DISCO: accurate Discrete Scale Convolutions

Ivan Sosnovik
i.sosnovik@uva.nl
Artem Moskalev
a.moskalev@uva.nl
Arnold Smeulders
a.w.m.smeulders@uva.nl

UvA-Bosch Delta Lab
University of Amsterdam
Netherlands

Abstract

Scale is often seen as a given, disturbing factor in many vision tasks. When doing so it is one of the factors why we need more data during learning. In recent work scale equivariance was added to convolutional neural networks. It was shown to be effective for a range of tasks. We aim for accurate scale-equivariant convolutional neural networks (SE-CNNs) applicable for problems where high granularity of scale and small kernel sizes are required. Current SE-CNNs rely on weight sharing and kernel rescaling, the latter of which is accurate for integer scales only. To reach accurate scale equivariance, we derive general constraints under which scale-convolution remains equivariant to discrete rescaling. We find the exact solution for all cases where it exists, and compute the approximation for the rest. The discrete scale-convolution pays off, as demonstrated in a new state-of-the-art classification on MNIST-scale and on STL-10 in the supervised learning setting. With the same SE scheme, we also improve the computational effort of a scale-equivariant Siamese tracker on OTB-13.

1 Introduction

Scale is a natural attribute of every object, as basic property as location and appearance. And hence it is a factor in almost every task in computer vision. In image classification, global scale invariance plays an important role in achieving accurate results [25]. In image segmentation, scale equivariance is important as the output map should scale proportionally to the input [1]. And in object detection or object tracking, it is important to be scale-agnostic [37], which implies the availability of both scale invariance as well as scale equivariance as the property of the method. Where scale invariance or equivariance is usually left as a property to learn in the training of these computer vision methods by providing a good variety of examples [31], we aim for accurate scale analysis for the purpose of needing less data to learn from.

Scale of the object can be derived externally from the size of its silhouette, e.g [51], or internally from the scale of its details, e.g [6]. External scale estimation requires the full object to be visible. It will easily fail when the object is occluded and/or when the object is

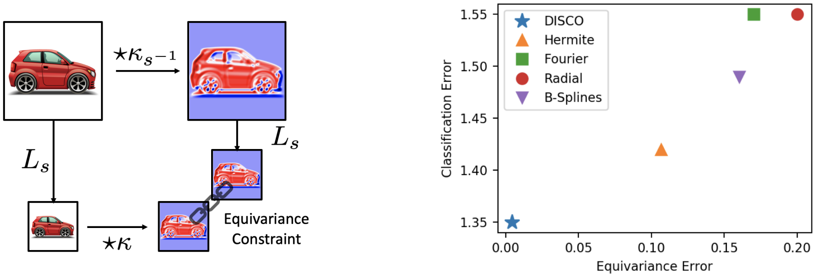

Figure 1: Left: the necessary constraint for scale-equivariance. When it is not satisfied an *equivariance error* appears. Right: Equivariance error vs. Classification error for scale-equivariant models on MNIST-scale. DISCO achieves the lowest equivariance error and this leads to the best classification accuracy. Alongside DISCO, we test SESN models with Hermite [41], Fourier [55], Radial [19] and B-Spline [3] bases.

amidst a cluttered background, for example for people in a crowd [40], when proper detection is hard. In contrast, internal scale estimation is build on the scale of common details [39], for example deriving the scale of a person from the scale of a sweater or a face. Where internal scale has better chances of being reliable, it poses heavier demands on the accuracy of assessment than external scale estimation. We focus on improvement of the accuracy of internal scale analysis.

We focus on accurate scale analysis on the generally applicable scale-equivariant convolutional neural networks [3, 41, 49]. A scale-equivariant network extends the equivariant property of conventional convolutions to the scale-translation group. It is achieved by rescaling the kernel basis and sharing weights between scales. While the weight sharing is defined by the structure of the group [9], the proper way to rescale kernels is an open problem. In [3, 41], the authors propose to rescale kernels in the continuous domain to project them later on a pixel grid. This permits the use of arbitrary scales, which is important to many application problems, but the procedure may cause a significant equivariance error [41]. Therefore, Worrall and Welling [49] model rescaling as a dilation, which guarantees a low equivariance error at the expense of permitting only integer scale factors. Due to the continuous nature of observed scale in segmentation, tracking or classification alike, integer scale factors may not cover the range of variations in the best possible way.

In the paper, we show how the equivariance error affects the performance of SE-CNNs. We make the following contributions:

- From first principles we derive the best kernels, which minimize the equivariance error.
- We find the conditions when the solution exists and find a good approximation when it does not exist.
- We demonstrate that an SE-CNN with the proposed kernels outperforms recent SE-CNNs in classification and tracking in both accuracy and compute time. We set new state-of-the-art results on MNIST-scale and STL-10.

