

UNIVERSITY OF ŽILINA IN ŽILINA
FACULTY OF MANAGEMENT SCIENCE AND INFORMATICS



Visual Object Tracking

DISSERTATION THESIS

2022

Ing. Milan Ondrašovič

Declaration

I sincerely declare that the thesis *Visual Object Tracking* has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except for the cases where stated otherwise, either by reference or an acknowledgment, the work presented is entirely my own created under the guidance of the aforementioned supervisors.

Author signature:

Acknowledgements

My deepest gratitude belongs to both my supervisors and closest family for showing understanding in situations where the health-related problems interfered with my life, and last but not least, with my research. The guidance and scrupulous cooperation I received made coping with the work easier.

Abstract

Add abstract in Slovak language.

Kľúčové slová: vizuálne trasovanie objektov, hlboké strojové učenie, Siamské neurónové siete, latentné priestory, mechanizmus pozornosti, analýza dopravy

Abstract

Visual object tracking is a challenging problem of current research and a central topic of this work. In this thesis, we disseminate our results concerning research in this area using deep machine learning tools. However, taking advantage of deep machine learning requires a large amount of data for model training and testing. With this in mind, we provide a short discussion of available and used datasets. More importantly, we provide a review of various approaches to tracking. As a result, object occlusion, one of the primary difficulties in object tracking, was selected as the problem to focus on. Besides tracking, this treatise includes similarity learning, a relevant and related area that is commonly approached using so-called Siamese neural networks. Consequently, the application of methods based on the aforementioned domain is not only discussed but also experimented with. Re-identification was explored to tackle object re-appearance after full occlusion. At the same time, to address the ramifications of the partial occlusion, we developed an attention-based approach. Latent spaces and embedding vectors were exploited to enhance the object ID propagation between frames. Moreover, one of the chapters is dedicated to our attempt to aid the tracking process using homography transformations. Throughout this dissertation, traffic analysis is frequently mentioned as the target real-world use case. Despite this specificity, the developed solutions and conclusions drawn from them apply to general object tracking with no specific constraints, although there is potential for additional improvement under concrete circumstances. Overall, few of the documented solutions did not yield the expected outcome. But, there are two contributions presented as two journal publications, namely an up-to-date comprehensive survey of Siamese-based tracking and the development of a new homography ranking method. The already promising results of the attention-based solution are still under proper validation and further evaluation.

Keywords: visual object tracking, deep machine learning, Siamese neural networks, latent spaces, attention mechanism, traffic analysis

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Visual Object Tracking (VOT) is one of the principal challenges in the field of computer vision is to locate a certain object in all frames of a video, given only its position in the first frame. An object is firstly detected in the image (frame) and a unique identifier is assigned to it. Subsequently, the same identifier has to be correctly assigned if the object is present in future images.

Object tracking is the task of following one or more objects in a scene, from their first appearance to their exit [?]. In general, this problem is still wide-open, with state-of-the-art (SOTA) performances lagging far behind human levels. However, there are successful real-world applications, particularly when a certain amount of control over the environment is possible, *e.g.* in industrial settings. Major difficulties stem from a change in object illumination, position, and orientation due to movement, object and camera viewpoint variations, and partial or full occlusion [?].

The goal of an object tracker is to produce a trajectory of a given object with respect to time using its position in every video frame. A practically unattainable (ideal) tracking algorithm should have the properties below [?]:

- properly detect all the objects that enter and exit the scene,
- differentiate between instances of multiple objects,
- consistently maintain the uniquely assigned identifier to each object,
- motion of the object or lack thereof should not influence the object tracking,
- partial or full object occlusion, even a long-term one, should be resolved.

From a methodological perspective, VOT may be tackled in two fundamentally distinct ways, namely *top-down* and *bottom-up* [?]. Using the top-down methodology, an attempt is made at every frame to locate the object given some initial hypothesis [?]. On the other hand, the bottom-up approach utilizes the object detection first and then the adequate correspondence is established with previously detected objects [?].

