Rotation Equivariant Siamese Networks for Tracking

Deepak K. Gupta*, Devanshu Arya†, Efstratios Gavves*
*QUVA Lab, University of Amsterdam, The Netherlands
†Informatics Institute, University of Amsterdam, The Netherlands
{d.k.gupta,d.arya,e.gavves}@uva.nl

Abstract

Rotation is among the long prevailing, yet still unresolved, hard challenges encountered in visual object tracking. The existing deep learning-based tracking algorithms use regular CNNs that are inherently translation equivariant, but not designed to tackle rotations. In this paper, we first demonstrate that in the presence of rotation instances in videos, the performance of existing trackers is severely affected. To circumvent the adverse effect of rotations, we present rotation-equivariant Siamese networks (RE-SiamNets), built through the use of group-equivariant convolutional layers comprising steerable filters. SiamNets allow estimating the change in orientation of the object in an unsupervised manner, thereby facilitating its use in relative 2D pose estimation as well. We further show that this change in orientation can be used to impose an additional motion constraint in Siamese tracking through imposing restriction on the change in orientation between two consecutive frames. For benchmarking, we present Rotation Tracking Benchmark (RTB), a dataset comprising a set of videos with rotation instances. Through experiments on two popular Siamese architectures, we show that RE-SiamNets handle the problem of rotation very well and outperform their regular counterparts. Further, RE-SiamNets can accurately estimate the relative change in pose of the target in an unsupervised fashion, namely the in-plane rotation the target has sustained with respect to the reference frame. Code and data can be accessed at https://github.com/dkgupta90/re-siamnet.

1. Introduction

The task of visual object tracking with Siamese networks [1, 29], also referred as Siamese tracking, transforms the problem of tracking into similarity estimation between a template frame and sampled regions from a candidate frame. Siamese trackers have recently gained popularity in the field of visual object tracking, especially because of their strong discriminative power obtained from similarity matching. This is the primary reason most of the state-of-the-art trackers are based on this framework [1, 11, 18, 19, 29].

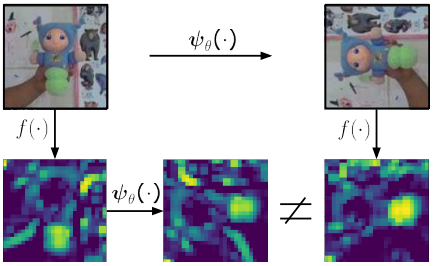

Figure 1: Example demonstrating rotation non-equivariance in regular CNN models used in object tracking, $\psi_\theta(f(\cdot)) \neq f(\psi_\theta(\cdot))$. Here $f(\cdot)$ and $\psi_\theta(\cdot)$ denote the neural network encoding function and rotation transform, respectively.

Although Siamese trackers are generally shown to work well, they are still prone to failure under challenges such as partial occlusion [16], scale change [27] or when one of the two inputs is rotated.

This paper focuses on handling the adverse affects of in-plane rotation of objects on the performance of Siamese trackers. Object rotation is considered to be amongst the hardest challenges of tracking with no effective solution till date. It can commonly occur in real-life scenarios, especially when the camera records from the top, as in drones, where either the object is rotating or the camera itself. Egocentric videos are another example, where large head rotations can cause the target to rotate.

The CNN architectures used in Siamese trackers are not inherently equivariant to in-plane rotations of the target. The implication is that the model may perform well on object orientations that are represented in the training set, but may fail on other previously unseen orientations. This happens because the latent encoding obtained from the network for such cases might not be representative of the input image itself. Example demonstrating this issue is shown in Figure 1. Further, even if it were equivariant, the cross-

correlation step in traditional Siamese trackers would still fail to perform an accurate matching between the template and candidate images due to rotational shift between them.

A straightforward approach to enforce learning of rotated variants is to use training datasets where in-plane rotations occur naturally or through data augmentation. However, as highlighted in [17], there are several limitations of data augmentation. First, such procedures would require learning separate representations for different rotated variants of the data. Second, the more variations are considered, the more flexible tracker model needs to be to capture them all. This means a significant increase in training data and computational budget. Further, such an approach would make the model invariant to rotations, thus making the predictions unreliable when the target is surrounded by similar objects, *e.g.*, tracking a fish in a school of fishes.

This paper aims at incorporating the property of rotation equivariance in the existing Siamese trackers. This built-in feature would then allow the trackers to capture the rotation variations from the start itself without the need of additional data augmentation. Rotation equivariant networks have been studied widely in the context of image classification [3, 4, 34, 35, 36]. Drawing inspiration from these works, we introduce rotation equivariance for the task of object localization in videos. We exploit the concept of group-equivariant CNNs [3], and use steerable filters [35] to make the Siamese trackers equivariant to rotations. This way of incorporating rotation equivariance induces built-in sharing of weights among the different groups of rotations and adds an internal notion of rotation in the model (referred further as *RE-SiamNet*).

Interpreting the template image as the static memory of the tracking model, RE-SiamNets know beforehand how the encoding should be represented for a discrete set of rotations. In the absence of other challenges such as illumination variation and occlusion, the target appearance would match exactly at one of the discrete rotations, and is expected to contain only small errors for other intermediate angles. This property increases the discriminative power of the trackers towards differences in orientation (in-plane rotation) of the target. Beyond this, RE-SiamNet can be used for relative 2D pose estimation of objects in videos, interchangeably also referred in this paper as relative orientation estimation of objects. RE-SiamNets are equivariant to translations and rotations, and these properties combined with the structure of Siamese networks allow capturing the change in pose of the target in 2D. Further, we propose an additional motion constraint on the rotational degree of freedom and demonstrate that it allows to obtain better temporal correspondence in videos.