The proposed approach contains [49] as a special case. Moreover, the proposed kernels can't be derived from [41] and vice versa. The union of our approach and the approach presented in [41] covers the whole set of possible SE-CNNs for a finite set of scales.

2 Related Work

Group Equivariant Networks. In recent years, various works on group-equivariant convolution neural networks have appeared. In majority, they consider the roto-translation group in 2D [9, 13, 24, 45, 47, 50], the roto-translation group in 3D [10, 26, 43, 46, 48], the compactified rotation-scaling group in 2D [21] and the rotation group 3D [10, 12, 18]. In [11, 27, 28] the authors demonstrate how to build convolution networks equivariant to arbitrary compact groups. All these papers cover group-equivariant networks for compact groups. In this paper, we focus the scale-translation group which is an example of a non-compact group.

Discrete Operators. Minimization of the discrepancies between the theoretical properties of continuous models and their discrete realizations has been studied for a variety of computer vision tasks. Lindeberg [32, 33] proposed a method for building a scale-space for discrete signals. The approach relied on the connection between the discretized version of the diffusion equation and the structure of images. While this method considered the scale symmetry of images and significantly improved computer vision models in the pre-deep-learning era, it is not directly applicable to our case of scale-equivariant convolutional networks.

In [16], Diaconu and Worrall demonstrate how to construct rotation-equivariant CNNs on the pixel grid for arbitrary rotations. The authors propose to learn the kernels which minimize the equivariance error of rotation-equivariant convolutional layers. The method relies on the properties of the rotation group and cannot be generalized to the scale-translation group. In this paper, we show how to minimize the equivariance error for scale-convolution without the use of extensive learning.

Scale-Equivariant CNNs. An early work of [25] introduced SI-ConvNet, a model where the input image is rescaled into a multi-scale pyramid. Alternatively, Xu *et al.* [52] proposed SiCNN, where a multi-scale representation is built from rescaling the network filters. While these modified convolutional networks significantly improve image classification, they require run-time interpolation. As a result they are several orders slower than standard CNNs.

In [3, 41, 55] the authors propose to parameterize the filters by a trainable linear combination of a pre-calculated, fixed multi-scale basis. Such a basis is defined in the continuous scale domain and projected on a pixel grid for the set of scale factors. The models do not involve interpolation during training nor inference. As a consequence, they operate within reasonable time. The continuous nature of the bases allows for the use of arbitrary scale factors, but it suffers from a reduced accuracy as the projection on the discrete grid causes an equivariance error.

Worral and Welling [49] propose to model filter rescaling by dilation. This solves the equivariance error of the previous method at the price of permitting only integer scale factors. That makes the method less suited for object tracking, depth analysis and fine-grained image classification, where subtle changes in the image scale are important in the performance. Our approach combines the best of the both worlds as it guarantees a low equivariance error for arbitrary scale factors.

Accurate Scale Analysis. Approaches based on feature pyramids are applied in many tasks [20, 31, 36, 44]. Their implementation require a significant specialisation of the network architecture. Models based on direct scale regression [7, 30, 37] have proved to be accurate in scale analysis, but they rely on a complicated training procedure. Scale-equivariant networks require only a drop-in replacement of the standard convolutions by scale-convolutions, while keeping the training procedure unchanged [3, 41, 42, 49]. We appreciate the universal applicability of scale-equivariant networks. We focus on this particular use in our implementation while the method we set out in this paper will also apply to other ways of using scale in computer vision.

Existing models for scale-equivariant networks bring computational overhead, which significantly slows down the training and the inference. In this paper, we present scale-equivariant models which allow for the accurate analysis of scale with a minimum computational overhead while retaining the advantage of being an easy replacement of convolutional layers to improve.

3 Method

Equivariance. A mapping g is equivariant under a transformation L if and only if there exists L' such that $g \circ L = L' \circ g$. If the mapping L' is identity, then g is invariant under transformation L.

Scale Transformations. Given a function $f : \mathbb{R} \to \mathbb{R}$ its scale transformation L_s is defined by

$$L_s[f](t) = f(s^{-1}t), \quad \forall s > 0 \tag{1}$$

We refer to cases with $s > 1$ as up-scalings and to cases with $s < 1$ as down-scalings, where $L_{1/2}[f]$ stands for a function down-scaled by a factor of 2.

