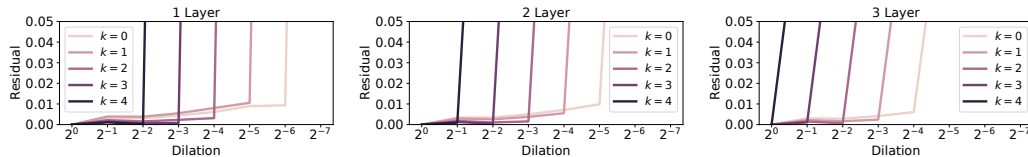


Figure 4: Equivariance quality. Left to right: 1, 2, and 3 layer DSSs. Each line represents the error as in Equation 25. We see the residual error is typically < 0.01 until boundary effects are present.

5 Related Work

In recent years, there have been a number of works on group convolutions, namely continuous roto-translation in 2D [Worrall et al., 2017] and 3D [Weiler et al., 2018a, Kondor et al., 2018, Thomas et al., 2018] and discrete roto-translations in 2D [Cohen and Welling, 2016a, Weiler et al., 2018b, Bekkers et al., 2018, Hoogetboom et al., 2018] and 3D [Worrall et al., 2017], continuous rotations on the sphere [Esteves et al., 2018, Cohen et al., 2018b], in-plane reflections [Cohen and Welling, 2016a], and even reverse-conjugate symmetry in DNA sequencing [Lunter and Brown, 2018]. Theory for convolutions on compact groups—used to model invertible transformations—also exists [Cohen and Welling, 2016a,b, Kondor and Trivedi, 2018, Cohen et al., 2018a], but to date the vast majority of work has focused on rotations.

For scale, there far fewer works. [Henriques and Vedaldi, 2017, Esteves et al., 2017] both perform a log-polar transform of the signal before passing to a standard CNN. Log-polar transforms reparameterize the input plane into angle and log-distance from a predefined origin. The transform is sensitive to origin positioning, which if done poorly breaks translational equivariance. [Marcos et al., 2018] use a group CNN architecture designed for roto-translation, but instead of rotating filters in the group correlation, they scale them. This seems to work on small tasks, but ignores large scale variations. [Hilbert et al., 2018] instead use filters of different sizes, but without any attention to equivariance.

6 Discussion, Limitations, and Future Works

We found the best performing architectures were composed mainly of correlations where the filters’ scale dimension is one, interleaved with correlations where the scale dimension is higher. This is similar to a network-in-network [Lin et al., 2013] architecture, where 3×3 convolutional layers are interleaved with 1×1 convolutions. We posit this is the case because of boundary effects, as were observed in Figure 4. We suspect working with dilations with smaller jumps in scale between levels of the scale-space would help, so $A_k = c^{-k}I$ for $1 < c < 2$. This is reminiscent of the tendency in the image processing community to use non-integer scales. This would, however, involve the development of non-integer dilations and hence interpolation. We see working on mitigating boundary effects as an important future work, not just for scale-equivariance, but CNNs as a whole.

Another limitation of the current model is the increase in computational overhead, since we have added an extra dimension to the activations. This may not be a problem long term, as GPUs grow in speed and memory, but the computational complexity of a correlation grows exponentially in the number of symmetries of the model, and so we need more efficient methods to perform correlation, either exactly or approximately.

We see semigroup correlations as an exciting new family of operators to use in deep learning. We have demonstrated a proof-of-concept on scale, but there are many semigroup-structured transformations left to be explored, such as causal shifts, occlusions, and affine transformations. Concerning scale, we are also keen to explore how multi-scale interactions can be applied on other domains as meshes and graphs, where symmetry is less well-defined.

7 Conclusion

We have presented deep scale-spaces a generalization of convolutional neural networks, exploiting the scale symmetry structure of conventional image recognition tasks. We outlined new theory for a generalization of the convolution operator on semigroups, the semigroup correlation. Then we showed how to derive the standard convolution and the group convolution [Cohen and Welling, 2016a] from the semigroup correlation. We then tailored the semigroup correlation to the scale-translation action used in classical scale-space theory and demonstrated how to use this in modern neural architectures.