It is important to note that most current datasets, especially in tracking, contain very limited to no instances of rotation. Thus, for benchmarking the performance of models in presence of in-plane rotations, we present Rotating Object Benchmark (ROB), a set of videos focusing on in-plane rotations. Annotations include bounding boxes of the target object as well as its orientation in every frame. To further summarize, the contributions of this paper are:

- We give a brief introduction to equivariant convolutions networks. We then extend the theory to obtain rotation-equivariant Siamese architectures (RE-SiamNets) that feature in-plane rotation equivariance.
- We show that RE-SiamNets estimate the relative 2D pose of any rotating object in a unsupervised manner. Further, we introduce an additional motion constraint to improve temporal correspondence in videos.
- For benchmarking, we present Rotating Object Benchmark (ROB), a novel dataset comprising sequences with significant in-plane rotations of the target.
- Through incorporating in two existing Siamese tracking methods, we show that rotation equivariance can provide significant improvements in tracking performance and accurately estimate the orientation changes.

2. Related Work

Siamese tracking. Object tracking aims at estimating the trajectory of an arbitrary target in a video given only its initial state in a video frame [15]. Most of the recent object tracking algorithms use Siamese networks and track the object based on similarity matching [6, 8, 10, 25, 31, 33, 38]. Such algorithms estimate a general similarity function between the feature representations learned for the target template and the candidate search region in a given frame.

The first Siamese trackers, SINT [29] and SiamFC [1], used twin subnetworks with shared parameters and calculated dot product similarities between the feature maps of the template and the candidate frame. Held et al. [13] introduced a detection-based Siamese tracker in which the similarity function was modeled as a fully-connected network. They applied extensive data augmentation for learning a generalized function for multiple object transformations. Valmadre et al. [30] introduced CFNet which expanded SiamFC using a differentiable correlation filter layer. All of these trackers were able to get good performance in terms of object deformation compared with the trackers without online updating, but were not suitable in fast tracking situations. Some of the subsequent methods such as [12, 19, 32, 40] discarded online updating, and turned to learn a robust feature representation instead. This allowed the aforementioned methods to perform high speed tracking using Siamese networks.

Challenges of tracking. There are several challenges encountered in visual object tracking that can affect the performance of the designed tracking algorithms. A detailed study highlighting some of the most important challenges

was presented in [26]. These include illumination variation, in-plane and out-of-plane rotations of the target, occlusion, clutter and confusion due to several similar objects, among others. With recent large-scale training datasets such as LaSOT [7] and TrackingNet [22], and state-of-the-art deep learning trackers, several of these challenges can be addressed up to a high degree of accuracy. For example, trackers such as SiamRPN++ [18] and DiMP [2] exhibit strong discriminative power with the use of deep CNN backbones, and have been found to tackle most of the challenges. However, some challenges such as occlusion and target rotation still remain to be solved. Recent works related to tackling occlusion tracking are [11] and [16]. In this paper, we focus on the challenge of target rotation.

Equivariant CNNs. Recently, several works have tried to directly incorporate equivariance into the network's architecture to capture various transformations. In this paper, we focus on rotation-equivariant CNNs which have gained popularity in image classification [5, 4], texture classification [21], boundary detection [36] and image segmentation [17]. Dieleman et al. [5] included 4 operations into existing networks to enrich both the batch- and feature dimension with their transformed versions. Cohen et al. [3] firstly introduced group-convolutional layers where feature maps resulting from transformed filters were treated as functions of the corresponding symmetry-group. However, in this method the computational cost was directly proportional to the group size, and this issue was resolved with steerable filters [4, 35]. A detailed study providing a general theory of equivariance across various existing methods is provided in [34]. In this paper, we study rotation equivariance in the context of object tracking.

In real-life scenarios, tracking a target object is very challenging, especially since it can undergo transformations beyond translation, such as in-plane and out-of-plane rotations, occlusion and scale change. Unless the network has an internal mechanism to handle these transformations, the template matching similarity can degrade significantly in a Siamese network. Recent Siamese trackers [18, 39] have implicitly or explicitly focused on making the trackers translational equivariant, i.e. a translation of the input image must result in the proportional translation of the corresponding feature space. The importance of translation equivariance is to reduce the positional bias during training, so that location of the target is easier to recover from the feature space. SiamRPN++ [18] proposed a training strategy which removes the spatial bias introduced in non fully-convolutional backbones. Further, [39] showed that existing tracking models induce positional bias, which breaks strict translation equivariance. Sosnovik et al. [27] introduced scale-equivariant Siamese trackers which are crucial when the camera zooms its lens or when the target moves into depth. We argue that in-plane rotation is also an important challenge in tracking, especially when the videos are recorded using drone cameras, other videos recorded from top view, cameras mounted on rotating objects and egocentric videos. To the best of our knowledge, rotation equivariance in the context of tracking has never been studied, and we address it in this paper.

3. Rotation Equivariant CNNs

We first provide some basic background knowledge on equivariance and rotation equivariance in CNNs required to formulate our tracker. For a more general overview we refer the interested reader to [35].

Equivariance. The property of equivariance requires functions to commute with the actions of a symmetry group acting on its domain and co-domain. For any given transformation group G, a mapping function $f : X \rightarrow Y$ is equivariant if it satisfies

$$f(\psi_g^X(x)) = \psi_g^Y(f(x)) \;\; g \in G, x \in X, \tag{1}$$

where $\psi_g^{(\cdot)}$ denotes a group action in the respective space. For invariance, $\psi_g^Y(\cdot)$ will be an identity mapping.