The scale-translation group. We are interested in equivariance under the scale-translation group H and its subgroups. It consists of the translations t and scale transformations s which preserve the position of the center. $H = \{(s,t)\} = S \rtimes T$ is a semi-direct product of a multiplicative group $S = (\mathbb{R}^+, +)$ and an additive group $T = (\mathbb{R}, +)$. For the multiplication of its elements we have $(s_2, t_2) \cdot (s_1, t_1) = (s_1 s_2, s_2 t_1 + t_2)$. Scale transformation of a function defined on group H consists of a scale transformation of its spatial part as it is defined in the Equation 1 and a corresponding multiplicative transformation of its scale part. In other words

$$L_{\hat{s}}[f](s,t) = f(s\hat{s}^{-1}, \hat{s}^{-1}t) \tag{2}$$

3.1 Scale-Convolution

A scale-convolution of f and a kernel κ both defined on scale s and translation t is given by: [41]:

$$[f \star_H \kappa](s,t) = \sum_{s'} [f(s', \cdot) \star \kappa_s(s^{-1}s', \cdot)](\cdot, t) \tag{3}$$

where κ_s stands for an s-times up-scaled kernel κ, $\star$ and $\star_H$ are convolution and scale-convolution. The exact way the up-scaling is performed depends on how the down-scaling of the input signal works.

Scale-convolution is equivariant to transformations $L_{\hat{s}}$ from the group H, therefore the following holds true by definition:

$$[L_{\hat{s}}[f] \star_H \kappa] = L_{\hat{s}}[f \star_H \kappa] \tag{4}$$

Expanding the left-hand side of this relation by using Equation 3, choosing $s = 1$ and replacing $s' \to s'\hat{s}$ we find:

$$[L_{\hat{s}}[f] \star_H \kappa](s,t) = \sum_{s'} [L_{\hat{s}}[f(s', \cdot)] \star \kappa(\hat{s}s', \cdot)](\cdot, t) \tag{5}$$

For the right-hand side we have:

$$L_{\hat{s}}[f \star_H \kappa](s,t) = \sum_{s'} L_{\hat{s}}[f(s',\cdot) \star \kappa_{\hat{s}^{-1}}(\hat{s}s',\cdot)](\cdot,t) \tag{6}$$

Equating the two sides and choosing f to be zero on all scales but $s = 1$, we obtain the equivariance constraint for the kernels

$$L_s[f] \star \kappa = L_s[f \star \kappa_{s^{-1}}], \quad \forall f, s \tag{7}$$

We have found that *the mapping defined by Equation 3 is scale-equivariant only if a kernel and its up-scaled versions satisfy Equation 7*. Thus, it proves to be the necessary condition for scale-equivariant convolutions. In [3, 41, 55] the opposite, sufficient condition was proved. As a whole it defines the relation between scale convolution and the constraints of its kernels.

3.2 Exact Solution

In the continuous domain, convolution is defined as an integral over the spatial coordinates. [3, 41, 55] derives a solution for Equation 7:

$$\kappa_s(t) = s^{-1}\kappa(s^{-1}t) \tag{8}$$

However, when such kernels are calculated and projected on the pixel grid, a discrepancy between the left-hand side and the right-hand side of Equation 7 will appear. We refer to such inequality as the *equivariance error*.

We aim at directly solving Equation 7 in the discrete domain. In general, for discrete signals down-scaling is a non-invertible operation. Thus L_s is well-defined only for $s < 1$. We start by solving Equation 7 for 1-dimensional discrete signals. We prove its generalization to the 2-dimensional case in supplementary materials. Figure 2 illustrates the approach.

Let us consider a discrete signal f represented as a vector $\boldsymbol{f}$ of length N_{in}. It is down-scaled to length $N_{\text{out}} < N_{\text{in}}$ by L_s, which is represented as a rectangular interpolation matrix $\boldsymbol{L}$ of size $N_{\text{out}} \times N_{\text{in}}$. A convolution with a kernel κ is represented as a multiplication with a matrix $\boldsymbol{K}$ of size $N_{\text{out}} \times N_{\text{out}}$, and with a kernel $\kappa_{s^{-1}}$ written as a matrix $\boldsymbol{K}_{s^{-1}}$ of size $N_{\text{in}} \times N_{\text{in}}$. Then Equation 7 can be rewritten in matrix form as follows:

$$\boldsymbol{KLf} = \boldsymbol{LK}_{s^{-1}}\boldsymbol{f}, \, \forall \boldsymbol{f} \quad \iff \quad \boldsymbol{KL} = \boldsymbol{LK}_{s^{-1}} \tag{9}$$

Without loss of generality we assume circular boundary conditions. Then the matrix representations $\boldsymbol{K}$ and $\boldsymbol{K}_{s^{-1}}$ are both circulant and their eigenvectors are the column-vectors of the Discrete Fourier Transform $\boldsymbol{F}$ [2, 5, 22]:

$$\boldsymbol{K}_{s^{-1}} = \boldsymbol{F}\text{diag}(\boldsymbol{F}\boldsymbol{\kappa}_{s^{-1}})\boldsymbol{F}^* \tag{10}$$

where $\boldsymbol{\kappa}_{s^{-1}}$ is a vector representation of $\kappa_{s^{-1}}$ padded with zeros. After substituting Equation 10 into Equation 9 and multiplying both sides by $\boldsymbol{F}$ from the right, we get:

$$\boldsymbol{KLF} = \boldsymbol{LF}\text{diag}(\boldsymbol{F}\boldsymbol{\kappa}_{s^{-1}}) \tag{11}$$

The left-hand side of the equation is obtained from $\boldsymbol{LF}$ by multiplying it with a diagonal matrix from the right. Thus, each column of the matrix $\boldsymbol{KLF}$ is proportional to the corresponding column of the matrix $\boldsymbol{LF}$. We prove in supplementary materials that *such a relation is possible if and only if the matrix* $\mathbf{L}$ *performs a down-scaling by an integer scale factor*.

