

Scale Equivariance Improves Siamese Tracking

Ivan Sosnovik* Artem Moskalev * Arnold Smeulders
 UvA-Bosch Delta Lab
 University of Amsterdam, Netherlands

{i.sosnovik, a.moskalev, a.w.m.smeulders}@uva.nl

Abstract

Siamese trackers turn tracking into similarity estimation between a template and the candidate regions in the frame. Mathematically, one of the key ingredients of success of the similarity function is translation equivariance. Non-translation-equivariant architectures induce a positional bias during training, so the location of the target will be hard to recover from the feature space. In real life scenarios, objects undergo various transformations other than translation, such as rotation or scaling. Unless the model has an internal mechanism to handle them, the similarity may degrade. In this paper, we focus on scaling and we aim to equip the Siamese network with additional built-in scale equivariance to capture the natural variations of the target a priori. We develop the theory for scale-equivariant Siamese trackers, and provide a simple recipe for how to make a wide range of existing trackers scale-equivariant. We present SE-SiamFC, a scale-equivariant variant of SiamFC built according to the recipe. We conduct experiments on OTB and VOT benchmarks and on the synthetically generated T-MNIST and S-MNIST datasets. We demonstrate that a built-in additional scale equivariance is useful for visual object tracking.

1. Introduction

Siamese trackers turn tracking into similarity estimation between a template and the candidate regions in the frame. The Siamese networks are successful because the similarity function is powerful: it can learn the variances of appearance very effectively, to such a degree that even the association of the frontside of an unknown object to its backside is usually successful. And, once the similarity is effective, the location of the candidate region is reduced to simply selecting the most similar candidate.

Mathematically, one of the key ingredients of the success of the similarity function is translation *equivariance*, i.e. a

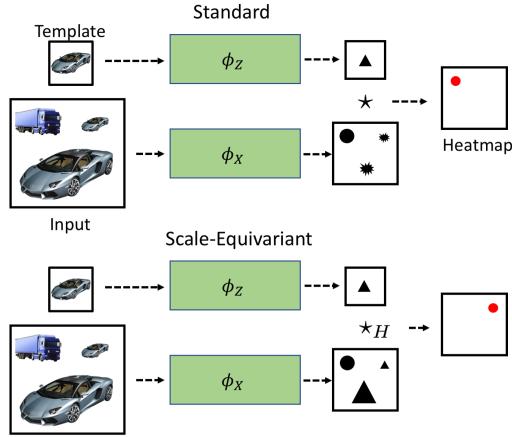


Figure 1: The standard version (top) and the scale-equivariant version (bottom) of a basic tracker. The scale-equivariant tracker has an internal notion of scale which allows for the distinction between similar objects which only differ in scale. The operator \star denotes convolution, and \star_H stands for scale-convolution.

translation in the input image is to result in the proportional translation in feature space. Non-translation-equivariant architectures will induce a positional bias during training, so the location of the target will be hard to recover from the feature space [21, 38]. In real-life scenarios, the target will undergo more transformations than just translation, and, unless the network has an internal mechanism to handle them, the similarity may degrade. We start from the position that equivariance to common transformations should be the guiding principle in designing conceptually simple yet robust trackers. To that end, we focus on scale equivariance for trackers in this paper.

Measuring scale precisely is crucial when the camera zooms its lens or when the target moves into depth. However, scale is also important in distinguishing among objects in general. In following a marching band or in analyzing a soccer game, or when many objects in the video have a similar appearance (a crowd, team sports), the sim-

*equal contribution

Source code: <https://github.com/isosnovik/SiamSE>

ilarity power of Siamese trackers has a hard time locating the right target. In such circumstances, spatial-scale equivariance will provide a richer and hence more discriminative descriptor, which is essential to differentiate among several similar candidates in an image. And, even, as we will demonstrate, when the sequence does not show variation over scale, proper scale measurement is important to keep the target bounding box stable in size.

The common way to implement scale into a tracker is to train the network on a large dataset where scale variations occur naturally. However, as was noted in [20], such training procedures may lead to learning groups of re-scaled duplicates of almost the same filters. As a consequence, inter-scale similarity estimation becomes unreliable, see Figure 1 top. Scale-equivariant models have an internal notion of scale and built-in weight sharing among different filter scales. Thus, scale equivariance aims to produce the same distinction for all sizes, see Figure 1 bottom.

In this paper, we aim to equip the Siamese network with spatial and scale equivariance built-in from the start to capture the natural variations of the target *a priori*. We aim to improve a broad class of tracking algorithms by enhancing their capacity of candidate distinction. We adopt recent advances [28] in convolutional neural networks (CNNs) which handle scale variations explicitly and efficiently.

While scale-equivariant convolutional models have lead to success in image classification [28, 33], we focus on their usefulness in *object localization*. Where scale estimation has been used in the localization for tracking, it typically relies on brute-force multi-scale detection with an obvious computational burden [8, 2], or on a separate network to estimate the scale [6, 22]. Both approaches will require attention to avoid bias and the propagation thereof through the network. Our new method treats scale and scale equivariance as a desirable fundamental property, which makes the algorithm conceptually easier. Hence, scale equivariance should be easy to merge into an existing network for tracking. Then, scale equivariance will enhance the performance of the tracker without further modification of the network or extensive data augmentation during the learning phase.

We make the following contributions:

- We propose the theory for scale-equivariant Siamese trackers and provide a simple recipe of how to make a wide range of existing trackers scale-equivariant.
- We propose building blocks necessary for efficient implementation of scale equivariance into modern Siamese trackers and implement a scale-equivariant extension of the recent SiamFC+ [38] tracker.
- We demonstrate the advantage of scale-equivariant Siamese trackers over their conventional counterparts on popular benchmarks for sequences with and without apparent scale changes.

2. Related Work

Siamese tracking The challenge of learning to track arbitrary objects can be addressed by deep similarity learning [2]. The common approach is to employ Siamese networks to compute the embeddings of the original patches. The embeddings are then fused to obtain a location estimate. Such formulation is general, allowing for a favourable flexibility in the design of the tracker. In [2] Bertinetto *et al.* employ off-line trained CNNs as feature extractors. The authors compare dot-product similarities between the feature map of the template with the maps coming from the current frame and measure similarities on multiple scales. Held *et al.* [14] suggest a detection-based Siamese tracker, where the similarity function is modeled as a fully-connected network. Extensive data augmentation is applied to learn a similarity function, which generalizes for various object transformations. Li *et al.* [22] consider tracking as a one-shot detection problem to design Siamese region-proposal-networks [25] by fusing the features from a fully-convolutional backbone. The recent ATOM [6] and DIMP [3] trackers employ a multi-stage tracking framework, where an object is coarsely localized by the online *classification* branch, and subsequently refined in its position by the *estimation* branch. From a Siamese perspective, in both [6, 3] the object embeddings are first fused to produce an initial location and subsequently processed by the IoU-Net [17] to enhance the precision of the bounding box.