For clarity, we take translation equivariance as an example. In this example, f stands for the convolutional neural network function and ψ_g denotes the translation group. Example actions from this group include for example, moving one pixel left, or one towards right, or an action comprising shift of several pixels. In this manner, an infinite number of actions can be defined within the translation group. Making the network equivariant to translations leads to reduced sample complexity and facilitates generalization of the model against translational variations.

It is important to note that there are several other transformations beyond translation that can be built in the model to improve robustness, if the effects of these transformations are present in the data and the task. Examples include rotations, reflections and scale change. For generalization over any of these transformations, equivariance needs to be enforced on the respective transformation group. In this work we focus on rotation equivariance.

Rotation equivariance. One of the more robust ways of enforcing rotation equivariance in CNNs is through the use of steerable filters [35]. Steerable filter CNNs (SFCNNs) extend the notion of weight sharing from the translation group to rotations as well. For rotation equivariance with steerable filters, the network must perform convolutions with different rotated versions of each filter. In this case weight sharing helps the model to generalize better.

Steerable filters not only facilitate efficiently computing responses for an arbitrary number of discrete filter rotations Λ, but they also exhibit strong expressive power as well. A filter Ψ is rotationally steerable if its rotation by an arbitrary angle θ can be expressed in terms of a fixed set of atomic

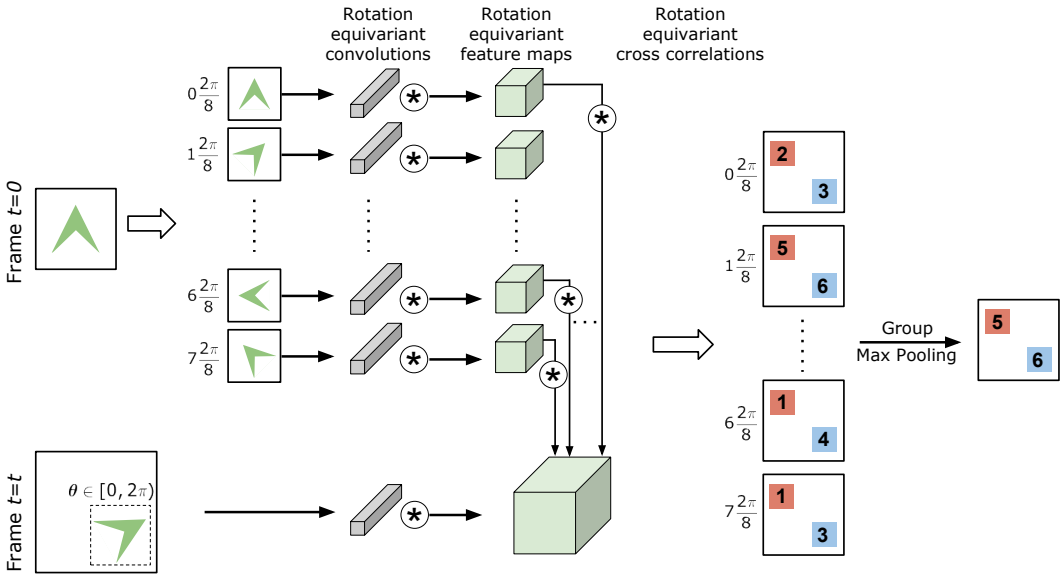

Figure 2: Schematic representation of RE-SiamNet typically designed for object tracking. On the template head, multiple equidistant rotated variants of the original template image are used.

functions [9, 35]. In our network, we employ circular harmonics ψ_{jk} defined as

$$\psi_{jk}(r, \phi) = \tau_j(r) e^{\mathrm{i}k\phi}, \tag{2}$$

where $\phi \in (-\pi, \pi]$ and $j = 1, 2, \ldots, J$ allows to control the radial part of the basis functions. Further, the (r, ϕ) refers to transformed version of (x_1, x_2) in polar coordinates and $k \in \mathbb{Z}$ denotes the angular frequency. The benefit of circular harmonics is that now we can simply express rotations on ψ_{jk} as a multiplication with a complex exponential,

$$\rho_\theta \psi_{jk}(x) = e^{-\mathrm{i}k\theta} \psi_{jk}(x). \tag{3}$$

Note that for clarity purpose, we express $\psi_{jk}(\cdot)$ as $\psi_{jk}(x)$.

Each learnt filter is then constructed as a linear combination of the elementary filters,

$$\Psi(x) = \sum_{j=1}^{J} \sum_{k=0}^{K} w_{jk} \psi_{jk}(x), \tag{4}$$

with weights $w_{jk} \in \mathbb{C}$. For rotation by θ, the composed filter can be steered through phase manipulation of the elementary filters,

$$\rho_\theta \Psi(x) = \sum_{j=1}^{J} \sum_{k=0}^{K} w_{jk} e^{-\mathrm{i}k\theta} \psi_{jk}(x). \tag{5}$$

A single orientation of the filter can be obtained by taking real part of Ψ, denoted as $\mathrm{Re}\Psi(x)$.

4. Rotation Equivariant Siamese Trackers

4.1. Proposed Formulation

For trackers that rely on similarity matching with Siamese networks, the resultant heatmap $h(z, x)$ is

$$h(z, x) = f(z) * f(x), \tag{6}$$

where z and x denote the template image and the candidate frames, respectively, $f(\cdot)$ is the encoding function of the Siamese network, and $*$ denotes the convolution operation.