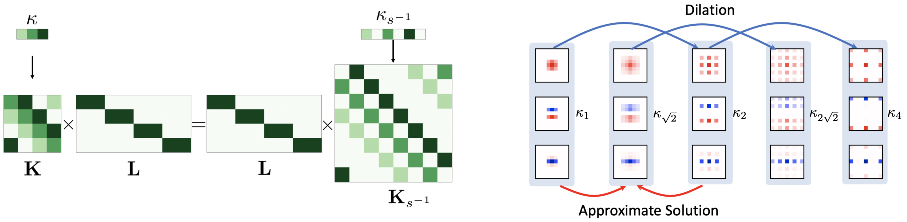

Figure 2: Left: a matrix representation of the 1-dimensional case of the equivariance constraint for $N_{\text{in}} = 8$ and $N_{\text{out}} = 4$. Right: a multi-scale kernel initialization. $\sqrt{2}$ is the smallest non-integer scale, for which the kernel is approximated by minimizing Equation 13, the rest of the kernels can be obtained with dilation.

When the requirement is satisfied, the solution with respect to $\boldsymbol{\kappa}_{s^{-1}}$ is the dilation of $\boldsymbol{\kappa}$ by factor s. Such a solution also known as the *à trous algorithm* [23]:

$$(\boldsymbol{\kappa}_{s^{-1}})_{is} = \sum_i \boldsymbol{F}^*_{ij} (\boldsymbol{KLF})_{1j} / (\boldsymbol{LF})_{1j} = \boldsymbol{\kappa}_i \tag{12}$$

3.3 Approximate solution

Let us consider a scale-convolutional layer. One of its hyper-parameters is the set of scales it operates on. For the cases of non-integer scale factors any kernels will introduce an equivariance error into the network. Thus, it is reasonable to use integer scales as reference points and add intermediate scales to cover the required range of scale factors best. Let us choose a set of scales $\{1, \sqrt{2}, 2, 2\sqrt{2}, 4, 4\sqrt{2}, \dots\}$. The set of corresponding kernels is $\{\kappa_1, \kappa_{\sqrt{2}}, \kappa_2, \kappa_{2\sqrt{2}}, \dots\}$. As the smallest kernel is known, all kernels defined on integer scales can be calculated as its dilated versions. And, when kernel $\kappa_{\sqrt{2}}$ is defined, all intermediate kernels $\kappa_{2\sqrt{2}}, \kappa_{4\sqrt{2}}, \dots$ can be calculated by using dilation as well. Thus, the only kernel yet unknown is kernel $\kappa_{\sqrt{2}}$.

The kernel $\kappa_{\sqrt{2}}$ can be calculated as a minimizer of the equivariance error based on the Equation 7 as follows:

$$\kappa_{\sqrt{2}} = \arg\min \mathbb{E}_f \|L[f] \star \kappa_1 - L[f \star \kappa_{\sqrt{2}}]\|^2_F + \|L[f] \star \kappa_{\sqrt{2}} - L[f \star \kappa_2]\|^2_F \tag{13}$$

where $L = L_{1/\sqrt{2}}$ is a down-scaling by a factor $\sqrt{2}$.

We demonstrate how to calculate approximate solution for the most general case in supplementary materials.

3.4 Implementation

To construct scale-equivariant convolution we parametrize the kernels as a linear combination of fixed multi-scale basis. The basis is then fixed and only corresponding coefficients are trained. The coefficients are shared for all scales.

We utilize the standard pixel basis on the smallest integer scale. The bases for the rest of the integer scales are computed as a dilation. The basis on the smallest non-integer scale is approximated by applying gradient descent to Equation 13. We note that it takes negligible time to compute all of the basis functions before training. See supplementary materials