The aforementioned references have laid the foundation for most of the state-of-the-art trackers. These methods share an implicit or explicit attention to translation equivariance for feature extraction. The decisive role of translation equivariance is noted in [2, 21, 38]. Bertinetto *et al.* [2] utilize fully-convolutional networks where the output directly commutes with a shift in the input image as a function of the total stride. Li *et al.* [21] suggest a training strategy to eliminate the spatial bias introduced in non-fully-convolutional backbones. Along the same line, Zhang and Peng [38] demonstrated that deep state-of-the-art models developed for classification are not directly applicable for localization. And hence these models are not directly applicable to tracking as they induce positional bias, which breaks strict translation equivariance. We argue that transformations, other than translation, such as rotation may be equally important for certain classes of videos like sports and following objects in the sea or in the sky. And we argue that scale transformation is common in the majority of sequences due to the changing distances between objects and the camera. In this paper, we take on the latter class of transformations for tracking.

Equivariant CNNs Various works on transformation-equivariant convolutional networks have been published recently. They extend the built-in property of translation-

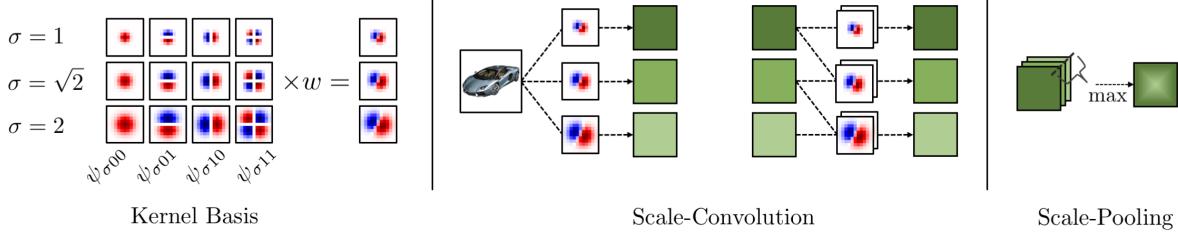


Figure 2: Left: convolutional kernels use a fixed kernel basis on multiple scales, each with a set of trainable weights. Middle: a representation of scale-convolution using Equation 6 for the first and all subsequent layers. Right: a scheme of scale-pooling, which transforms a 3D-signal into a 2D one without losing scale equivariance. As an example, we use a basis of 4 functions and 3 scales with a step of $\sqrt{2}$. Only one channel of each convolutional layer is demonstrated for simplicity.

equivariance of conventional CNNs to a broader set of transformations. Mostly considered was roto-translation, as demonstrated on image classification [4, 5, 15, 30, 27, 26], image segmentation [31] and edge detection [33].

One of the first works on scale-translation-equivariant convolutional networks was by Marcos *et al.* [24]. In order to process images on multiple scales, the authors resize and convolve the input of each layer multiple times, forming a stack of features which corresponds to variety of scales. The output of such a convolutional layer is a vector whose length encodes the maximum response in each position among different scales. The direction of the vector is derived from the scale, which gave the maximum. The method has almost no restrictions in the choice of admissible scales. As this approach relies on rescaling the image, the obtained models are significantly slower compared to conventional CNNs. Thus, this approach is not suitable for being applied effectively in visual object tracking.

Worrall & Welling [32] propose Deep Scale-Spaces, an equivariant model which generalizes the concept of scale-space to deep networks. The approach uses filter dilation to analyze the images on different scales. It is almost as fast as a conventional CNN with the same width and depth. As the method is restricted to integer scale factors it is unsuited to applications in tracking where the scene dictates arbitrary scale factors.

Almost simultaneously, three papers [28, 1, 39] were proposed to implement scale-translation-equivariant networks with arbitrary scales. What they have in common is that they use a pre-calculated and fixed basis defined on multiple scales. All filters are then calculated as a linear combination of the basis and trainable weights. As a result, no rescaling is used. We prefer to use [28], as Sosnovik *et al.* propose an approach for building general scale-translation-equivariant networks with an algorithm for the fast implementation of the scale-convolution.

To date, the application of scale-equivariant networks was mostly demonstrated in image classification. Almost

no attention was paid to tasks that involve object localization, such as visual object tracking. As we have noted above, it is a fundamentally different case. To the best of our knowledge, we demonstrate the first application of transformation-equivariant CNNs to visual object tracking.

3. Scale-Equivariant Tracking

In this work, we consider a wide range of modern trackers which can be described by the following formula:

$$h(z, x) = \phi_X(x) \star \phi_Z(z) \quad (1)$$

where z, x are the template and the input frame, and ϕ_X, ϕ_Z are the functions which process them, and \star is the convolution operator which implements a connection between two signals. The resulting value $h(z, x)$ is a heatmap that can be converted into a prediction by relatively simple calculations. Functions ϕ_X, ϕ_Z here can be parametrized as feed-forward neural networks. For our analysis, it is both suitable if the weights of these networks are fixed or updated during training or inference. This pipeline describes the majority of Siamese trackers such as [2, 22, 21] and the trackers based on correlation filters [7, 8].

3.1. Convolution is all you need

Let us consider some mapping g . It is equivariant under a transformation L if and only if there exists L' such that $g \circ L = L' \circ g$. If L' is the identity mapping, then the function g is invariant under this transformation. A function of multiple variables is equivariant when it is equivariant with respect to each of the variables. In our analysis, we consider only transformations that form a transformation group, in other words, $L \in G$.

Theorem 1. *A function given by Equation 1 is equivariant under a transformation L from group G if and only if ϕ_X and ϕ_Z are constructed from G -equivariant convolutional layers and \star is the G -convolution.*

The proof of Theorem 2 is given in the supplementary material. A simple interpretation of this theorem is that *a tracker is equivariant to transformations from G if and only if it is fully G -convolutional*. The necessity of fully-convolutional trackers is well-known in tracking community and is related to the ability of the tracker to capture the main variations in the video — the translation. In this paper, we seek to extend this ability to scale variations as well. Which, due to Theorem 2 boils down to using scale-convolution and building fully scale-translation convolutional trackers.