Figure 2 presents the schematic representation of our RE-SiamNet framework for object tracking. Architecturally, we start from and modify the basic SiamFC [1] model due to its simple design. The basic SiamFC comprises the following modular layers: input, convolutional layers, and a cross-correlation of the outputs from the two Siamese heads. For our rotational Siamese tracker, we replace these layers with rotation equivariant modules. Further, we introduce a group max pooling module that selects the cross-correlation encoding for the most appropriate orientations among the multiple heatmaps generated in our setup. Details related to these modules follow below.

Rotation equivariant input. The candidate head of the network takes a single search image as input. However, the template head is modified to not just take one template image as an input, rather a set of its Λ rotated variants defined by the set Z, where $Z = \{z_1, z_2, \ldots, z_\Lambda\}$. Instead of taking all possible rotation versions Z of the template target, we could also first compute the feature $f(z)$ of the original target, then rotate $f(z)$. In theory, this is supported by rotation equivariant networks. In practice, however, the spatial

resolution of $f(z)$ is very low, typically 6×6 or 7×7 pixels. As a result, there will be artifacts at the corners and edges because of the crudeness of the transformation. Instead, it yields more accurate feature maps if, when creating Z in the first frame, we first rotate the whole frame (not just the target) centering about the target, and then crop. Since this is only performed on the target branch, it can be pre-computed during the inference phase.

Each input image I, as stated above, comprises C channels, where each channel is represented as I_c and $c \in \{1, 2, \ldots, C\}$. This input is then convolved with $\hat{C}$ rotated filters $\rho_\theta \Psi_{\hat{c}c}^{(1)}$, where $\hat{c} \in \{1, 2, \ldots, \hat{C}\}$. Based on Eq. 5, the resultant features obtained before applying nonlinear activation will be

$$y_{\hat{c}}^{(1)}(x,\theta) = \text{Re}\sum_{c=1}^{C}\sum_{j=1}^{J}\sum_{k=0}^{K_j} w_{\hat{c}cjk} e^{-\text{i}k\theta}(I_c * \psi_{jk})(x), \quad (7)$$

where the filters are then rotated variants at equidistant orientations θ represented by the set $\Theta = \{(i-1)\cdot 360/\Lambda\}_{i=1}^{\Lambda}$. The bias term $\beta_{\hat{c}}^{(1)}$ and nonlinearity σ are then applied to obtain the feature map at the first layer $\zeta_{\hat{c}}^{(1)}$.

Rotation equivariant convolutions. Feature maps resulting from Eq. 7 are processed further using group convolutions, generalizing spatial convolutions over a wider set of transformation groups. Similar to the first layer, steerable filters are defined on the group as

$$y_{\hat{c}}^{(l)}(x,\theta) = \text{Re}\sum_{c=1}^{C}\sum_{\phi\in\Theta}\sum_{j,k} w_{\hat{c}cjk,\theta-\phi} e^{-\text{i}k\theta}(\zeta_c^{(l-1)}(\cdot,\phi) * \psi_{jk})(x). \quad (8)$$

The additional index $\theta-\phi$ introduced in Eq. 8 for the weight tensor facilitates the group convolution operation along the rotation dimension. It involves transforming the functions on the group through rotating them spatially.

Rotation equivariant pooling. The output of the last group convolutional layer is further processed through pooling over the rotation dimension. Unlike the conventional classification tasks, pooling is not performed along the spatial dimension to preserve the rotation equivariance.

Rotation equivariant cross-correlation. From the two subnetworks of the Re-SiamNet module, we obtain two sets of feature maps, $\{\phi(z)\}$ and $\phi(x)$, where $\{\phi(z)\}$ is the set containing feature maps at Λ orientations. Next, $\{\phi(z)\}$ and $\phi(x)$ are convolved to obtain $\{\hat{h}(z,x)\}$, a set of Λ heatmaps, where $h_i(z,x) = \phi(z_i) * \phi(x)$. Next, $\{\hat{h}(z,x)\}$ is processed with a global maxpooling operation to obtain the final output heatmap $h(Z,x)$. The global maxpooling operation identifies the maximum value in $\{\hat{h}(z,x)\}$ and selects the feature map that contains it.

By introducing the aforementioned modules, we obtain the rotation equivariant Siamese tracker. Again, we emphasize that the tracker is equivariant to *in-plane rotations*, as out-of-plane rotations require knowledge of the 3D scene to be integrated in the network. Next, we describe the training and inference of rotation equivariant Siamese trackers.

4.2. Constructing RE-SiamNet Framework

We outline below the steps to design RE-SiamNet framework using the rotation equivariant modules described in the earlier section.

1. Identify the precision of the tracker in terms of discriminating between different orientations of the rotational degree of freedom. We consider here Λ rotation groups, based on which RE-SiamNets would be perfectly equivariant to angles defined by the set $\Theta = \{(i-1)\cdot 360/\Lambda\}_{i=1}^{\Lambda}$.
2. Define the non-parametric encoding $\phi(\cdot)$ based on existing Siamese trackers. Based on the choice of $\phi(\cdot)$, the discriminative power of trackers varies.
3. Replace all the convolutional layers of $\phi(\cdot)$ with the rotation-equivariant modules [1].
4. Instead of a single convolution to generate $h(z,x)$, Λ convolutions are performed to generate Λ different heatmaps.
5. Perform Global max-pooling over the feature maps to generate $h(Z,x)$, which is then processed to localize the target.

Note that depending on the choice of the tracker head, processing operation on $h(Z,x)$ can differ. For example, for trackers such as SINT [29] and SiamFC [1], target instance from the previous frame is fitted at different scales and aspect ratios, and the best among them is chosen. For other trackers such as SiamRPN [19] and SiamRPN++ [18], a region-proposal module is added that regresses the bounding box using a neural network head. In our tracking architecture, rotation equivariance needs to be only maintained up to $h(Z,x)$, thus it can work with any of these methods.