Model	Basis	MNIST	MNIST+	Equi. error	# Params.
CNN	-	2.02 ± 0.07	1.60 ± 0.09	-	495 K
SiCNN	-	2.02 ± 0.14	1.59 ± 0.03	-	497 K
SI-ConvNet	-	1.82 ± 0.11	1.59 ± 0.10	-	495 K
SEVF	-	2.12 ± 0.13	1.81 ± 0.09	-	475 K
DSS	Dilation	1.97 ± 0.08	1.57 ± 0.09	0.0	494 K
SS-CNN	Radial	1.84 ± 0.10	1.76 ± 0.07	-	494 K
SESN	Hermite	1.68 ± 0.06	1.42 ± 0.07	0.107	495 K
SESN	B-Spline	1.74 ± 0.08	1.49 ± 0.05	0.163	495 K
SESN	Fourier	1.88 ± 0.07	1.55 ± 0.07	0.170	495 K
SESN	Radial	1.74 ± 0.07	1.55 ± 0.10	0.200	495 K
DISCO	Discrete	$\mathbf{1.52 \pm 0.06}$	$\mathbf{1.35 \pm 0.05}$	0.004	495 K

Table 1: The classification error of various methods on the MNIST-scale dataset, lower is better. We test both the regime with and without data augmentation, where scaling data augmentation is denoted by "+". All results are reported as mean ± std over 6 different, fixed realizations of the dataset. The best results are **bold**.

for more details. We refer to scale-convolutions with the proposed bases as Discrete Scale Convolutions or shortly DISCO. As DISCO kernels are sparse, they allow for lower computational complexity.

4 Experiments

4.1 Equivariance Error

To quantitatively evaluate the equivariance error of DISCO versus other methods for scale-convolution [3, 41, 55], we follow the approach proposed in [41]. In particular, we randomly sample images from the MNIST-Scale dataset [41] and pass in through the scale-convolution layer. Then, the equivariance error is calculated as follows:

$$\Delta = \sum_s \|L_s \Phi(f) - \Phi(L_s f)\|_2^2 / \|L_s \Phi(f)\|_2^2 \tag{14}$$

where Φ is scale-convolution with weights initialized randomly.

The equivariance error for each model is reported in Table 1 and in Figure 1. Note that we can not directly compare against [49] as it only permits integer scale factors. As can be seen, there exists a correlation between an equivariance error and classification accuracy. DISCO model attains the lowest equivariance error.

4.2 Image Classification

We conduct several experiments to compare various methods for scale analysis in image classification. Alongside DISCO, we test SI-ConvNet [25], SS-CNN [19], SiCNN [52], SEVF [34], DSS [49] and SESN [41]. By relying on the code provided by the authors we additionally reimplement SESN models with other bases such as B-Splines [3], Fourier-Bessel Functions [55] and Log-Radial Harmonics [19, 35].

Model	WRN	SiCNN	SI-ConvNet	DSS	SS-CNN	SESN	DISCO
Basis	-	-	-	Dilation	Radial	Hermite	Discrete
Time, s	10	110	55	40	15	165	50
Error	11.48	11.62	12.48	11.28	25.47	8.51	**8.07**

Table 2: The classification error on STL-10. The best results are in **bold**. The average compute time per epoch is reported in seconds. DISCO sets a new state-of-the-art result in the supervised learning setting.

Equi. Error	STL-10 Error
0.240	8.63
0.082	8.25
0.003	**8.07**

Table 3: Classification accuracy on STL-10 and the equivariance error for the DISCO model with different filters. The first and the second rows correspond to the cases when the basis for the intermediate scale is not optimized.

MNIST-scale. Following [41] we conduct experiments on the MNIST-scale dataset. The dataset consists of 6 splits, each of which contains 10,000 images for training, 2,000 for validation and 50,000 for testing. Each image is a randomly rescaled version of the original from MNIST [29]. The scaling factors are uniformly sampled from the range of $0.3-1.0$.

As a baseline model we use the SESN model, which holds the state-of-the-art result on this dataset. Both SESN and DISCO use the same set of scales in scale convolutions: $\{1, 2^{1/3}, 2^{2/3}, 2\}$ and are trained in exactly the same way. As can be seen from Table 1, our DISCO model outperforms other scale equivariant networks in accuracy and equivariance error and sets a new state-of-the-art result.

STL-10. To demonstrate how accurate scale equivariance helps when the training data is limited, we conduct experiments on the STL-10 [8] dataset. This dataset consists of just 8,000 training and 5,000 testing images, divided into 10 classes. Each image has a resolution of 96×96 pixels.

As a baseline we use WideResNet [53] with 16 layers and a widening factor of 8. Scale-equivariant models are constructed according to [41]. All models have the same number of parameters, the same set of scales $\{1, \sqrt{2}, 2\}$ and are trained for the same number of steps. For testing the disco model we use exactly the same setup as described by the authors of [41]. All the models are trained on NVidia GTX 1080 Ti.

The models are trained for 1000 epochs using the SGD optimizer with a Nesterov momentum of 0.9 and a weight decay of $5 \cdot 10^{-4}$. For DISCO, we increase the weight decay to $1 \cdot 10^{-4}$. Tuning weight decay for the other models did not bring any improvement. The learning rate is set to 0.1 at the start and decreased by a factor of 0.2 after the epochs 300, 400, 600 and 800. The batch size is set to 128. During training, we additionally augment the dataset with random crops, horizontal flips and cutout [13].