3.2. Scale Modules

Given a function $f : \mathbb{R} \rightarrow \mathbb{R}$, a scale transformation is defined as follows:

$$L_s[f](t) = f(s^{-1}t), \quad \forall s \geq 0 \quad (2)$$

where cases with $s > 1$ are referred to as upscale and with $s < 1$ as downscale. Standard convolutional layers and convolutional networks are translation equivariant but not scale-equivariant [28].

Parametric Scale-Convolution In order to build scale-equivariant convolutional networks, we follow the method proposed by Sosnovik *et al.* [28]. We begin by choosing a complete basis of functions defined on multiple scales. Choosing the center of the function to be the point $(0, 0)$ in coordinates (u, v) , we use functions of the following form:

$$\psi_{\sigma nm}(u, v) = A \frac{1}{\sigma^2} H_n\left(\frac{u}{\sigma}\right) H_m\left(\frac{v}{\sigma}\right) e^{-\frac{u^2+v^2}{2\sigma^2}} \quad (3)$$

Here H_n is a Hermite polynomial of the n -th order, and A is a constant used for normalization. In order to build a basis of N functions, we iterate over increasing pairs of n and m . As the basis is complete, the number of functions N is equal to the number of pixels in the original filter. We build such a basis for a chosen set of equidistant scales σ and fix it:

$$\Psi_\sigma = \left\{ \psi_{\sigma 00}, \psi_{\sigma 01}, \psi_{\sigma 10}, \psi_{\sigma 11} \dots \right\} \quad (4)$$

Kernels of convolutional layers are parametrized by trainable weights w in the following way:

$$\kappa_\sigma = \sum_i \Psi_{\sigma i} w_i \quad (5)$$

As a result, each kernel is defined on multiple scales and no image interpolation is used. Given a function of scale and translation $f(s, t)$ and a kernel $\kappa_\sigma(s, t)$, a scale convolution is defined as:

$$[f \star_H \kappa_\sigma](s, t) = \sum_{s'} [f(s', \cdot) \star \kappa_{s \cdot \sigma}(s^{-1}s', \cdot)](t) \quad (6)$$

The result of this operation is a stack of features each of which corresponds to a different scale. We end up with a 3-dimensional representation of the signal — 2-dimensional translation + scale. We follow [28] and denote scale-convolution as \star_H in order to distinguish it with the standard one. Figure 2 demonstrates how a kernel basis is formed and how scale-convolutional layers work.

Fast 1×1 Scale-Convolution An essential building block of many backbone deep networks such as ResNets [13] and Wide ResNets [36] is a 1×1 convolutional layer. We follow the interpretation of these layers proposed in [23] — it is a linear combination of channels. Thus, it has no spatial resolution. In order to build a scale-equivariant counterpart of 1×1 convolution, we do not utilize a kernel basis. As we pointed out before, the signal is stored as a 3 dimensional tensor for each channel. Therefore, for a kernel defined on N_S scales, the convolution of the signal with this kernel is just a 3-dimensional convolution with a kernel of size 1×1 in spatial dimension, and with N_S values in depth. This approach for 1×1 scale-convolution is faster than the special case of the algorithm proposed in [28].

Padding Although zero padding is a standard approach in image classification for saving the spatial resolution of the image, it worsens the localization properties of convolutional trackers [21, 38]. Nevertheless, a simple replacement of standard convolutional layers with scale-equivariant ones in very deep models is not possible without padding. Scale-equivariant convolutional layers have kernels of a bigger spatial extent because they are defined on multiple scales. For these reasons, we use circular padding during training and zero padding during testing in our models.

The introduced padding does not affect the feature maps which are obtained with kernels defined on small scales. It does not violate the translation equivariance of a network. We provide an experimental proof in supplementary material.

Scale-Pooling In order to capture correlations between different scales and to transform a 3-dimensional signal into a 2-dimensional one, we utilize global max pooling along the scale axis. This operation does not eliminate the scale-equivariant properties of the network. We found that it is useful to additionally incorporate this module in the places where conventional CNNs have spatial max pooling or strides. The mechanism of scale-pooling is illustrated in Figure 2.

Non-parametric Scale-Convolution The convolutional operation which results in the heatmap of a tracker is non-parametric. Both the input and the kernel come from neural networks. Thus, the approach described in Equation 6

is not suitable for this case. Given two functions f_1, f_2 of scale and translation the non-parametric scale convolution is defined as follows:

$$[f_1 \star_H f_2](s, t) = L_{s^{-1}}[L_s[f_1] \star f_2](t) \quad (7)$$

Here L_s is rescaling implemented as bicubic interpolation. Although it is a relatively slow operation, it is used only once in the tracker and does not heavily affect the inference time. The proof of the equivariance of this convolution is provided in supplementary material.

3.3. Extending a Tracker to Scale Equivariance

We present a recipe to extend a tracker to scale equivariance.

1. The first step is to estimate to what degree objects change in size in this domain, and then to select a set of scales $\sigma_1, \sigma_2, \dots, \sigma_N$. This is a domain-specific hyperparameter. For example, a domain with significant scale variations requires a broader span of scales, while for more smooth sequences, the set may consist of just 3 scales around 1.
2. For a tracker which can be described by Equation 1, derive ϕ_X and ϕ_Z .
3. For the networks represented by ϕ_X and ϕ_Z , all convolutional layers need to be replaced with scale-convolutional layers. The basis for these layers is based on the chosen scales $\sigma_1, \sigma_2, \dots, \sigma_N$.
4. (Optional) Scale-pooling can be included to additionally capture inter-scale correlations between all scales.
5. The connection operation \star needs to be replaced with a non-parametric scale-convolution.
6. (Optional) If the tracker only searches over spatial locations, scale-pooling needs to be included at the very end.

The obtained tracker produces a heatmap $h(z, x)$ defined on scale and translation. Therefore, each position is assigned a vector of features that has both the measure of similarity and the scale relation between the candidate and the template. If additional scale-pooling is included, then all scale information is just aggregated in the similarity score.

Note that the overall structure of the tracker, as well as the training and inference procedures are not changed. Thus, the recipe allows for a simple extension of a tracker with little cost of modification.

4. Scale-Equivariant SiamFC

While the proposed algorithm is applicable to a wide range of trackers, in this work, we focus on Siamese trackers. As a baseline we choose SiamFC [2]. This model serves as a starting point for modifications for the many modern high-performance Siamese trackers.

4.1. Architecture

Given the recipe, here we discuss the actual implementation of the scale-equivariant SiamFC tracker (SE-SiamFC).