5. Unsupervised Relative Rotation Estimation

Unsupervised 2D pose estimation. The inherent design of RE-SiamNets allows to obtain an estimate of the relative change of 2D pose of the target in a fully unsupervised manner. This information can be obtained from the result of the group maxpooling step. Let $i \in \{1, 2, \ldots, \Gamma\}$ denote one of Λ orientations of the template image. Then, i is the number of rotation groups by which the pose of the template differs from that of its appearance in the candidate image, if:

$$h(Z,x) = \hat{h}(z_i,x) = \text{group-maxpool}(\{h(z,x)\}). \quad (9)$$

[1] For implementing rotation-equivariant modules in this paper, we use the `e2CNN` Pytorch library [34] available at `https://github.com/QUVA-Lab/e2cnn`

The difference in pose expressed in terms of rotational angle θ_{diff} is then $i \cdot 360/\Gamma$. Assuming that the actual in-plane rotation of the target is θ_c, the error in prediction in degrees is bounded as $|\theta_{\text{diff}} - \theta_c| \leq \frac{360}{2\Lambda}$. Thus, for larger values of Λ, error in the estimation of pose decreases.

Rotational Motion Consistency. An important advantage is that RE-SiamNets provide a novel motion constraint that can be used to improve temporal correspondence in object tracking. To reiterate, Siamese trackers are mostly based on similarity matching with only weak temporal correspondence introduced through localizing the search area in any candidate frame around the target location in the previous frame and penalizing the changes in translation and scale between two consecutive frames. With RE-SiamNets, we explore the applicability in improving the temporal consistency through imposing restrictions on the rotational motion. This is achieved during the selection of $\theta_{\text{opt}} \in \Theta$ among the Λ orientations. Let $\theta_{t,\text{opt}} = \theta_{t,i}$, where $\theta_{t,i}$ refers to the i^{th} orientation in frame t. For the next frame, rather than selecting $\theta_{t+1,\text{opt}}$ from the full set Θ, a constraint can be imposed such that $\theta_{t+1,\text{opt}} \in \{\theta_i\}$. Index i here is constrained to the set $\{i_{t,\text{opt}} - \gamma, \ldots, i_{t,\text{opt}} - 1, i_{t,\text{opt}}, i_{t,\text{opt}} + 1, \ldots, i_{t,\text{opt}} + \gamma\}$ such that γ is the maximum change in number of orientations allowed in either directions between two consecutive frames. This constraint ensures that the orientation does not change by more than γ groups between two successive frames.

6. Rotating Objects Benchmark (ROB)

State-of-the-art benchmarks mostly do not contain rotation annotations. To evaluate RE-SiamNets as well as to enable future benchmarking on rotation sensitive tracking, unsupervised rotation estimation and rotation stabilization. We present Rotating Objects Benchmark (ROB) consisting of real world video sequences with large-scale variations in in-plane rotation of target objects.

ROB dataset is a collection of short video clips comprising multiple objects in diverse scenarios, where the target object undergoes rotation due to a rotating camera or/and an in-plane rotation of the object itself. In each video, the camera moves around the objects, capturing its different angles of rotation. The dataset consists of 35 video sequences with over 10,000 annotated frames and 15 object categories, ranging from a wide range of real-world scenarios such as livestock monitoring, cycling and aeroplanes.

Sequences from ROB dataset are densely annotated in a semi-automated manner, with each frame providing objects location using bounding box coordinates, as well as information about its orientation with respect to the frame. To annotate orientation change, a one-head arrow is drawn along one of the axes of target in the first frame, and consistently followed in rest of of the frames. This allows to compute the orientation change between the appearances of the target in any two frames of the sequences.

7. Experiments

We validate rotation equivariant Siamese trackers in tracking and estimation of relative 2D orientation changes. We first compare with the non-rotation equivariant version of the trackers, including SiamFC and SiamFCv2 [1] and SiamRPN++ [18]. The proposed design philosophy, however, is general and any Siamese tracker can benefit. Moreover, we compare with DiMP [2] that attains SOTA results on standard tracking benchmarks.

Training. All rotation equivariant variants of SiamFC are trained on the GOT-10k [14] training set. To train SiamRPN++, we trained a rotation equivariant version of ResNet50 architecture on ImageNet. The SiamRPN++ model was then trained using this backbone on sets of COCO [20], ImageNet DET [24], ImageNet VID and YouTube-BoundingBoxes Dataset [23]

Evaluation. To evaluate how well the proposed RE-SiamNets perform in presence of frequent in-plane rotations, we test them on ROB, Rot-OTB100 and Rot-MNIST datasets. Rot-OTB100 dataset is built by rotating each frame of OTB100 videos by 0.5 degree with respect to its previous frame. Rot-MNIST involves superposition of 3-5 MNIST digits on GOT-10k image backgrounds, and the digits translate and rotate randomly but in a smooth manner. Details related to the generation of these two datasets, as well as results on ROT-MNIST are provided in the supplementary section of this paper. To demonstrate that adding RE-SiamNets do not degrade the performance of trackers with respect to other challenges, we test them on tracking benchmarks that include OTB100 [37] and GOT-10k [14].

Implementation Details To design RE-SiamNets, we adapt the existing models by replacing the regular CNN layers with rotation equivariant layers and using a group-pooling layer to output features at single orientation for every input. These rotation equivariant modules are added using the `e2cnn` pytorch library [34]. For base Siamese trackers, we use SiamFC [1], its variant SiamFCv2, and SiamRPN++ [18]. Here and henceforth, we use the prefix 'RE-' to refer to the rotation equivariant version of a tracker.