Recently, the most influential approaches to VOT were the ones involving the modern tools of machine learning, especially deep machine learning. Deep learning can perform global generalization, making it a valuable tool as a function approximator. This revolutionary idea which has in the past decade reaped an upsurge in its utility will be the key element of this thesis. Krizhevsky *et al.* [?] showed that an outstanding tool, when it comes to the application of deep learning in computer vision, are Convolutional Neural Networks (CNNs) (Section ??). It is a predominant approach to extract valuable visual features from the pixel space of images as well as a fundamental building block of our work.

Visual tracking of single or multiple objects is often just an intermediate step for various ends. In this work, the traffic analysis, specifically, tracking vehicles, is considered to be the primary target of our applied research, even though, our developed methods are generally applicable regardless of this specific use case. Besides traffic analysis [?], whether from a static camera or as part of self-driving cars, there are pedestrian detection and tracking [?], activity understanding [?], and imitation based on a video [?]. While this is an important problem, the current SOTA solutions still lack high accuracy in unconstrained scenarios with potential object occlusion [?]. Understandably, the object might re-emerge after the occlusion in a significantly different form, thus it might be mistaken for a new object. Occlusion comes in three separate types [?]:

- *self-occlusion*, where the object occludes itself (a person holding a phone),
- *intra-object occlusion*, in which multiple different objects occlude each other (a small vehicle passing behind a truck),
- or *object-background occlusion*, when the occlusion is caused by a static object in the background (a tree occluding a cat).

A key element for a tracking algorithm to hold onto when dealing with occlusion is to discern between new and previously seen objects. For this purpose, a repeated identifi-

cation of some object, or Re-identification (ReID), is indispensable. Various advances in the creation of latent spaces and *embeddings* using deep learning (Section ??) have shown promising results [?, ?]. One use case of embeddings is to create a metric space into which the tracked visual objects are encoded as vectors. It is a form of dimensionality reduction, where similar objects are mapped onto a manifold closer together while distinct objects are mapped further away from each other [?].

A broad range of real-life applications requires tracking multiple objects, which only adds complexity to an already tough problem itself. But [?] shows that approaching the problem of vehicle ReID using embeddings (Section ??) trained with contrastive or triplet loss (Section ??) brings substantial improvement. We plan to explore this idea and utilize it as a basis for VOT when it comes to repeatedly re-identify occluded objects.

The primary objective of this thesis is to explore, implement, and experiment with methods for VOT by use of deep learning. Considering the performance of SOTA approaches as well as the practical demand for an accurate tracking outcome, be it traffic or other scenarios [?], we think that an emphasis should be put on occlusion handling, whether partial or full, as it causes major difficulties for existing methods [?].

The secondary objective is to extend the current knowledge in the field of computer vision and deep learning regarding dynamic scenes involving VOT. Nowadays, at the time of writing this document, there is still a lack of freely available implementations. Widely used open-source libraries such as OpenCV [?] provide only tracking algorithms for single objects, as opposed to Multi-Object Tracking (MOT), which is demanded in practice, yet concrete solutions exist but are not ubiquitous and easy to use.

The rest of the document is organized as follows. The main goals of this dissertation thesis are described in Chapter ??, discussed next. To equip the reader with the necessary foundational knowledge we provide Chapter ?. Right after this chapter, we composed a short treatise on available and used datasets in Chapter ?. At this point, we start with our first relevant experiment related to homography transformations, to which an entire Chapter ?? is dedicated. This chapter is a one-to-one re-write and expansion of our published paper, so this work can be treated as finished and successfully published. Then, we have Chapter ??, focused on our developed approaches to VOT. It is by far the most important chapter and presents the greatest part of our work. Our achieved results are discussed subsequently in Chapter ??, which closes the entire document with

a conclusion.