References

- E. Barnard and D. Casasent. Invariance and neural nets. *IEEE Transactions on Neural Networks*, 2(5):498–508, Sep. 1991. ISSN 1045-9227. doi: 10.1109/72.134287.
- Erik J. Bekkers, Maxime W. Lafarge, Mitko Veta, Koen A. J. Eppenhof, Josien P. W. Pluim, and Remco Duits. Roto-translation covariant convolutional networks for medical image analysis. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pages 440–448, 2018. doi: 10.1007/978-3-030-00928-1_50.
- Bernhard Burgeth, Stephan Didas, and Joachim Weickert. The besel scale-space. In *Deep Structure, Singularities, and Computer Vision, First International Workshop, DSSCV 2005, Maastricht, The Netherlands, June 9-10, 2005, Revised Selected Papers*, pages 84–95, 2005a. doi: 10.1007/11577812_8. URL https://doi.org/10.1007/11577812_8.
- Bernhard Burgeth, Stephan Didas, and Joachim Weickert. Relativistic scale-spaces. In *Scale Space and PDE Methods in Computer Vision, 5th International Conference, Scale-Space 2005, Hofgeismar, Germany, April 7-9, 2005, Proceedings*, pages 1–12, 2005b. doi: 10.1007/11408031_1. URL https://doi.org/10.1007/11408031_1.
- Peter J Burt. Fast filter transform for image processing. *Computer graphics and image processing*, 16(1):20–51, 1981.
- Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2018. doi: 10.1109/TPAMI.2017.2699184.
- Taco Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 2990–2999, 2016a.
- Taco Cohen, Mario Geiger, and Maurice Weiler. A general theory of equivariant cnns on homogeneous spaces. *CoRR*, abs/1811.02017, 2018a.
- Taco S. Cohen and Max Welling. Steerable cnns. *CoRR*, abs/1612.08498, 2016b.
- Taco S. Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical cnns. *CoRR*, abs/1801.10130, 2018b.
- Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 3213–3223, 2016. doi: 10.1109/CVPR.2016.350.
- James L Crowley, Olivier Riff, and Justus H Piater. Fast computation of characteristic scale using a half octave pyramid. In *International Conference on Scale-Space Theories in Computer Vision*, 2002.
- Remco Duits and Bernhard Burgeth. Scale spaces on lie groups. In *Scale Space and Variational Methods in Computer Vision, First International Conference, SSVM 2007, Ischia, Italy, May 30 - June 2, 2007, Proceedings*, pages 300–312, 2007. doi: 10.1007/978-3-540-72823-8_26. URL https://doi.org/10.1007/978-3-540-72823-8_26.
- Remco Duits, Michael Felsberg, Luc Florack, and Bram Platel. alpha scale spaces on a bounded domain. In *Scale Space Methods in Computer Vision, 4th International Conference, Scale-Space 2003, Isle of Skye, UK, June 10-12, 2003, Proceedings*, pages 494–510, 2003. doi: 10.1007/3-540-44935-3_34. URL https://doi.org/10.1007/3-540-44935-3_34.
- Remco Duits, Luc Florack, Jan de Graaf, and Bart M. ter Haar Romeny. On the axioms of scale space theory. *Journal of Mathematical Imaging and Vision*, 20(3):267–298, 2004. doi: 10.1023/B:JMIV.0000024043.96722.aa. URL <https://doi.org/10.1023/B:JMIV.0000024043.96722.aa>.