For most experiments in this paper, we use RE-SiamFC. The base tracker SiamFCv2 differs from SiamFC in terms of the filter sizes and the number of convolutional layers. The former comprises only 4 convolutional layers with filter sizes of 9, 7, 7 and 6. The reason behind choosing this variant is to experiment with models involving largers filters, since these are known to work well for rotation equivariant CNNs [34]. Full details on SiamFC and SiamFCv2 are provided in the supplementary section. We further point out we will occasionally refer SiamFC and SiamFCv2 under the same name of SiamFC. We experiment with rotation groups of $\Lambda = 4, 8, 16$ for SiamFC and $\Lambda = 4$ for SiamRPN++.

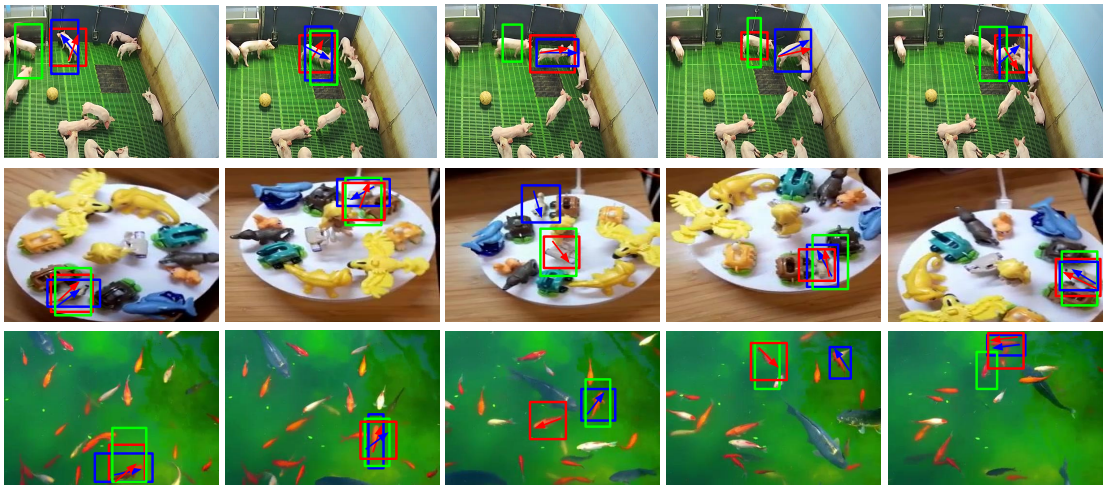

Figure 3: Example frames obtained with different template images from 3 sequences of ROB dataset. Images show the ground truth bounding box (blue), and predictions obtained using SiamFC[1](green) and RE-SiamFC using 8 rotation groups (red). Further, blue and red arrows show the ground truth pose estimate and the prediction obtained using RE-SiamFC.

Model	Type	Rot-OTB100 Success	Rot-OTB100 Precision	OTB100 Success	OTB100 Precision
SiamFC	-	0.315	0.523	0.578	0.765
	R4	0.360	0.629	0.567	0.745
	R8	0.423	0.676	0.566	0.749
SiamFCv2	-	0.288	0.473	0.540	0.724
	R4	0.348	0.622	0.526	0.710
	R8	0.425	0.678	0.532	0.717
	R16	0.423	0.688	0.514	0.705
SiamFCv2	aug	0.317	0.541	0.533	0.718
SiamRPN++	-	0.461	0.634	0.696	0.914
SiamRPN++	R4	0.485	0.679	0.691	0.903
DiMP18	-	0.429	0.643	0.660	-
DiMP50	-	0.447	0.668	0.684	-
DiMP50	R4	0.490	0.701	0.673	0.908

Table 1: Performance scores (success rate 'Succ' and precision 'Pr' of OPE) for object tracking using different Siamese trackers with regular CNNs as well as RE-SiamNets on Rot-OTB100 and OTB datasets. Further, 'aug' refers to inclusion of rotation augmentation during training.

All RE-SiamNet implementations described in this paper are trained using stochastic gradient descent method. The methods follow the same training configurations as those of their base trackers. Exceptions include training of RE-SiamFC with R16 for 150 epochs using batch size of 16. Further, the rotation equivariant ResNet50 backbone was trained on ImageNet for only 50 epochs due to limited computational time. All models were trained on machines equipped with either 1 or 4 GPU Titan X GPUs. Details on optimization can be found in the supplementary material.

7.1. Rotation Equivariance in Tracking

Rot-OTB100. Table 1 shows that adding rotations in the tracked sequences makes tracking considerably harder. Thus, compared to the performance obtained on standard OTB100, the precision and success scores for SiamFC drop by 24.2% and 26.3%, respectively. Further, for SiamRPN++, these scores drop by 23.5% and 28.0%, respectively. Even with just 4 rotational groups RE-SiamNet outperforms both variants of SiamFC comfortably. Importantly, rotation equivariant Siamese trackers are notably better than standard trackers trained on data with additional rotation augmentations. With 16 rotation groups, there does not seem to be any improvement in performance. The reason is that for the same number of parameters, 16 quantizations permit relatively very few channels per layer. When doubling the number of channels per layer in SiamFCv2-16, success and precision scores increase to 0.437 and 0.698, respectively. Going beyond 16 quantizations requires many more parameters and is susceptible to overfitting for SiamFC and slow to train with SiamRPN++. We also note that with 16 bins, we have fine-grained angle resolution (22.5^o, *i.e.*, ± 11.25 around the heading, similar to pose estimation [28]). Interestingly, adding rotation equivariance brings improvements even to deep siamese trackers such as SiamRPN++ [19] and DiMP [2]. See supplementary section for plots on AUC for precision and success scores.