As can be seen from Table 2, the proposed DISCO model outperforms the other scale-equivariant networks and sets a new state-of-the-art result in the supervised learning setting. Moreover, DISCO is more than 3 times faster than the second-best SESN-model.

We additionally check how accuracy degrades if the basis for the scale of $\sqrt{2}$ is not

Model	SiamFC [4]	TriSiam [17]	SiamFC+[54]	SE-SiamFC+ [42]	DISCO
FPS	-	-	56	14	28
AUC	0.61	0.62	0.67	**0.68**	**0.68**

Table 4: Performance comparisons on the OTB-13 tracking benchmark. The best results are **bold**. We report the average number of framer per second (FPS) per sequence. Higher FPS and AUC are better.

correctly calculated. While the optimal basis is a minimizer of Equation 13, it is possible to stop the stop optimization procedure before convergence and generate then a non-optimal basis. We generated two non-optimal bases which correspond to different moments of the optimization procedure. We report the equivariance error and the classification error on the STL-10 dataset for DISCO with such bases functions in Table 3. It can be seen that lower equivariance errors correspond to lower classification errors.

4.3 Tracking

To test the ability of DISCO to deliver accurate scale estimation, we choose the task of visual object tracking. We take the recent SE-SiamFC+ [54] tracker and follow the recipe provided in [42] to make it scale-equivariant. We employ the standard one-pass evaluation protocol to compare our method with conventional Siamese trackers and SE-SiamFC+ [42] with a Hermite basis for the scale convolutions. The trackers are evaluated by the usual area-under-the-success-curve (AUC).

The scale-equivariant tracker with DISCO matches the performance of the state-of-the-art SE-SiamFC+, but twice faster as can be seen in Table 4. FPS is measured on Nvidia GTX 1080 Ti for all models.

4.4 Scene Geometry by Contrasting Scales

We demonstrate the ability of DISCO to propagate scale information through the layers of the network, by presenting a simple approach for geometry estimation of a scene through the use of the intrinsic scale. This is possible because in the DISCO model, we can use high granularity of scale factors and process them more accurately and faster compared to other scale-equivariant models.

We construct a scale-equivariant network with DISCO layers. The weights are initialized from an ImageNet-pretrained network [14] following the approach described in [42]. Next, we strip the classification head of the network and apply global spatial average-pooling. The resulting feature map thus has a dimension $B \times C \times S \times 1 \times 1$, where B, C, S are the batch, channel and scale dimensions respectively. To decode the scale information, we sample the argmax along the scale dimension. Such a tensor has shape $B \times C$ where each element is a scalar that encodes the argmax for each of the objects on each of the channels. Then the tensor is passed to a shallow network, which produces a scale estimate for the input image. The feature extraction network followed by the shallow scale estimator network is denoted as F_θ, where θ is the parameters of the shallow scale estimator, so we do not train the parameters of the feature extractor.

At the core of the method is the scale-contrastive learning algorithm. The model is trained to predict how much one image should be interpolated to match the other. Such an

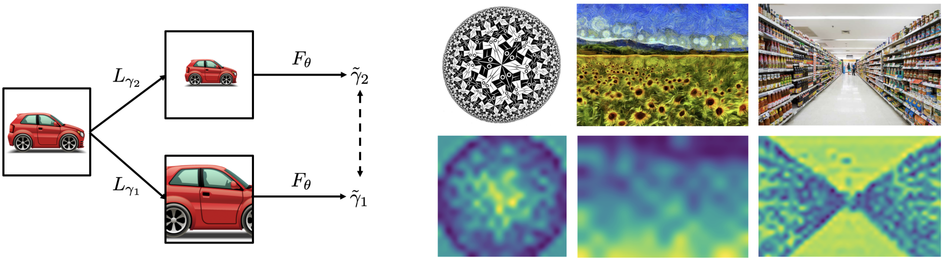

Figure 3: Left: the network is trained to predict the scale difference between an object and its resized version. Right: images and their scale fields produced by the DISCO model trained to contrast scales.

approach does not require any dedicated depth or scale labels. The algorithm is illustrated in Figure 3. First, we sample randomly two scale factors $\gamma_1, \gamma_2 \sim U[0.5, 2.0]$ and apply interpolations $L_{\gamma_1}, L_{\gamma_2}$ to the image $\mathcal{I}$. The transformed images are fed into the network F_θ, which predicts scale estimates $\tilde{\gamma}_1, \tilde{\gamma}_2$ (Figure 3). Then, we minimize the following loss by using the Adam optimizer:

$$\mathcal{L}_{\text{scale}} = \mathbb{E}_{\mathcal{I}} \left[\frac{\gamma_2}{\gamma_1} - \frac{\tilde{\gamma}_2}{\tilde{\gamma}_1} \right]^2 = \mathbb{E}_{\mathcal{I}} \left[\frac{\gamma_2}{\gamma_1} - \frac{F_\theta(L_{\gamma_2}(\mathcal{I}))}{F_\theta(L_{\gamma_1}(\mathcal{I}))} \right]^2 \longrightarrow \min_\theta \tag{15}$$