- Carlos Esteves, Christine Allen-Blanchette, Xiaowei Zhou, and Kostas Daniilidis. Polar transformer networks. *CoRR*, abs/1709.01889, 2017.
- Carlos Esteves, Christine Allen-Blanchette, Ameesh Makadia, and Kostas Daniilidis. Learning SO(3) equivariant representations with spherical cnns. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII*, pages 54–70, 2018. doi: 10.1007/978-3-030-01261-8_4.
- Luc Florack, Bart M. ter Haar Romeny, Jan J. Koenderink, and Max A. Viergever. Scale and the differential structure of images. *Image Vision Comput.*, 10(6):376–388, 1992. doi: 10.1016/0262-8856(92)90024-W.
- Luc Florack, Bart M. ter Haar Romeny, Jan J. Koenderink, and Max A. Viergever. Linear scale-space. *Journal of Mathematical Imaging and Vision*, 4(4):325–351, 1994. doi: 10.1007/BF01262401. URL <https://doi.org/10.1007/BF01262401>.
- Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 580–587, 2014. doi: 10.1109/CVPR.2014.81.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.
- João F. Henriques and Andrea Vedaldi. Warped convolutions: Efficient invariance to spatial transformations. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 1461–1469, 2017.
- Adam Hilbert, Bastiaan Veeling, and Henk Marquering. Data-efficient convolutional neural networks for treatment decision support in acute ischemic stroke. *International conference on Medical Imaging with Deep Learning*, 2018.
- Emiel Hoogeboom, Jorn W. T. Peters, Taco S. Cohen, and Max Welling. Hexaconv. *CoRR*, abs/1803.02108, 2018.
- Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2261–2269, 2017. doi: 10.1109/CVPR.2017.243.
- Taizo Iijima. Basic theory of pattern observation. *Technical Group on Automata and Automatic Control*, pages 3–32, 1959.
- Iasonas Kokkinos. Pushing the boundaries of boundary detection using deep learning. *International Conference on Learning Representations (ICLR)*, 2015.
- Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 2752–2760, 2018.
- Risi Kondor, Zhen Lin, and Shubhendu Trivedi. Clebsch-gordan nets: a fully fourier space spherical convolutional neural network. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 10138–10147, 2018.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, pages 1106–1114, 2012.

- Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998. ISSN 0018-9219. doi: 10.1109/5.726791.
- Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *CoRR*, abs/1312.4400, 2013.
- Tony Lindeberg. Scale-space for discrete signals. *IEEE Trans. Pattern Anal. Mach. Intell.*, 12(3): 234–254, 1990. doi: 10.1109/34.49051.
- Tony Lindeberg. On the axiomatic foundations of linear scale-space. In *Gaussian Scale-Space Theory*, pages 75–97, 1997. doi: 10.1007/978-94-015-8802-7_6.
- Yun Liu, Krishna Gadepalli, Mohammad Norouzi, George E. Dahl, Timo Kohlberger, Aleksey Boyko, Subhashini Venugopalan, Aleksei Timofeev, Philip Q. Nelson, Gregory S. Corrado, Jason D. Hipp, Lily Peng, and Martin C. Stumpe. Detecting cancer metastases on gigapixel pathology images. *CoRR*, abs/1703.02442, 2017.
- Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3431–3440, 2015. doi: 10.1109/CVPR.2015.7298965.
- Gerton Lunter and Richard Brown. An equivariant bayesian convolutional network predicts recombination hotspots and accurately resolves binding motifs. *bioRxiv*, 2018. doi: 10.1101/351254.
- Stéphane Mallat. *A Wavelet Tour of Signal Processing - The Sparse Way, 3rd Edition*. Academic Press, 2009. ISBN 978-0-12-374370-1.
- Diego Marcos, Benjamin Kellenberger, Sylvain Lobry, and Devis Tuia. Scale equivariance in cnns with vector fields. *CoRR*, abs/1807.11783, 2018.
- Eric J. Pauwels, Luc J. Van Gool, Peter Fiddelaers, and Theo Moons. An extended class of scale-invariant and recursive scale space filters. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17(7):691–701, 1995. doi: 10.1109/34.391411.
- Alfons H. Salden, Bart M. ter Haar Romeny, and Max A. Viergever. Linear scale-space theory from physical principles. *Journal of Mathematical Imaging and Vision*, 9(2):103–139, 1998. doi: 10.1023/A:1008300826001. URL <https://doi.org/10.1023/A:1008300826001>.
- Jure Sokolic, Raja Giryes, Guillermo Sapiro, and Miguel R. D. Rodrigues. Generalization error of invariant classifiers. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, pages 1094–1103, 2017.
- Nathaniel Thomas, Tess Smidt, Steven M. Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation- and translation-equivariant neural networks for 3d point clouds. *CoRR*, abs/1802.08219, 2018.
- Bastiaan S. Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II*, pages 210–218, 2018. doi: 10.1007/978-3-030-00934-2_24.
- Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco Cohen. 3d steerable cnns: Learning rotationally equivariant features in volumetric data. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 10402–10413, 2018a.
- Maurice Weiler, Fred A. Hamprecht, and Martin Storath. Learning steerable filters for rotation equivariant cnns. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 849–858, 2018b.
- Andrew P. Witkin. Scale-space filtering. In *Proceedings of the 8th International Joint Conference on Artificial Intelligence. Karlsruhe, FRG, August 1983*, pages 1019–1022, 1983.