ROB. We benchmark rotation equivariance also on natural in-plane rotations on the ROB dataset, see Figure 4. It shows the performance plots obtained on ROB dataset using SiamFCv2, SiamRPN++ and their RE-SiamNet equivalents. We make similar observations as in Rot-OTB100. Adding rotation equivariance makes both SiamFC and SiamRPN++ more capable to handle natural rotations and overall, the precision and success rates improve. We provide qualitative examples in Figure 3, showcasing the ben-

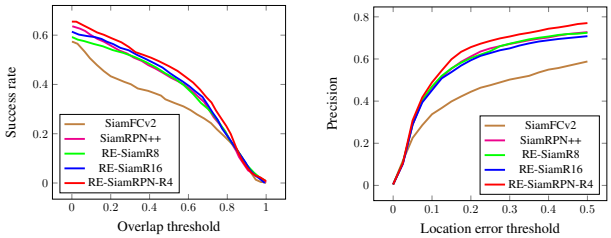

Figure 4: Performance curves for ROB dataset obtained using SiamFCv2 and RE-SiamNet with different choices of equivariant rotation groups.

Type	Range	ROB $SR_{0.5}$	ROB $SR_{0.7}$	Rot-OTB100 $SR_{0.5}$	Rot-OTB100 $SR_{0.7}$
Baselines	$\pm\frac{\pi}{4}$	0.25	0.25	0.25	0.25
	$\pm\frac{\pi}{8}$	0.125	0.125	0.125	0.125
	$\pm\frac{\pi}{16}$	0.062	0.062	0.0625	0.062
R4	$\pm\frac{\pi}{4}$	0.57	0.66	0.61	0.73
R8	$\pm\frac{\pi}{8}$	0.55	0.64	0.60	0.73
	$\pm\frac{\pi}{4}$	0.71	0.82	0.79	0.87
R16	$\pm\frac{\pi}{16}$	0.10	0.14	0.16	0.32
	$\pm\frac{\pi}{8}$	0.15	0.21	0.22	0.38
	$\pm\frac{\pi}{4}$	0.31	0.46	0.38	0.51

Table 2: Performance values for RE-SiamFC with R8 on the task of 2D relative pose estimation for ROT-OTB100 and ROB datasets. Scores reported are in terms of success rate (SR) at IoU thresholds of 0.5 and 0.7. Reported baselines are computed assuming equal probability for each orientation in the dataset.

efits of inducing rotation equivariance in Siamese trackers.

OTB100 and GOT-10k. To further analyze if the rotation equivariant formulation can have adverse effects on other tracking challenges, we compared the results of RE-SiamFC with 4 rotation groups to that of the base Siamese model on OTB100 and GOT-10k. For both the cases, drops in performance scores were within 2% of the original values. Such minor drop is expected given that the rotation equivariant trackers use lesser number of channels for the same number of parameters, thereby exhibiting slightly lower discriminative power in general.

7.2. Unsupervised Pose Estimation

We experimentally demonstrate that RE-SiamNets can extract the relative 2D pose of the target over time, using the first frame as a reference. We provide results in Table 2 on the Rot-OTB100 and ROB datasets. In this experiment, we measure the success rate SR_α as the fraction of frames for which the actual and predicted orientations are within the specified range at an IoU threshold of α.

We observe that rotation equivariant trackers recover the relative orientation change with average accuracy above 60%, well beyond the random baseline. With 8 rotational groups, RE-SiamNets can even predict angles within a confidence of $\pm\frac{\pi}{8}$ at a similar accuracy. For finer rotations within $\pm\frac{\pi}{16}$ there is a significant drop, with accuracies ranging between 0.1 and 0.3. The problem is that by increasing the rotation groups, we trade the parameters required for better tracking with parameters that are required for finer rotational bases, thus reducing the final discriminative capacity of our trackers. We include some qualitative examples in Figure 3 to show the orientations predicted by our rotation equivariant tracker.

Type	Range	Orientation Estimation $SR_{0.3}$	Orientation Estimation $SR_{0.5}$	Orientation Estimation $SR_{0.7}$	Tracking Pr	Tracking Succ
R8	$\pm\frac{\pi}{4}$	0.72	0.79	0.87	0.42	0.68
c-R8		0.75	0.80	0.88	0.43	0.69
R16	$\pm\frac{\pi}{4}$	0.34	0.38	0.51	0.42	0.69
c-R16		0.36	0.42	0.54	0.43	0.69

Table 3: Accuracy of orientation estimation and performance scores for object tracking on Rot-OTB100 dataset obtained for RE-SiamFC with (denoted with prefix 'c-') and without imposing constraint on rotational motion. Here, 'Range' refers to permissible change in orientation between two consecutive frames of any video, 'SR_X' refers to success rate at an IoU threshold of X.

7.3. Rotational-based Motion Constraints

Last, we explore briefly whether the predictions of orientation estimates can be used to improve tracking by an additional constraint to encourage smooth orientation changes over time. We present results in Table 3. Adding the rotation constraint on rotational motion has a modest yet positive influence on tracking performance, while the benefits regarding robustness are higher (data not shown). We conjecture that introducing other types of equivariance to place more constraints on the attainable types of motion in videos would yield even more robust trackers.