Chapter 2

Dissertation Thesis Goals

VOT has been increasingly studied in recent years due to its real-world applications. The research area is vast, very active, and encompasses a broad range of approaches. Deep learning has occupied a great deal of the time that we spent studying and researching. The intention to build a working solution and incrementally advance the domain of object tracking has been the main incentive behind the choice for the topic of this dissertation thesis. With this in mind, **the general goal of this dissertation thesis is to improve the accuracy of visual object tracking using deep machine learning tools.**

Tracking of objects using visual features may produce the output in many forms, the usual axis-aligned bounding boxes (BBOXes), rotated BBOXes [?], segmentation masks [?], or even object contours [?]. We plan to primarily focus on solutions producing axis-aligned BBOXes since they are ubiquitous. We have mentioned our intention to deal with traffic-related scenarios. So far, there have been no restrictions as to why our tracker would have to run at real-time speed. Many of our real-world applications have involved camera-recorded videos that were processed offline. The decision to deal with traffic-related scenarios is also driven by projects supported by the University of Žilina. Companies of the private or public sector are interested in the automated analysis of traffic. Currently, a great deal of work such as vehicle counting is performed with human intervention. The author of this thesis as well as one of the supervisors were actively involved in solving problems related to tracking and traffic analysis using recorded videos during the research performed as part of this dissertation.

There are diverse factors that contribute to the overall performance of a tracker. Our analysis has led us to narrow our focus to occlusion handling. Occlusion can impose a huge

precision penalty since the tracked object may be lost, or worse, the tracker’s attention may be dragged away to a different object (a so-called drift problem). Many of the SOTA solutions lack explicit occlusion handling [?, ?, ?] and we have identified this to be one of the leading causes of failure. Thus, **the specific goal is to propose a solution to problems regarding visual tracking that stem from the presence of occlusion, the transformation of the tracked objects as well as varying lighting conditions in the scene.**

Given what has been presented so far, we have chosen the path of similarity learning and a closely related ReID (Section ??). Despite the existence of different techniques, we have been convinced during this initial research phase about the great potential of metric spaces and their properties that seem to fit our needs [?]. Some works argue that a well-built similarity function based upon metric learning in combination with a simple matching algorithm on the level of BBOXes can produce a reasonable performance [?]. We think that even humans would be capable of discerning between objects when shown their pictures from distinct times in a video even dozens of seconds apart. The inherent visual clue about the object that makes it stand out among the set of others should be present most of the time. However, it may not be always possible to unambiguously identify an object given its appearance, especially when it comes to vehicles. Special marks such as customized paintings, decorations, or even scratches become relevant [?]. Consequently, the use of attention-based [?] approaches has yielded promising results. In addition, attention may serve the purpose of enhancing the tracker’s discriminating power when it comes to detecting objects undergoing a partial occlusion. Thus, expanding upon the foundation of the specific goal, we will strive to propose and implement a solution that can improve the performance of a tracker by considering the obstacles mentioned above. Therefore, **the goal from a methodological perspective is the application of approaches based on attention or similarity learning to handle object occlusion of varying intensity, change in the position and viewpoint of the tracked object, and fluctuations in the scene illumination.**

Chapter 3

Theoretical Foundations

3.1 Neural Networks

3.1.1 Artificial Neural Networks

Artificial neural networks (or just neural networks for short) are computing systems that are inspired by, but not identical to, biological neural networks that constitute animal brains. Such systems essentially *learn* to perform tasks by considering multiple samples, generally without being programmed with task-specific rules. They form a basis for deep machine learning, and are also known in the literature as the multilayer perceptron [?], feedforward neural networks, or deep feedforward neural networks [?].

In mathematical terms, the goal of a neural network is to approximate some unknown function f . For instance, when considering a classifier, the transformation $y = f(\mathbf{x})$ maps the given input \mathbf{x} to a category y . Such a network, therefore, defines a mapping and learns the value of the parameters that result in the best function approximation [?].