We train the model on the STL-10 dataset [8] and evaluate it on random images found on the Internet. To infer the scene geometry of the image, we split the image into overlapping patches. For each of them we predict the scale. We provide qualitative results in Figure 3. While the proposed methods was never trained on whole images, it captures the global geometry of the scenes, be it a road or a supermarket.

We provide more detailed information for each of the experiments in supplementary materials.

5 Discussion

In this work, we demonstrate that the equivariance error affects the performance of equivariant networks. We introduce DISCO, a new class of kernels for scale-convolution, so the equivariance error is minimized. We develop a theory to derive an optimal rescaling to be used in DISCO and analyze under what conditions an optimal rescaling is possible and how to find a good approximation if these conditions do not hold. We also demonstrate how to efficiently incorporate DISCO into an existing scale-equivariant network.

We experimentally demonstrate that DISCO scale-equivariant networks outperform conventional and other scale-equivariant models, setting the new state-of-the-art on the MNIST-Scale and STL-10 datasets. In the visual object tracking experiment, DISCO matches the state-of-the-art performance of SE-SiamFC+ on OTB-13, however, works 2 times faster.

We suppose that the DISCO would be the most useful in problems, where an accurate scale analysis is required, such as multi-object tracking for autonomous vehicles, where the scale of objects can rapidly change due to the relative motion. We additionally want to highlight that the approach presented in this paper can be used to construct scale-equivariant self-attention models with reduced complexity [38].

References

[1] Helen L Anderson, Ruzena Bajcsy, and Max Mintz. Adaptive image segmentation. 1988.

[2] Bassam Bamieh. Discovering transforms: A tutorial on circulant matrices, circular convolution, and the discrete fourier transform. *arXiv preprint arXiv:1805.05533*, 2018.

[3] Erik J Bekkers. B-spline cnns on lie groups. *arXiv preprint arXiv:1909.12057*, 2019.

[4] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pages 850–865. Springer, 2016.

[5] Chu-Yin Chang, Anthony A Maciejewski, and Venkataramanan Balakrishnan. Fast eigenspace decomposition of correlated images. *IEEE Transactions on Image Processing*, 9(11):1937–1949, 2000.

[6] Jason Chang and John W Fisher. Analysis of orientation and scale in smoothly varying textures. In *2009 IEEE 12th International Conference on Computer Vision*, pages 881–888. IEEE, 2009.

[7] Zedu Chen, Bineng Zhong, Guorong Li, Shengping Zhang, and Rongrong Ji. Siamese box adaptive network for visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6668–6677, 2020.

[8] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223, 2011.

[9] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999, 2016.

[10] Taco Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Convolutional networks for spherical signals. *arXiv preprint arXiv:1709.04893*, 2017.

[11] Taco Cohen, Mario Geiger, and Maurice Weiler. A general theory of equivariant cnns on homogeneous spaces. *arXiv preprint arXiv:1811.02017*, 2018.

[12] Taco Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional networks and the icosahedral cnn. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2019.

[13] Taco S Cohen and Max Welling. Steerable cnns. *arXiv preprint arXiv:1612.08498*, 2016.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[15] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[16] Nichita Diaconu and Daniel Worrall. Learning to convolve: A generalized weight-tying approach. In *International Conference on Machine Learning*, pages 1586–1595. PMLR, 2019.

[17] Xingping Dong and Jianbing Shen. Triplet loss in siamese network for object tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 459–474, 2018.

[18] Carlos Esteves, Christine Allen-Blanchette, Ameesh Makadia, and Kostas Daniilidis. Learning so (3) equivariant representations with spherical cnns. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 52–68, 2018.

[19] Rohan Ghosh and Anupam K Gupta. Scale steerable filters for locally scale-invariant convolutional neural networks. *arXiv preprint arXiv:1906.03861*, 2019.

[20] Dongyoon Han, Jiwhan Kim, and Junmo Kim. Deep pyramidal residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5927–5935, 2017.

[21] Joao F Henriques and Andrea Vedaldi. Warped convolutions: Efficient invariance to spatial transformations. In *International Conference on Machine Learning*, pages 1461–1469. PMLR, 2017.

[22] Joao F Henriques, Pedro Martins, Rui F Caseiro, and Jorge Batista. Fast training of pose detectors in the fourier domain. *Advances in neural information processing systems*, 27:3050–3058, 2014.