Daniel E. Worrall, Stephan J. Garbin, Daniyar Turmukhambetov, and Gabriel J. Brostow. Harmonic networks: Deep translation and rotation equivariance. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 7168–7177, 2017. doi: 10.1109/CVPR.2017.758.

Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *CoRR*, abs/1511.07122, 2015.

Fisher Yu, Vladlen Koltun, and Thomas A. Funkhouser. Dilated residual networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 636–644, 2017. doi: 10.1109/CVPR.2017.75.

Supplementary Material

In this supplementary material we elaborate on where the scale-space action comes from and the architectures used in our paper.

A Scale-spaces

Here we provide some extra information on scale-spaces for the interested reader. For in depth literature, we suggest Florack et al. [1994, 1992], Pauwels et al. [1995], Lindeberg [1990, 1997], Crowley et al. [2002], Salden et al. [1998], Duits et al. [2004, 2003], Duits and Burgeth [2007], Burgeth et al. [2005a,b].

A.1 1D Gaussian Scale-space

We are given an initial 1D signal f_0 with intrinsic bandlimit, or *zero-scale*, defined by s_0 i.e., we impose a maximum frequency, such that there is a correspondence with discretized signals. Note that $\sqrt{s_0}$ is inversely proportional to frequency content of the signal. We wish to downsize it isotropically by a factor a , which we call the *dilation*. For this we introduce the downsizing action $L_a[f](x) = f(a^{-1}x)$ for $a \leq 1$. We model the bandlimit of the initial signal as the result of convolving some other signal f with a Gaussian of width s_0 , so

$$f_0(x) = [G(\cdot, s_0) *_{\mathbb{R}^d} f](x). \quad (26)$$

Now the result of the downscaling action of f_0 is as follows

$$L_a[f_0](x) = f_0(a^{-1}x) = \int_{\mathbb{R}^d} G(a^{-1}x - y, s_0) f(y) dy = \int_{\mathbb{R}^d} G(x - ay, a^2 s_0) f(y) a dy \quad (27)$$

$$= \int_{\mathbb{R}^d} G(x - z, a^2 s_0) f(a^{-1}z) dz = [G(\cdot, a^2 s_0) *_{\mathbb{R}^d} L_a[f]](x). \quad (28)$$

From the first to second lines we have performed a change of variables $z = ay$. So we see that the effect of downsizing a bandlimited signal by a shifts the bandlimit from s_0 to $a^2 s_0$. Since $a \leq 1$, this means the blurring Gaussian is narrower and so the frequency content of the signal has been shifted higher. The key relation to bear in mind is the shift $s_0 \mapsto a^2 s_0$.

For a proper scaling, we want the result of the downscaling to have the *same bandlimit* as the original signal f_0 . This is because if we are representing the signal on a discrete grid, then we have a physically defined maximum frequency content we can store, given by the pixel separation. The solution is to convolve the signal $L_a[f_0]$ with a correcting Gaussian of width $s_0 - a^2 s_0$. Note that this is possible, since $a \leq 1$, so $s_0 - a^2 s_0 > 0$. Alternatively, we want to find a correcting Gaussian to blur before downsizing. Say this correcting Gaussian has bandlimit t , then we have

$$L_a[G(\cdot, t) * f_0] = L_a[G(\cdot, t + s_0) * f] = [G(\cdot, a^2(t + s_0)) * L_a[f]]. \quad (29)$$