The name *feedforward* is derived from the fact that the information flows through the function being evaluated from \mathbf{x} , through the intermediate computations, and finally to the output \mathbf{y} . Such a model can be described with a directed acyclic graph denoting the sequential composition of several functions. More concretely, we might have three functions $f^{(1)}$, $f^{(2)}$ and $f^{(3)}$, forming a chain, $f(\mathbf{x}) = f^{(3)}(f^{(2)}(f^{(1)}(\mathbf{x})))$. These chain structures are the most commonly used structures in neural networks [?]. Deep machine learning consists of multiple such layers of neurons stacked onto each other that are trained using the *backpropagation* algorithm [?].

3.1.2 Convolutional Neural Networks

This type of neural network has gained popularity in the computer vision community thanks to a never-seen-before performance on image classification task [?]. This approach processes images or other high dimensional, grid-like input and then learns the importance (weights and biases) of various aspects of the input data.

The successful ability of these networks to capture spatial properties via learned convolutional filters is the fundamental principle. Let f and g be functions. Then, the operation of convolution denoted by \star produces a third function, as a result of the following computation (demonstrating the commutativity property, too) [?]:

$$(f \star g)(t) = \int_{-\infty}^{\infty} f(\tau) g(t - \tau) d\tau = \int_{-\infty}^{\infty} f(t - \tau) g(\tau) d\tau. \quad (3.1)$$

In this setting, f is the input, and g is the kernel. The output of this operation is a feature map. During the training, the aim is to learn the weights of the kernel matrix that produces a feature map based on which the model can solve the given task according to some objective function.

Let I be a two-dimensional input image and K be a two-dimensional kernel. Then, for a given position (i, j) in the input image I , the discrete convolution can be written as

$$(f \star g)(i, j) = \sum_m \sum_n I(m, n) K(i - m, j - n). \quad (3.2)$$

Many machine learning libraries implement the cross-correlation operation, not the convolution operation by its strict definition. This operation is the same except for the fact that the kernel is not flipped. Thus, the result of the cross-correlation is given by [?]

$$(f \star g)(i, j) = \sum_m \sum_n I(i + m, j + n) K(m, n). \quad (3.3)$$

An indispensable outcome of CNNs is the ability to capture hierarchical relations. Layers placed near the input of the model capture low-level features such as edges, colors, gradient orientations, and so on. On the other hand, layers placed further, deeper in the model, highlight semantic, abstract features that are specific to the task at hand. We will refer to this fact many times in our work since visual object trackers need to accurately discriminate between the foreground and background. The exploitation of the mentioned

qualities of convolutional layers will be demonstrated in Section ??.

Another prominent use case of CNNs is *transfer learning*, where a pre-trained model is adopted for a new task, utilizing the already learned features. These pre-trained models may come in various flavors, but typical ones are pre-trained for an image classification task using the ImageNet dataset [?]. The reasoning is that visual features such as edges and contours are vital to general object recognition tasks, hence it is not needed to learn coarse, rudimentary, low-level features from scratch all the time. Our work will certainly require the *transfer learning*, and additional *fine-tuning* of the pre-trained model.

3.2 Object Detection

Since the topic of this thesis concerns the task of *visual* object tracking, an indispensable step in the processing pipeline will unquestionably be object detection. Holistically, there are *fast* object detectors, such as the most prominent one, You Look Only Once (YOLO) [?] (Section ??), or Single Shot Detector (SSD) [?]. Here, we use the term *fast* to denote that the detector can operate on high Frames per Second (FPS). This comes at a cost of relatively lower accuracy, though, as compared to *slower* but more accurate approaches based on Region Proposal Networks (RPNs), such as various versions of Faster Region-based Convolutional Neural Network (Faster R-CNN) (Section ??).