In the first step of the recipe, we assess the range of scales in the domain (dataset). In sequences presented in most of the tracking benchmarks, like OTB or VOT, objects change their size relatively slowly from one frame to the other. The maximum scale change usually does not exceed a factor of $1.5 - 2$. Therefore, we use 3 scales with a step of $\sqrt{2}$ as the basis for the scale-convolutions. The next step in the recipe is to represent the tracker as it is done in Equation 1. SiamFC localizes the object as the coordinate *argmax* of the heatmap $h(z, x) = \phi_Z(z) \star \phi_X(x)$, where $\phi_Z = \phi_X$ are convolutional Siamese backbones. Next, in step number 3, we modify the backbones by replacing standard convolutions by scale-equivariant convolutions. We follow step 4 and utilize scale-pooling in the backbones in order to capture additional scale correlations between features of various scales. According to step 5, the connecting correlation is replaced with non-parametric scale-convolution. SiamFC computes its similarity function as a 2-dimensional map, therefore, we follow step 6 and add extra scale-pooling in order to transform a 3-dimensional heatmap into a 2-dimensional one. Now, we can use exactly the same inference algorithm as in the original paper [2]. We use the standard approach of scale estimation, based on the greedy selection of the best similarity for 3 different scales.

4.2. Weight Initialization

An important ingredient of a successful model training is the initialization of its weights. A common approach is to use weights from an Imagenet [9] pre-trained model [22, 38, 21]. In our case, however, this requires additional steps, as there are no available scale-equivariant models pre-trained on the Imagenet. We present a method for initializing a scale-equivariant model with weights from a pre-trained conventional CNN. The key idea is that a scale-equivariant network built according to Section 3.3 contains a sub-network that is identical to the one of the non-scale-equivariant counterpart. As the kernels of scale-equivariant models are parameterized with a fixed basis and trainable weights, our task is to initialize these weights.

We begin by initializing the inter-scale correlations by setting to 0 all weights responsible for these connections.

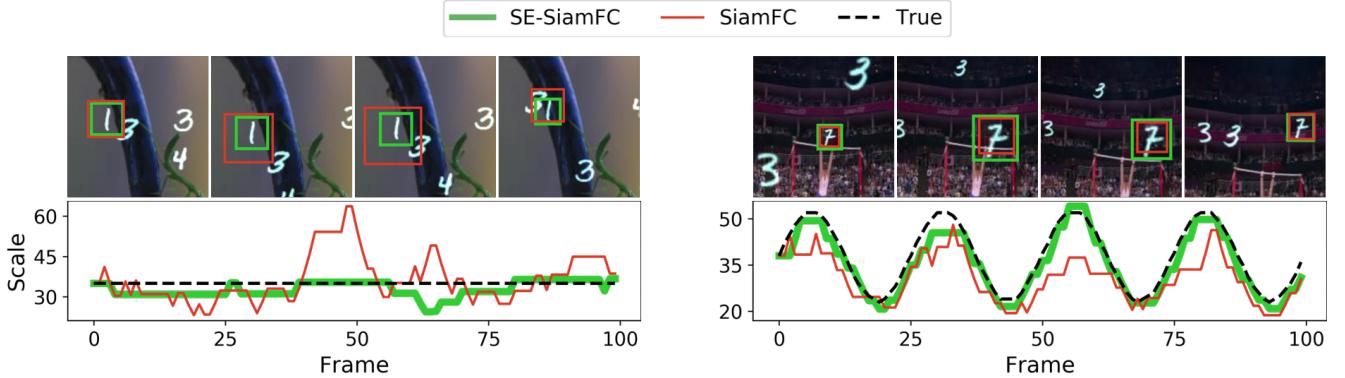


Figure 3: Top: examples of simulated T-MNIST and S-MNIST sequences. Bottom: scale estimation for equivariant and non-equivariant models. In the S-MNIST example, SE-SiamFC can estimate the scale more accurately. In the T-MNIST example, our model better preserves the scale of the target unchanged, while the non-scale-equivariant model is prone to oscillations in its scale estimate.

At this moment, up to scale-pooling, the scale-equivariant model consists of several networks parallel to, yet disconnected from one another, where the only difference is the size of their filters. For the convolutional layers with a non-unitary spatial extent, we initialize the weights such that the kernels of the smallest scale match those of the source model. Given a source kernel $\kappa'(u, v)$ and a basis $\Psi_{\sigma i}(u, v)$ with $\sigma = 1$, weights w_i are chosen to satisfy the linear system derived from Equation 5:

$$\kappa_1(u, v) = \sum_i \Psi_{1i}(u, v) w_i = \kappa'(u, v), \quad \forall u, v \quad (8)$$

As the basis is complete by construction, its matrix form is invertible. The system has a unique solution with respect to w_i :

$$w_i = \sum_{u, v} \Psi_{1i}^{-1}(u, v) \kappa'(u, v) \quad (9)$$

All 1×1 scale-convolutional layers are identical to standard 1×1 convolutions after zeroing out inter-scale correlations. We copy these weights from the source model. We provide an additional illustration of the proposed initialization method in the supplementary material.

5. Experiments and Results

5.1. Translation-Scaling MNIST

To test the ability of a tracker to cope with translation and scaling, we conduct an experiment on a simulated dataset with controlled factors of variation. We construct the datasets of translating (T-MNIST) and translating-scaling (S-MNIST) digits.

In particular, to form a sequence, we randomly sample up to 8 MNIST digits with backgrounds from the GOT10k

Tracker	T/T	T/S	S/T	S/S	# Params
SiamFC	0.64	0.62	0.64	0.63	999 K
SE-SiamFC	0.76	0.69	0.77	0.70	999 K

Table 1: AUC for models trained on T-MNIST and S-MNIST. T/S indicates that the model was trained on T-MNIST and tested on S-MNIST datasets. Bold numbers represent the best result for each of the training/testing scenarios.

dataset [16]. Then, on each of the digits in the sequence independently, a smoothed Brownian motion model induces a random translation. Simultaneously, for S-MNIST, a smooth scale change in the range $[0.67, 1.5]$ is induced by the sine rule:

$$s_i(t) = \frac{h - l}{2} \left[\sin\left(\frac{t}{4} + \beta_i\right) + 1 \right] + l \quad (10)$$

where $s_i(t)$ is the scale factor of the i -th digit in the t -th frame, h, l are upper and lower bounds for scaling, and $\beta_i \in [0, 100]$ is a phase, sampled randomly for each of the digits. In total, we simulate 1000 sequences for training and 100 for validation. Each sequence has a length of 100 frames. We compare two configurations of the tracker: (i) SiamFC with a shallow backbone and (ii) its scale-equivariant version SE-SiamFC. We conduct the experiments according to 2×2 scenarios: the models are trained on either S-MNIST or T-MNIST and are subsequently tested on either of them. The results are listed in Table 1. See supplementary material for a detailed description of the architecture, training, and testing procedures.