But we also want the bandlimit of the downsampled signal to be s_0 , so we have the relation

$$a^2(t + s_0) = s_0. \quad (30)$$

Thus a 1D Gaussian scale-space, parameterized by dilation, can be built by setting

$$f(t(a, s_0), x) = [G(\cdot, t(a, s_0)) * f_0](x), \quad t(a, s_0) := \frac{s_0}{a^2} - s_0 \quad (31)$$

$$f(0, x) = f_0(x) \quad (32)$$

A.2 ND Gaussian Scale-space

We now explore downscaling in N -dimensions. We first of all represent an initial image f_0 as

$$f_0 = G(\cdot, \Sigma_0) *_{\mathbb{R}^d} f. \quad (33)$$

We use the Gaussian to represent the fact that f_0 should have an intrinsic bandlimit, usually defined by the resolution at which it is sampled. Now let's introduce the affine action:

$$L_{A,z}[f](x) = f(A^{-1}(x - z)). \quad (34)$$

It simply applies an affine transformation to our signal. Now using a similar logic to in the 1D case, if we concatenate the affine action with bandlimiting we get

$$L_{A,z}[G(\cdot, \Sigma_0) *_{\mathbb{R}^d} f_0] = G(\cdot, A\Sigma_0 A^\top) *_{\mathbb{R}^d} L_{A,z}[f_0]. \quad (35)$$

So we see that resizing a bandlimited signal shifts the bandlimit according to $\Sigma_0 \mapsto A\Sigma_0 A^\top$. Since we would like to have the same bandlimit on our signal before and after the resizing (since we can only represent the signal at constant resolution), we introduce a second bandlimiting by convolving with a Gaussian of width $\Sigma_0 - A\Sigma_0 A^\top$. To save space, we write $G_\Sigma = G(\cdot, \Sigma)$. So

$$G_{\Sigma_0} *_{\mathbb{R}^d} L_{A,z}[f] = G_{\Sigma_0 - A\Sigma_0 A^\top} *_{\mathbb{R}^d} L_{A,z}[f_0] = L_{A,z}[G_{A^{-1}\Sigma_0 A^{-\top} - \Sigma_0} *_{\mathbb{R}^d} f_0] \quad (36)$$

From the first to the second equality, we have exchanged the order of the Gaussian convolution and the affine action and altered the bandwidth from $\Sigma_0 - A\Sigma_0 A^\top$ to $A^{-1}\Sigma_0 A^{-\top} - \Sigma_0$, which comes from the relation established in Equation 35. Thus the affine action for an affine scale-pyramid is defined as

$$L_{A,z}[G_{A^{-1}\Sigma_0 A^{-\top} - \Sigma_0} *_{\mathbb{R}^d} f](x) = [G_A^{\Sigma_0} *_{\mathbb{R}^d} f](A^{-1}(x - z)) \quad (37)$$

where we have defined $G_A^{\Sigma_0} = G_{A^{-1}\Sigma_0 A^{-\top} - \Sigma_0}$.

For what values of A and z is this action valid? Let's first focus on A . To maintain the zero-scale of Σ_0 , we had to convolve with a Gaussian of width $\Delta = \Sigma_0 - A\Sigma_0 A^\top$. Now we know that covariance matrices have to be symmetric $\Delta = \Delta^\top$ and positive definite $\Delta \succ 0$. We see already that it is symmetric, but it is not necessary positive definite. If the base bandlimit is of the form $\Sigma_0 = \sigma_0^2 I$ (the original image is isotropically bandlimited), then we can rearrange to

$$\Delta = \Sigma_0 - A\Sigma_0 A^\top = (I - AA^\top)\Sigma_0 \succ 0 \quad (38)$$

This expression is only positive definite if $I - AA^\top \succ 0$; that is

$$I \succ AA^\top. \quad (39)$$

This condition implies that A^\top is a contraction because

$$v^\top (I - AA^\top) v = \|v\|_2^2 - \|A^\top v\|_2^2 \geq 0 \implies \|v\|_2^2 \geq \|A^\top v\|_2^2, \quad \forall v \in \mathbb{R}^N. \quad (40)$$

Another way of phrasing this is that the singular values of A may not exceed unity. Note that rotations do not break this constraint. So we see this model naturally aligns with our notion that we can only model image downscalings, and that upscalings are prohibited.

A.3 Other Scale-space variants

We have presented the Gaussian scale-space in ND , but there also exists a zoo of other scales-spaces. The most prominent are: the α -scale-spaces [Pauwels et al., 1995], the discrete Gaussian scale-spaces [Lindeberg, 1990], and the binomial scale-spaces [Burt, 1981]. In the following, we give a brief introduction to each, exhibited in 1D.