Object detection is, by its very nature, a difficult task, because the number of objects is unknown in advance, which means, that the number of outputs of the model is variable. A plethora of attempts have been proposed to evade this inherent shortage of standard neural networks. An obvious solution is to only produce a constant number of BBOXes, as utilized by SSD and YOLO, for instance. But methods based on RPNs try to circumvent the obstacle of having to predict only a fixed set of BBOXes. Other differences between object detectors stem from the architecture itself, whether the training is an end-to-end pipeline or the model consists of various parts. Fully convolutional architectures are also becoming more prevalent [?].

3.2.1 Non-Maximum Suppression

Object detectors have hugely profited from the end-to-end learning paradigm in which features, object proposals, and the classifier become part of one model [?]. A proposal

Algorithm 1 Non-Maximum Suppression

```

1: function NMS( $\mathcal{B}$ ,  $\mathcal{S}$ ,  $\lambda$ )
2:    $\mathcal{B}_{nms} \leftarrow \emptyset$                                  $\triangleright$  initialize the output (filtered) set of region proposals
3:   while  $\mathcal{B} \neq \emptyset$  do                                 $\triangleright$  loop until all the proposals are processed
4:      $m \leftarrow \arg \max_{i \in \{1,2,\dots,|\mathcal{S}|\}} \mathcal{S}$        $\triangleright$  find an index of a proposal with the highest score
5:      $\mathcal{B} \leftarrow \mathcal{B} - \mathbf{b}_m$ ,  $\mathcal{S} \leftarrow \mathcal{S} - s_m$            $\triangleright$  remove the proposal
6:      $\mathcal{B}_{nms} \leftarrow \mathcal{B}_{nms} \cup \mathbf{b}_m$             $\triangleright$  save the proposal with the highest score
7:     for  $i \leftarrow 1$  to  $|\mathcal{B}|$  do                       $\triangleright$  iterate through remaining proposals
8:       if  $\text{IOU}(\mathbf{b}_m, \mathbf{b}_i) \geq \lambda$  then         $\triangleright$  IoU (Equation ??) exceeds the threshold
9:          $\mathcal{B} \leftarrow \mathcal{B} - \mathbf{b}_i$ ,  $\mathcal{S} \leftarrow \mathcal{S} - s_i$            $\triangleright$  remove the proposal
10:      end if
11:    end for
12:  end while
13:  return  $\mathcal{B}_{nms}$ 
14: end function

```

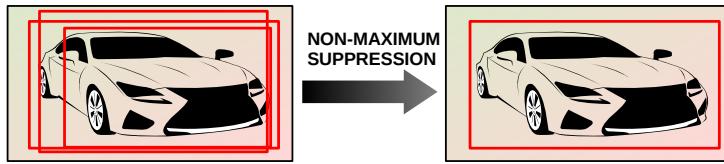


Fig. 3.1: An illustration of a potential effect of the NMS algorithm on BBOXes. Multiple proposals (left) are filtered so that only the ones with the highest detection score (right) remain while satisfying the condition that the overlap does not exceed a specific threshold.

is nothing but a region containing a potential object of interest. However, the number of proposals may grow considerably, outnumbering the real count of present objects. Moreover, these proposals may have a large overlapping region as measured by Intersection over Union (IoU) (Section ??), rendering most of them useless in terms of conveying new information. To filter such proposals, the Non-Maximum Suppression (NMS) algorithm introduced below is used (Fig. ??). The objective is to iteratively select only proposals the IoU of which is below a specific threshold.

Let $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n\}$ be a set of n region proposals described by n BBOXes. Scores for each detection are contained in a set $\mathcal{S} = \{s_1, s_2, \dots, s_n\}$, where s_i denotes a detection score for the i -th box, \mathbf{b}_i . Let λ , such that $0 \leq \lambda < 1$, denote the threshold for the maximum allowed portion of the overlap between regions. \mathcal{B}_{nms} is the set of filtered proposal instances from the set \mathcal{B} produced using NMS (Algorithm ??).