As can be seen from Table 1, the equivariant version outperforms its non-equivariant counterpart in all scenarios.

Tracker	Year	OTB-2013		OTB-2015		VOT2016			VOT2017		
		AUC	Prec.	AUC	Prec.	EOA	A	R	EOA	A	R
SINT [29]	2016	0.64	0.85	-	-	-	-	-	-	-	-
SiamFC [2]	2016	0.61	0.81	0.58	0.77	0.24	0.53	0.46	0.19	0.50	0.59
DSiam [12]	2017	0.64	0.81	-	-	-	-	-	-	-	-
StructSiam [37]	2018	0.64	0.88	0.62	0.85	0.26	-	-	-	-	-
TriSiam [10]	2018	0.62	0.82	0.59	0.78	-	-	-	0.20	-	-
SiamRPN [22]	2018	-	-	0.64	0.85	0.34	0.56	0.26	0.24	0.49	0.46
SiamFC+ [38]	2019	0.67	0.88	0.64	0.85	0.30	0.54	0.38	0.23	0.50	0.49
SE-SiamFC	Ours	0.68	0.90	0.66	0.88	0.36	0.59	0.24	0.27	0.54	0.38

Table 2: Performance comparisons on OTB-2013, OTB-2015, VOT2016, and VOT2017 benchmarks. Bold numbers represent the best result for each of the benchmarks.

The experiment on S-MNIST, varying the scale of an artificial object, shows that the scale-equivariant model has a superior ability to precisely follow the change in scale compared to the conventional one. The experiment on T-MNIST shows that (proper) measurement of scale is important even in the case when the sequence does not show a change in scale, where the observed scale in SE-SiamFC fluctuates much less than it does in the baseline (see Figure 3).

5.2. Benchmarking

We compare the scale-equivariant tracker against a non-equivariant baseline on popular tracking benchmarks. We test SE-SiamFC with a backbone from [38] against other popular Siamese trackers on OTB-2013, OTB-2015, VOT2016, and VOT2017. The benchmarks are chosen to allow direct comparison with the baseline [38]. We compare against the results published in the original paper. Although additional results are presented online¹, we couldn’t reproduce them in a reasonable amount of time.

Implementation details The parameters of our model are initialized with weights pre-trained on Imagenet by a method described in Section 4.2. We use the same training procedure as in the baseline. See supplementary material for a detailed description of the architecture.

The pairs for training are collected from the GOT10k [16] dataset. We adopt the same prepossessing and augmentation techniques as in [38]. The inference procedure remains unchanged compared to the baseline.

OTB We test on the OTB-2013 [34] and OTB-2015 [35] benchmarks. Each of the sequences in the OTB datasets carries labels from 11 categories of difficulty in tracking the sequence. Examples of these labels include: occlusion, scale variation, in-plane rotation, *etc.* We employ a standard one-pass evaluation (OPE) protocol to compare our method with

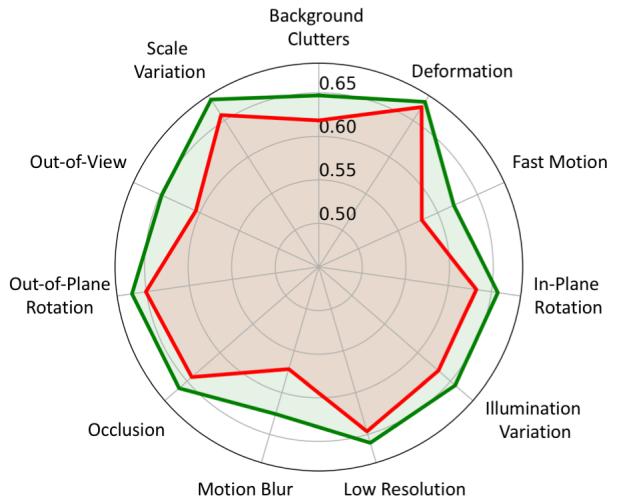


Figure 4: Comparison of AUC on OTB-2013 with different factors of variations. The red polygon corresponds to the baseline SiamFC+ and the green polygon — to SE-SiamFC.

other trackers by the area under the success curve (AUC) and precision.

The results are reported in Table 2. Our scale-equivariant tracker outperforms its non-equivariant counterpart by more than 3% on OTB-2015 in both AUC and precision, and by 1.4% on OTB-2013. When summarized at each label of difficulty (see Figure 4), the proposed scale-equivariant tracker is seen to improve all sequence types, not only those labeled with “scale variation”.

We attribute this to the fact that the “scale variation” tag in the OTB benchmark only indicates the sequences with a relatively big change in scale factors, while up to a certain degree, scaling is present in almost any video sequence. Moreover, scaling may be present implicitly, in the form of

¹<https://github.com/researchmm/SiamDW>

Model	Pretrained weights	AUC
SiamFC+ (aug. 5%)	✓	0.668
SiamFC+ (aug. 20%)	✓	0.668
SiamFC+ (aug. 50%)	✓	0.664
SE-SiamFC $\sigma = 1.2$	✓	0.677
SE-SiamFC $\sigma = 1.3$	✓	0.680
SE-SiamFC $\sigma = 1.4$	✓	0.681
SE-SiamFC $\sigma = 1.5$	✓	0.678
SE-SiamFC $\sigma = 1.4$	✗	0.553

Table 3: Ablation study on the OTB-2013 benchmark. The parameter σ stands for the step between scales in scale-equivariant models. Bold numbers represent the best result.

the same patterns being observed on multiple scales. An ability of our model to exploit this leads to better utilization of trainable parameters and a more discriminative Siamese similarity as a result.

VOT We next evaluate our tracker on VOT2016 and VOT2017 datasets [19]. The performance is evaluated in terms of average bounding box overlap ratio (A), and the robustness (R). These two metrics are combined into the Expected Average Overlap (EAO), which is used to rank the overall performance.

The results are reported in Table 2. On VOT2016 our scale-equivariant model shows an improvement from 0.30 to 0.36 in terms of EAO, which is a 20% gain compared to the non-equivariant baseline. On VOT2017, the increase in EAO is 17%.