[23] Matthias Holschneider, Richard Kronland-Martinet, Jean Morlet, and Ph Tchamitchian. A real-time algorithm for signal analysis with the help of the wavelet transform. In *Wavelets*, pages 286–297. Springer, 1990.

[24] Emiel Hoogeboom, Jorn WT Peters, Taco S Cohen, and Max Welling. Hexaconv. *arXiv preprint arXiv:1803.02108*, 2018.

[25] Angjoo Kanazawa, Abhishek Sharma, and David Jacobs. Locally scale-invariant convolutional neural networks. *arXiv preprint arXiv:1412.5104*, 2014.

[26] Risi Kondor. N-body networks: a covariant hierarchical neural network architecture for learning atomic potentials. *arXiv preprint arXiv:1803.01588*, 2018.

[27] Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *International Conference on Machine Learning*, pages 2747–2755. PMLR, 2018.

[28] Leon Lang and Maurice Weiler. A wigner-eckart theorem for group equivariant convolution kernels. *arXiv preprint arXiv:2010.10952*, 2020.

[29] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[30] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4282–4291, 2019.

[31] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[32] Tony Lindeberg. Scale-space for discrete signals. *IEEE transactions on pattern analysis and machine intelligence*, 12(3):234–254, 1990.

[33] Tony Lindeberg. *Scale-space theory in computer vision*, volume 256. Springer Science & Business Media, 2013.

[34] Diego Marcos, Benjamin Kellenberger, Sylvain Lobry, and Devis Tuia. Scale equivariance in cnns with vector fields. *arXiv preprint arXiv:1807.11783*, 2018.

[35] Hanieh Naderi, Leili Goli, and Shohreh Kasaei. Scale equivariant cnns with scale steerable filters. In *2020 International Conference on Machine Vision and Image Processing (MVIP)*, pages 1–5. IEEE, 2020.

[36] Siyuan Qiao, Liang-Chieh Chen, and Alan Yuille. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. *arXiv preprint arXiv:2006.02334*, 2020.

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

[38] David W Romero and Jean-Baptiste Cordonnier. Group equivariant stand-alone self-attention for vision. *arXiv preprint arXiv:2010.00977*, 2020.

[39] Henry Schneiderman and Takeo Kanade. Object detection using the statistics of parts. *International Journal of Computer Vision*, 56(3):151–177, 2004.

[40] Arnold WM Smeulders, Dung M Chu, Rita Cucchiara, Simone Calderara, Afshin Deghghan, and Mubarak Shah. Visual tracking: An experimental survey. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1442–1468, 2013.

[41] Ivan Sosnovik, Michał Szmaja, and Arnold Smeulders. Scale-equivariant steerable networks. *arXiv preprint arXiv:1910.11093*, 2019.

[42] Ivan Sosnovik, Artem Moskalev, and Arnold W.M. Smeulders. Scale equivariance improves siamese tracking. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2765–2774, January 2021.

[43] Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018.

[44] Xinjiang Wang, Shilong Zhang, Zhuoran Yu, Litong Feng, and Wayne Zhang. Scale-equalizing pyramid convolution for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13359–13368, 2020.

[45] Maurice Weiler and Gabriele Cesa. General $e(2)$-equivariant steerable cnns. *arXiv preprint arXiv:1911.08251*, 2019.

[46] 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*, pages 10381–10392, 2018.

[47] Maurice Weiler, Fred A Hamprecht, and Martin Storath. Learning steerable filters for rotation equivariant cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 849–858, 2018.

[48] Daniel Worrall and Gabriel Brostow. Cubenet: Equivariance to 3d rotation and translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 567–584, 2018.

[49] Daniel Worrall and Max Welling. Deep scale-spaces: Equivariance over scale. In *Advances in Neural Information Processing Systems*, pages 7366–7378, 2019.

[50] Daniel E Worrall, Stephan J Garbin, Daniyar Turmukhambetov, and Gabriel J Brostow. Harmonic networks: Deep translation and rotation equivariance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5028–5037, 2017.

[51] Yiran Wu, Sihao Ying, and Lianmin Zheng. Size-to-depth: a new perspective for single image depth estimation. *arXiv preprint arXiv:1801.04461*, 2018.

[52] Yichong Xu, Tianjun Xiao, Jiaxing Zhang, Kuiyuan Yang, and Zheng Zhang. Scale-invariant convolutional neural networks. *arXiv preprint arXiv:1411.6369*, 2014.

[53] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016.

[54] Zhipeng Zhang and Houwen Peng. Deeper and wider siamese networks for real-time visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4591–4600, 2019.

[55] Wei Zhu, Qiang Qiu, Robert Calderbank, Guillermo Sapiro, and Xiuyuan Cheng. Scale-equivariant neural networks with decomposed convolutional filters. *arXiv preprint arXiv:1909.11193*, 2019.