3.2.2 YOLO

YOLO is a very popular single-stage object detector thanks to its ability to run in real-time and yet be sufficiently accurate. Its speed is primarily a consequence that *it looks only once* at a given input image. Compared to RPN-based approaches, the authors of YOLO devised a CNN model capable of performing extraction of region proposals as well as classification in a single run [?]. Besides, the backbone CNN model that is responsible for handling the visual input processes an entire image during the training and test time, allowing the implicit inclusion of contextual information about classes together with their visual representation. During the testing phase, after the single forward pass through the model, the NMS algorithm is employed to filter predictions to make sure that each object instance is detected just once. Since the initial introduction of this approach [?], multiple updates have been brought forward, either from the main author himself [?, ?] or from other researchers [?, ?]. Despite its popularity, for our experiments, we employed the paragon of two-stage object detection, discussed next.

3.2.3 Faster R-CNN

The Faster R-CNN [?] is the most prominent *two-stage* object detector. The first stage consists of generating region proposals using the RPN [?]. These class-agnostic proposals, a set of rectangular BBOXes, are produced from an input of arbitrary size (allowed by the fact that this process is modeled by a fully convolutional network) (Fig. ??). As far as the second stage is concerned, the proposed regions (usually 300) serve as a basis for subsequent cropping of features from the same intermediate feature maps that are then fed to the remaining feature extractor to predict the class. Based on this class prediction, the proposed box is further refined.

As the endeavor is to diminish unnecessary computations, the proposals are not cropped explicitly from the input image, and then processed by the feature extractor. Nevertheless, there is a part of the computation that has to be executed once per each proposed region, so the performance depends on the number of regions generated by the RPN.

This architecture is especially important for this treatise as it formed a base object detector for the multi-object tracker with which we conducted the majority of our object tracking research.

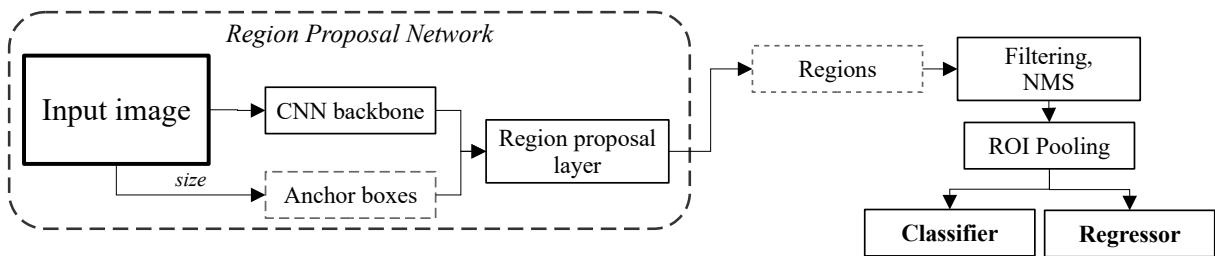


Fig. 3.2: A conceptual processing diagram of the Faster R-CNN model.

3.3 Latent Spaces and Embeddings

3.3.1 Learning Metric Embedding

As Hermans *et al.* [?] describe, the goal of learning metric embedding is to learn a function $f_\theta(x) : \mathbb{R}^F \rightarrow \mathbb{R}^D$, which maps semantically similar points from the data manifold in \mathbb{R}^F onto metrically close points in \mathbb{R}^D . Analogously, $f_\theta(\cdot)$ should map semantically different points in \mathbb{R}^F onto metrically distant points in \mathbb{R}^D .

Suppose the use of this transformation for vehicle ReID. The corresponding embedding vector would be produced by a learned function that would map the images of vehicles into a latent space, where images of the same vehicle would be mapped closer together. Moreover, such mapping should be invariant to variations in lighting conditions, vehicle rotations, and many others. Among other things, embedding trained this way can be used to produce a feature vector for classification, one-shot learning tasks [?], clustering [?], face recognition [?] and last, but not least, object ReID [?].