We qualitatively investigated the sequences with the largest performance gain and observed that the most challenging factor for our baseline is the rapid scaling of the object. Even when the target is not completely lost, the imprecise bounding box heavily influences the overlap with the ground truth and the final EAO. Our scale-equivariant model better adapts to the fast scaling and delivers tighter bounding boxes. We provide qualitative results in the supplementary material.

5.3. Ablation Study

We conduct an ablation study on the OTB-2013 benchmark to investigate the impact of scale step, weight initialization, and fast 1×1 scale-convolution. We also test the baseline SiamFC+ model with various levels of scale data augmentation during the training. We follow the same training and testing procedure as in Section 5.2 for all experiments. In the weight initialization experiment, however, we do not use gradual weights unfreezing, but train the whole model end-to-end from the first epoch.

Scale step We investigate the impact of scale step σ , which defines a set of scales our model operates on. We train and test SE-SiamFC with various scale steps. Results are shown in Table 3. It can be seen that the resulting method outperforms the baseline on a range of scale steps. We empirically found that $\sigma = 1.4$ achieves the best performance.

Scale data augmentation Data augmentation is a common way to improve model generalization over different variations. Since our method is focused on scale, we compare SE-SiamFC against a baseline trained with different levels of scale data augmentation. Our results indicate (Table 3) that scale augmentation does not improve the performance of the conventional non-equivariant tracker.

Weight Initialization We train and test SE-SiamFC model, where weights initialized randomly [11, 31]. As can be seen from Table 3, random initialization results in a 19% performance drop compared to the proposed initialization technique.

Fast 1×1 scale-convolution We compare the speed of 1×1 scale-convolution from [28] and the proposed fast implementation. Implementation from [28] requires $450 / 1650 \mu\text{s}$, while our implementation requires $67 / 750 \mu\text{s}$ for forward / backward pass respectively, which is more than 6 times faster. In our experiments, the usage of fast 1×1 scale-convolution results in $30 - 40\%$ speedup of a tracker.

6. Discussion

In this work, we argue about the usefulness of additional scale equivariance in visual object tracking for the purpose of enhancing Siamese similarity estimation. We present a general theory that applies to a wide range of modern Siamese trackers, as well as all the components to turn an existing tracker into a scale-equivariant version. Moreover, we prove that the presented components are both necessary and sufficient to achieve built-in scale-translation equivariance. We sum up the theory by developing a simple recipe for extending existing trackers to scale equivariance. We apply it to develop SE-SiamFC — a scale-equivariant modification of the popular SiamFC tracker.

We experimentally demonstrate that our scale-equivariant tracker outperforms its conventional counterpart on OTB and VOT benchmarks and on the synthetically generated T-MNIST and S-MNIST datasets, where T-MNIST is designed to keep the object at a constant scale, and S-MNIST varies the scale in a known manner.

The experiments on T-MNIST and S-MNIST show the importance of proper scale measurement for all sequences, regardless of whether they have scale change or not. For the standard OTB and VOT benchmarks, our tracker proves the power of scale equivariance. It is seen to not only improves

the tracking in the case of scaling, but also when other factors of variations are present (see Figure 4). It affects the performance in two ways: it prevents erroneous jumps to similar objects at a different size and it provides a better consistent estimate of the scale.

Acknowledgments

We thank Thomas Andy Keller, Konrad Groh, Zenglin Shi and Deepak Gupta for valuable comments, discussions and help with the project. We appreciate the help of Zhipeng Zhang and Houwen Peng in reproducing the experiments from [38].

References

- [1] Erik J Bekkers. B-spline cnns on lie groups. In *International Conference on Learning Representations*, 2020.
- [2] 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.
- [3] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.
- [4] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999, 2016.
- [5] Taco S Cohen and Max Welling. Steerable cnns. *arXiv preprint arXiv:1612.08498*, 2016.
- [6] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Atom: Accurate tracking by overlap maximization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [7] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Eco: Efficient convolution operators for tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6638–6646, 2017.
- [8] Martin Danelljan, Andreas Robinson, Fahad Shahbaz Khan, and Michael Felsberg. Beyond correlation filters: Learning continuous convolution operators for visual tracking. In *European conference on computer vision*, pages 472–488. Springer, 2016.
- [9] 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.
- [10] 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.
- [11] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. volume 9 of *Proceedings of Machine Learning Research*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.
- [12] Qing Guo, Wei Feng, Ce Zhou, Rui Huang, Liang Wan, and Song Wang. Learning dynamic siamese network for visual object tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1763–1771, 2017.
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [14] David Held, Sebastian Thrun, and Silvio Savarese. Learning to track at 100 fps with deep regression networks. In *European Conference Computer Vision (ECCV)*, 2016.
- [15] Emiel Hoogeboom, Jorn WT Peters, Taco S Cohen, and Max Welling. Hexaconv. *arXiv preprint arXiv:1803.02108*, 2018.
- [16] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–1, 2019.
- [17] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yunling Jiang. Acquisition of localization confidence for accurate object detection. In *The European Conference on Computer Vision (ECCV)*, September 2018.
- [18] Risi Kondor and Shubhendu Trivedi. On the generalization of equivariance and convolution in neural networks to the action of compact groups. *arXiv preprint arXiv:1802.03690*, 2018.
- [19] M. Kristan, A. Leonardis, J. Matas, and M. Felsberg et. The visual object tracking vot2017 challenge results. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 1949–1972, 2017.
- [20] Dmitry Laptev, Nikolay Savinov, Joachim M Buhmann, and Marc Pollefeys. Ti-pooling: transformation-invariant pooling for feature learning in convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 289–297, 2016.
- [21] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. *arXiv preprint arXiv:1812.11703*, 2018.
- [22] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8971–8980, 2018.
- [23] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.
- [24] Diego Marcos, Benjamin Kellenberger, Sylvain Lobry, and Devis Tuia. Scale equivariance in cnns with vector fields. *arXiv preprint arXiv:1807.11783*, 2018.
- [25] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.
- [26] David W Romero, Erik J Bekkers, Jakub M Tomczak, and Mark Hoogendoorn. Attentive group equivariant convolutional networks. *arXiv preprint arXiv:2002.03830*, 2020.