α -scale-spaces α -scale-spaces Pauwels et al. [1995] are a generalization of the Gaussian space-space in the continuous domain. They are easiest to understand by considering their form in Fourier-space. We begin by considering the Fourier transform of the Gaussian space-space over the spatial dimension

$$\hat{f}(t, \omega) = \hat{G}(\omega, t) \cdot \hat{f}_0(\omega) \quad (41)$$

$$\hat{f}(0, \omega) = \hat{f}_0(\omega), \quad (42)$$

where \hat{f} is the Fourier transform of f . We are interested in finding a collection of filters like G , closed under convolution. In the Fourier domain this corresponds to finding a collection of filters, like \hat{G} closed under multiplication. The Fourier transform of the Gauss-Weierstrass kernel is

$$G(x, t) = \frac{1}{(4\pi t)^{1/2}} \exp\left\{-\frac{x^2}{4t}\right\} \xleftrightarrow{\text{FT}} \hat{G}(\omega, t) = \exp\{-\omega^2 t\}. \quad (43)$$

The collection $\{\hat{G}(\omega, t)\}_{t \geq 0}$ is indeed closed under multiplication and forms a semigroup. To form the α -scale-spaces we notice that the Fourier kernel

$$\hat{G}^\alpha(\omega, t) = \exp\{-\omega^{2\alpha} t\} \quad (44)$$

is also closed under multiplication and defines a semigroup. The range of α is typically taken to be $(0, 1]$, to make sure that higher levels in the α -scale-space are blurrier. Notice that for $\alpha = 1$ we return to the standard Gaussian scale-space.

Binomial Scale-space The binomial scale-space Crowley et al. [2002] is a discrete scale-space in both the spatial and scale dimensions. It is generated by convolution in \mathbb{Z} with the binomial kernel

$$B(x, N) = {}^N C_x \bigg/ \sum_{x=0}^N {}^N C_x, \quad {}^N C_x = \frac{N!}{(N-x)!x!}, \quad (45)$$

where $N > 0$ is the width of the kernel and $0 \leq x \leq N$ is the spatial location of the filter tap. Thus the scale-space is

$$f(N, x) = [B(\cdot, N) *_{\mathbb{Z}} f_0](x) \quad N > 0. \quad (46)$$

As N grows $B(N, x)$ rapidly converges to a Gaussian kernel of variance $\sigma^2 = N/4$. The Binomial filters are closed under convolution obeying the semigroup property

$$[B(\cdot, N) *_{\mathbb{Z}} B(\cdot, M)](x) = B(x, N + M - 1). \quad (47)$$

Discrete Gaussian Scale-space The discrete Gaussian scale-space Lindeberg [1990] is discrete in the spatial dimension but continuous in the scale dimension, which makes it popular to work with in many practical scale-spaces with non-integer dilation. The scale-space is generated by convolution in \mathbb{Z} with the discrete Gaussian kernel

$$G(x, t) = e^{-t} I_{|x|}(t), \quad I_x(t) = \sum_{m=0}^{\infty} \frac{\left(\frac{1}{2}x\right)^{2m+\alpha}}{m! \Gamma(m + \alpha + 1)} \quad (48)$$

where the term $I_k(t)$ is a modified Bessel function of the first kind. These can be implemented easily using `scipy.special.i0e`. The scale-space is formed in the usual way as

$$f(t, x) = [G(\cdot, t) *_{\mathbb{Z}} f_0](x) \quad t > 0 \quad (49)$$

$$f(0, x) = f_0(x). \quad (50)$$

B Architectures In The Experiments

In the experiments, we use a DenseNet Huang et al. [2017] and a ResNet He et al. [2016]. The architectures are as follows. For the scale equivariant versions, we use 4 scales of a discrete Gaussian scale-space Lindeberg [1990].