8. Conclusions

This paper addresses the challenge of in-plane rotations of the target in visual object tracking. We demonstrated that frequent in-plane rotations can have an adverse effect on conventional trackers, for which data augmentations do not suffice. To address this, we introduce rotation equivariant Siamese trackers, specifically for SiamFC and SiamRPN++, that can adapt to rotation changes at no extra parameter cost due to shared weights. Results show that rotation equivariant Siamese trackers can track accurately under the presense of artificial and natural rotations, they can accurately recognize the relative orientation changes of the target with respect to the first reference frame, and they can even be made more robust by placing additional rotational motion constraints.

References

[1] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *Computer Vision – ECCV 2016 Workshops*, pages 850–865, 2016. 1, 2, 4, 5, 6, 7

[2] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6182–6191, 2019. 3, 6, 7

[3] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999, 2016. 2, 3

[4] Taco S Cohen and Max Welling. Steerable cnns. *International Conference on Learning Representations (ICLR)*, 2017. 2, 3

[5] Sander Dieleman, Jeffrey De Fauw, and Koray Kavukcuoglu. Exploiting cyclic symmetry in convolutional neural networks. *International Conference on Machine Learning (ICML)*, 2016. 3

[6] Xingping Dong, Jianbing Shen, Dongming Wu, Kan Guo, Xiaogang Jin, and Fatih Porikli. Quadruplet network with one-shot learning for fast visual object tracking. *IEEE Transactions on Image Processing*, 28(7):3516–3527, 2019. 2

[7] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5374–5383, 2019. 3

[8] Heng Fan and Haibin Ling. Siamese cascaded region proposal networks for real-time visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7952–7961, 2019. 2

[9] William T Freeman, Edward H Adelson, et al. The design and use of steerable filters. *IEEE Transactions on Pattern analysis and machine intelligence*, 13(9):891–906, 1991. 4

[10] 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. 2

[11] Deepak K Gupta, Efstratios Gavves, and Arnold WM Smeulders. Tackling occlusion in siamese tracking with structured dropouts. In *Proceedings of the IEEE conference on pattern recognition*, 2020. 1, 3

[12] Anfeng He, Chong Luo, Xinmei Tian, and Wenjun Zeng. A twofold siamese network for real-time object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4834–4843, 2018. 2

[13] David Held, Sebastian Thrun, and Silvio Savarese. Learning to track at 100 fps with deep regression networks. In *European Conference on Computer Vision*, pages 749–765. Springer, 2016. 2

[14] 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*, 2019. 6

[15] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Luka Cehovin, Gustavo Fernandez, Tomas Vojir, Gustav Hager, Georg Nebehay, and Roman Pflugfelder. The visual object tracking vot2015 challenge results. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 1–23, 2015. 2

[16] Thijs P Kuipers, Devanshu Arya, and Deepak K Gupta. Hard occlusions in visual object tracking. In *Computer Vision – ECCV 2020 Workshops*, 2020. 1, 3

[17] 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. 2, 3

[18] 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 Conference on Computer Vision and Pattern Recognition*, pages 4282–4291, 2019. 1, 3, 5, 6

[19] 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. 1, 2, 5, 7

[20] T. Y. Lin, M. Maire, S. Belongie, J. Heys, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pages 740–755, 2014. 6

[21] Diego Marcos, Michele Volpi, and Devis Tuia. Learning rotation invariant convolutional filters for texture classification. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2012–2017. IEEE, 2016. 3

[22] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 300–317, 2018. 3

[23] E. Real, J. Shlens, S. Mazzocchi, X. Pan, and V. Vanhoucke. Youtube-BoundingBoxes: A large high-precision human-annotated data set for object detection in video. In *CVPR*, pages 7464–7473, 2017. 6

[24]). Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. volume 15, pages 211–252, 2015. 6

[25] Jianbing Shen, Xin Tang, Xingping Dong, and Ling Shao. Visual object tracking by hierarchical attention siamese network. *IEEE transactions on cybernetics*, 2019. 2

[26] Arnold WM Smeulders, Dung M Chu, Rita Cucchiara, Simone Calderara, Afshin Dehghan, and Mubarak Shah. Visual tracking: An experimental survey. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1442–1468, 2013. 3

[27] Ivan Sosnovik, Artem Moskalev, and Arnold Smeulders. Scale equivariance improves siamese tracking. *arXiv preprint arXiv:2007.09115*, 2020. 1, 3

[28] H. Su *et al.* Render for CNN: Viewpoint estimation in images using CNNs trained with rendered 3d model views. In *ICCV*. 7

[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. 1, 2, 5

[30] Jack Valmadre, Luca Bertinetto, Joao Henriques, Andrea Vedaldi, and Philip HS Torr. End-to-end representation learning for correlation filter based tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2805–2813, 2017. 2

[31] Guangting Wang, Chong Luo, Zhiwei Xiong, and Wenjun Zeng. Spm-tracker: Series-parallel matching for real-time visual object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3643–3652, 2019. 2

[32] Qiang Wang, Zhu Teng, Junliang Xing, Jin Gao, Weiming Hu, and Stephen Maybank. Learning attentions: residual attentional siamese network for high performance online visual tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4854–4863, 2018. 2

[33] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1328–1338, 2019. 2

[34] Maurice Weiler and Gabriele Cesa. General e (2)-equivariant steerable cnns. In *Advances in Neural Information Processing Systems*, pages 14334–14345, 2019. 2, 3, 5, 6

[35] 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. 2, 3, 4

[36] 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. 2, 3

[37] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2411–2418, 2013. 6

[38] 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. 2

[39] 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. 3

[40] Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, and Weiming Hu. Distractor-aware siamese networks for visual object tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 101–117, 2018. 2