- [27] David W Romero and Mark Hoogendoorn. Co-attentive equivariant neural networks: Focusing equivariance on transformations co-occurring in data. *arXiv preprint arXiv:1911.07849*, 2019.
- [28] Ivan Sosnovik, Michał Szmaja, and Arnold Smeulders. Scale-equivariant steerable networks. *arXiv preprint arXiv:1910.11093*, 2019.
- [29] Ran Tao, Efstratios Gavves, and Arnold WM Smeulders. Siamese instance search for tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1420–1429, 2016.
- [30] Maurice Weiler and Gabriele Cesa. General e (2)-equivariant steerable cnns. In *Advances in Neural Information Processing Systems*, pages 14334–14345, 2019.
- [31] 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.
- [32] Daniel Worrall and Max Welling. Deep scale-spaces: Equivariance over scale. In *Advances in Neural Information Processing Systems*, pages 7364–7376, 2019.
- [33] Daniel E Worrall, Stephan J Garbin, Daniyar Turmukhametov, 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.
- [34] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.
- [35] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015.
- [36] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.
- [37] Yunhua Zhang, Lijun Wang, Jinqing Qi, Dong Wang, Mengyang Feng, and Huchuan Lu. Structured siamese network for real-time visual tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 351–366, 2018.
- [38] 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.
- [39] 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.

A. Proofs

A.1. Convolution is all you need

In the paper we consider trackers of the following form

$$h(z, x) = \phi_X(x) \star \phi_Z(z) \quad (11)$$

where ϕ_X and ϕ_Z are parameterized with feed-forward neural networks.

Theorem 2. *A function given by Equation 11 is equivariant under a transformation L from group G if and only if ϕ_X and ϕ_Z are constructed from G -equivariant convolutional layers and \star is the G -convolution.*

Proof. Let us fix $z = z_0$ and introduce a function $h_X = h(x, z_0) = \phi_X(x) \star \phi_Z(z_0)$. This function is a feed-forward neural network. All its layers but the last one are contained in ϕ_X and the last layer is a convolution with $\phi_Z(z_0)$. According to [18] a feed-forward neural network is equivariant under transformations from G if and only if it is constructed from G -equivariant convolutional layers. Thus, the function h_X is equivariant under transformations from G if and only if

- The function ϕ_X is constructed from G -equivariant convolutional layers
- The convolution \star is the G -convolution

If we then fix $x = x_0$, we can show that a function $h_Z = h(x_0, z) = \phi_X(x_0) \star \phi_Z(z)$ is equivariant under transformations from G if and only if

- The function ϕ_Z is constructed from G -equivariant convolutional layers
- The convolution \star is the G -convolution

The function h is equivariant under G if and only if both the function h_X and the function h_Z are equivariant. \square

A.2. Non-parametric scale-convolution

Given two functions f_1, f_2 of scale and translation the non-parametric scale convolution is defined as follows:

$$[f_1 \star_H f_2](s, t) = L_{s^{-1}}[L_s[f_1] \star f_2](t) \quad (12)$$

Lemma 1. *A function given by Equation 12 is equivariant under scale-translation.*

Proof. A function given by Equation 12 is equivariant under scale transformations of f_1 , indeed

$$\begin{aligned} [L_{\hat{s}}[f_1] \star_H f_2](s, t) &= L_{s^{-1}}[L_{s\hat{s}}[f_1] \star f_2](t) \\ &= L_{\hat{s}}L_{(s\hat{s})^{-1}}[L_{s\hat{s}}[f_1] \star f_2](t) \quad (13) \\ &= L_{\hat{s}}[f_1 \star_H f_2](s\hat{s}, t) \end{aligned}$$

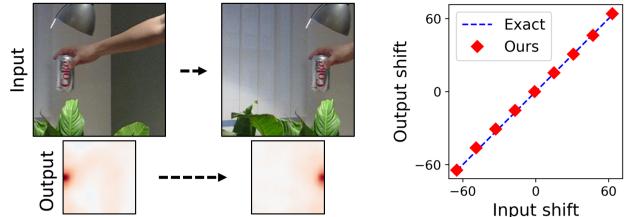


Figure 5: Left: two samples from the simulated sequence. The input image is a translated and cropped version of the source image. The output is the heatmap produced by the proposed model. The red color represents the place where the object is detected. Right: correspondence between the input and the output shifts.

For a pair of scale and translation s, \hat{t} we have the following property of the joint transformation $L_s T_{\hat{t}} = T_{\hat{t}s} L_s$ from [28], where $T_{\hat{t}}$ is the translation operator defined as $T_{\hat{t}}[f](t) = f(t - \hat{t})$. Now we can show the following:

$$\begin{aligned} [T_{\hat{t}}[f_1] \star_H f_2](s, t) &= L_{s^{-1}}[L_s[T_{\hat{t}}[f_1]] \star f_2](t) \\ &= L_{s^{-1}}[T_{\hat{t}s}L_s[f_1] \star f_2](t) \\ &= L_{s^{-1}}T_{\hat{t}s}[L_s[f_1] \star f_2](t) \quad (14) \\ &= T_{\hat{t}}L_{s^{-1}}[L_s[f_1] \star f_2](t) \\ &= T_{\hat{t}}[f_1 \star_H f_2](t) \end{aligned}$$

Therefore, a function given by Equation 12 is also equivariant under translations of f_1 . The equivariance of the function with respect to a joint transformation follows from the equivariance to each of the transformations separately [28].

We proved the equivariance with respect to f_1 . The proof with respect to f_2 is analogous. \square

B. Weight initialization

The proposed weight initialization scheme from a pre-trained model is depicted in Figure 6.

C. Experiments

C.1. Padding

We conduct an experiment to verify that the proposed padding technique does not violate translation equivariance of convolutional trackers. We choose an image and select a sequence of translated and cropped windows inside of it. We process this sequence with a deep model that consists of the proposed convolutional layers and follows the inference procedure described in [38]. We derive the predicted location of the object and compare its value to the input shift. Figure 5 demonstrates that the input and the output translations have nearly identical values.

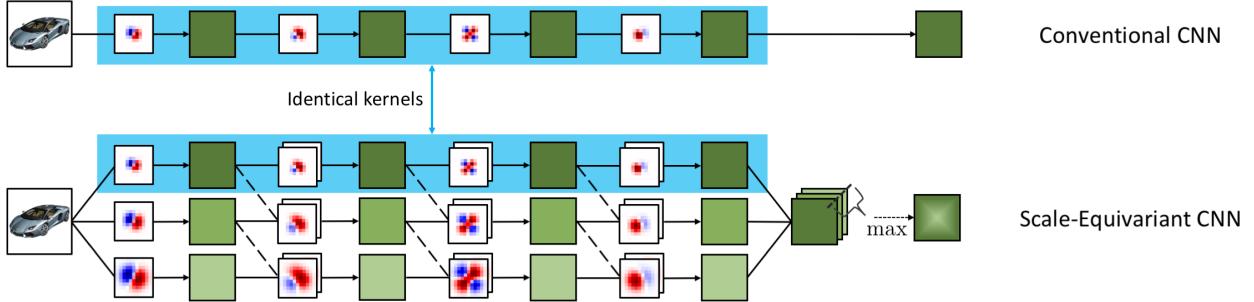


Figure 6: The visualization of the weight initialization scheme from a pretrained model. Dashed connections are initialized with 0.