ResNet The residual network consists of a concatenation of residual blocks. A single residual block implements the following

$$y = x + \mathcal{F}(x) \quad (51)$$

where on the RHS we refer to x as the skip connection and $\mathcal{F}(x)$ as the residual connection. If x has fewer channels than $\mathcal{F}(x)$, then we pad the missing dimensions with zeros. Each residual connection uses a concatenation of two scale-equivariant correlation interleaved with batch normalization (BN) and a ReLU (ReLU) nonlinearity. These are composed as follows (input left, output right)

$$\text{corr}[1, 3, 3] - \text{BN} - \text{ReLU} - \text{corr}[k, 3, 3] - \text{BN}. \quad (52)$$

where $\text{corr}[k, h, w]$ refers to a scale correlation with kernel size $[k, h, w]$ and where k is the number of scale channels, h is the spatial height of the filter, and w is its spatial width. We denote the entire residual block as $\text{res}[k, h, w]$.

The model we use is given in Table 2. It follows the practice of Yu et al. [2017], who use a bilinear upsampling at the end of the network, since segmentations do not tend to contain high frequency details. In our experiments we use the models shown in Table 3

DenseNet The Dense network Huang et al. [2017] consists of a concatenation of 3 dense blocks. Each dense block is composed of layers of the form

$$y_{N+1} = \mathcal{H}([y_1, y_2, \dots, y_N]) \quad (53)$$

Table 2: A residual network. Input at the top. A horizontal line denotes spatial average pooling of stride 2, kernel size 2. Shape is displayed as [scale-space levels, height, width, channels out]. The no-res block denotes a residual block without the skip connection, i.e. $y = \mathcal{F}(x)$. Scale pooling denotes an averaging over all scale dimensions.

Layer type	Shape
res[$k, 3, 3$]	$[S, 992, 992, N]$
res[$k, 3, 3$]	$[S, 496, 496, 2N]$
res[$1, 3, 3$]	$[S, 248, 248, 4N]$
res[$k, 3, 3$]	$[S, 248, 248, 4N]$
res[$1, 3, 3$]	$[S, 124, 124, 8N]$
res[$1, 3, 3$]	$[S, 124, 124, 8N]$
res[$1, 3, 3$]	$[S, 124, 124, 8N]$
res[$k, 3, 3$]	$[S, 124, 124, 8N]$
no-res[$k, 3, 3$]	$[S, 124, 124, 8N]$
scale-pool	$[1, 124, 124, 8N]$
corr[$1, 1, 1$],	$[1, 124, 124, 19]$
bilinear upsample	$[1, 992, 992, 19]$

where $[y_1, y_2, \dots, y_N]$ is the concatenation of all the previous layers' outputs. Each layer \mathcal{H} is the composition (input left, output right)

$$\text{BN} - \text{ReLU} - \text{corr}[k, 3, 3] \quad (54)$$

where $\text{corr}[k, 3, 3]$ was described in the previous section. We use the notation $\text{dense}_C[k, h, w] \times N$ to denote a dense block with N layers and C output channels per layer. The number of channel outputs remains constant per layer within a dense block. Between dense blocks, we insert transition layer which have the form

$$\text{dense}[1, 1, 1] \times 1 - \text{pool} - \text{dense}[k, h, w] \times 1. \quad (55)$$

Here we use a 1×1 convolution to halve the number of output channels, and then perform a spatial average pooling with kernel size 2 and stride 2, followed by a second dense layer. We denote these as $\text{transition}[k, h, w]$. We also use long skip connection between transition layers. The network we use is shown in Table 4.

Table 3: We match model settings with their names from the paper. Settings are displayed as $[k, S, N]$ or [kernels scale dim., num scales, number of channels].

Model	Settings
S-ResNet, multiscale interaction	$[2, 4, 16]$
S-ResNet no interaction	$[1, 4, 16]$
ResNet, matched channels	$[1, 1, 16]$
ResNet, matched parameters	$[1, 1, 18]$

Table 4: A DenseNet. Input at the top. For shape, we show the number of scales, the height, the width, and the number of channels. S denotes the number of scales used per layer.

Layer type	Shape
dense ₁₂ [1, 3, 3] \times 3	[S , 96, 96, 39]
transition[3, 3, 3]	[S , 48, 48, 19]
dense ₂₄ [1, 3, 3] \times 3	[S , 48, 48, 94]
transition[3, 3, 3]	[S , 24, 24, 47]
dense ₄₈ [1, 3, 3] \times 3	[S , 24, 24, 213]
Global average pooling	[1, 1, 213]
Linear layer	[1, 1, 2]