C.2. Translating-Scaling MNIST

Stage	SiamFC	SE-SiamFC
Conv1	$[3 \times 3, 96, s = 2]$	
Conv2	$[3 \times 3, 128, s = 2]$	
Conv3	$[3 \times 3, 256, s = 2]$	
Conv4	$[3 \times 3, 256, s = 1]$	
Connect.	Cross-correlation	Non-parametric scale-convolution
# Params	999 K	999 K

Table 4: Architectures used in T/S-MNIST experiment. All convolutions in SE-SiamFC are scale-convolutions.

For both T-MNIST and S-MNIST, we use architectures described in Table 4. 2D BatchNorm and ReLU are inserted after each of the convolutional layers except the last one. We do not use max pooling to preserve strict translation-equivariance.

We train both models for 50 epochs using SGD with a mini-batch of 8 images and exponentially decay the learning rate from 10^{-2} to 10^{-5} . We set the momentum to 0.9 and the weight decay to 0.5^{-4} . A binary cross-entropy loss as in [2] is used. The inference algorithm is the same for both SiamFC and SE-SiamFC and follows the original implementation [2].

C.3. OTB and VOT

For OTB and VOT experiments we used architectures described in Table 5. We use the baseline [38] with Cropping Inside Residual (CIR) units. SE-SiamFC is constructed directly from the baseline as described in the paper.

In Table 5 the kernel size refers to the smallest scale $\sigma = 1$ in the network. The sizes of the kernels, which correspond to bigger scales are 9×9 for Conv1 and 5×5 for other layers. Figure 7 gives a qualitative comparison of the proposed method and the baseline.

Stage	SiamFC+	SE-SiamFC
Conv1	$[7 \times 7, 64, s = 2]$	$[7 \times 7, 64, s = 2]$
	max pool $[2 \times 2, s = 2]$	
Conv2	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 64, i = 2 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
Conv3	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 128, sp \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 3$
Connect.	Cross-correlation	Non-parametric scale-convolution
# Params	1.44 M	1.45 M

Table 5: Architectures used in OTB/VOT experiments. All convolutions in SE-SiamFC are scale-convolutions. s refers to stride, sp denotes scale pooling, i — is the size of the kernel in a scale dimension.

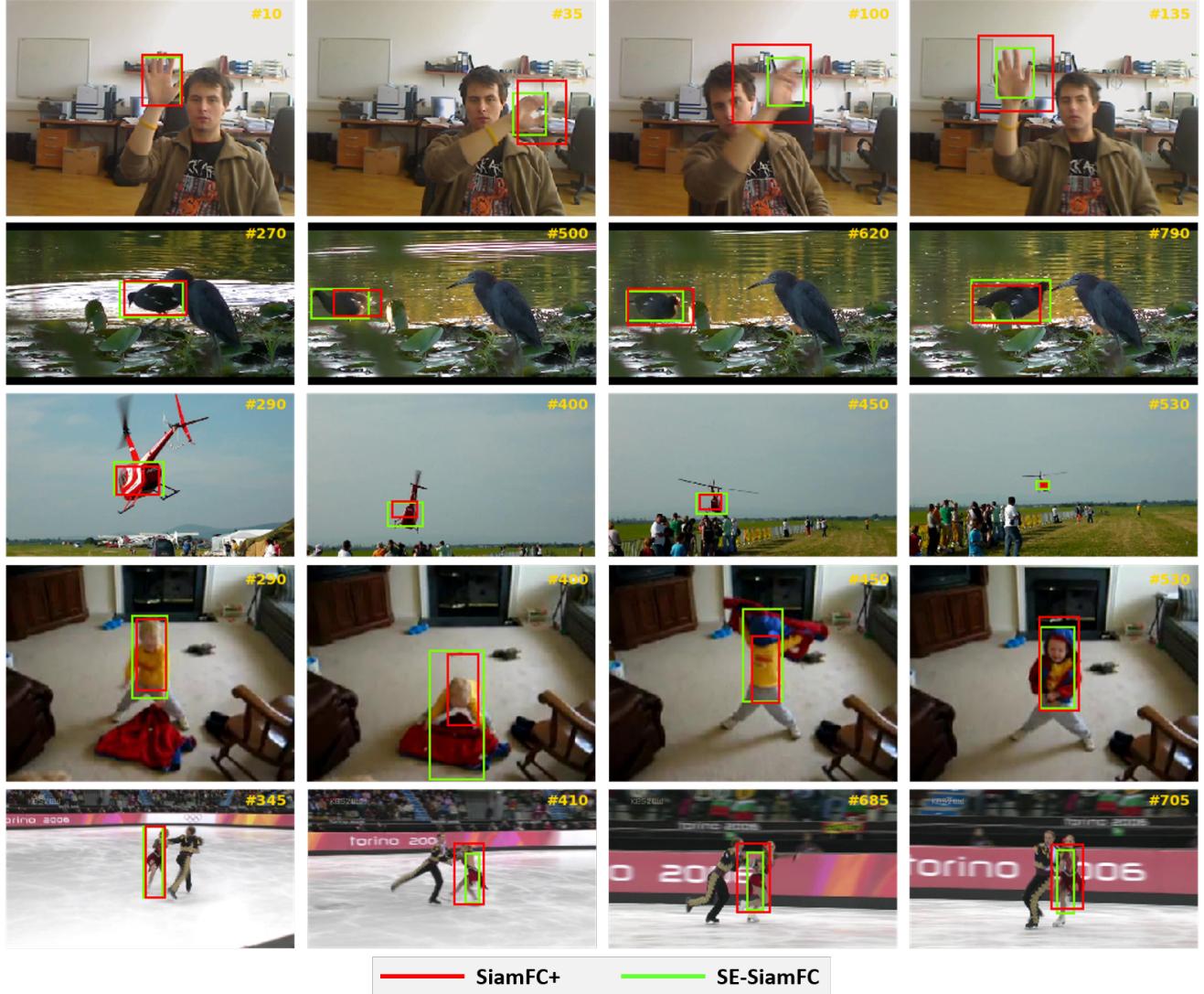


Figure 7: Qualitative comparison of SE-SiamFC with SiamFC+ on VOT2016/2017 sequences.