3.3.2 Embedding Vector Similarity

There are various ways to evaluate the degree of similarity between embedding vectors. For simplicity as well as our purposes, we will discuss the two most common approaches, *i.e.*, Euclidean distance and cosine similarity.

Let \mathbf{u} and \mathbf{v} be arbitrary D -dimensional vectors representing our embedding vectors. The Euclidean distance between the vectors \mathbf{u} and \mathbf{v} is defined as

$$\|\mathbf{u} - \mathbf{v}\|_2 = \sqrt{\sum_{i=0}^D (\mathbf{u}_i - \mathbf{v}_i)^2}, \quad (3.4)$$

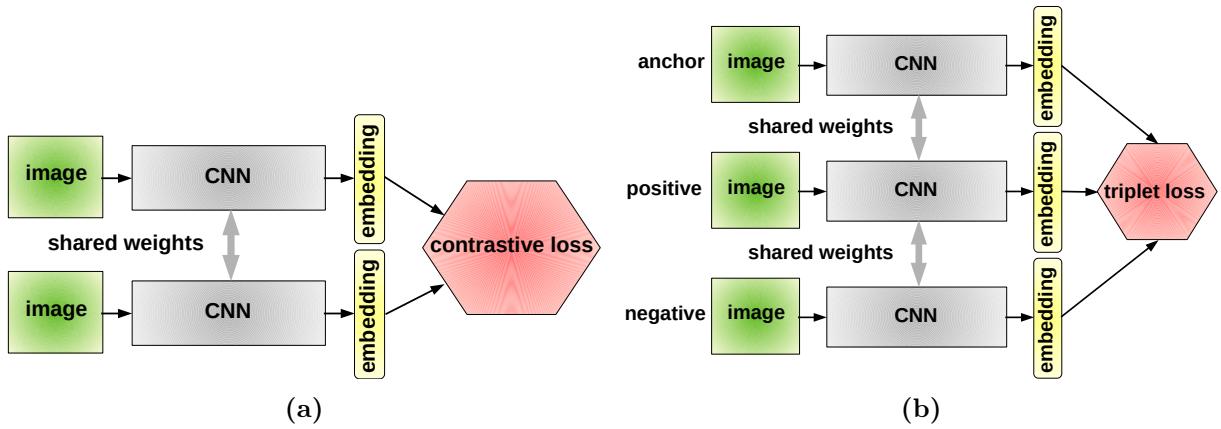


Fig. 3.3: Comparison of the Siamese (a) and the triplet (b) network architectures. The concept of weight sharing implies that only one set of model weights is trained.

and the cosine similarity is defined as

$$\cos \angle (\mathbf{u}, \mathbf{v}) = \cos (\theta) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|_2 \|\mathbf{v}\|_2}, \quad (3.5)$$

where θ is the angle between the vectors \mathbf{u} and \mathbf{v} .

3.3.3 Siamese and Triplet Networks

For the upcoming discussion, let $D(x, y) : \mathbb{R}^D \times \mathbb{R}^D \rightarrow \mathbb{R}$ be a metric function measuring distances in the embedding space. Without a loss of generality, we resort to use of the Euclidean distance (L_2 norm), so $D(x, y) = \|x - y\|_2$.

Contrastive Loss

Consider a sample (x_0, x_1, y) , where x_0 and x_1 represent the input, and the label $y = 1$ if x_0 and x_1 belong to the same category, otherwise $y = 0$. $D(\cdot)$ is the previously defined metric function, and α is the margin representing the minimum distance in the metric space to separate positive from negative samples. Then the contrastive function for any sample is defined as [?]

$$\mathcal{L}_{contr}(\theta) = \frac{1}{2}yD(f_\theta(x_0), f_\theta(x_1))^2 + \frac{1}{2}(1-y)([\alpha - D(f_\theta(x_0), f_\theta(x_1))]_+)^2. \quad (3.6)$$

The two inputs x_0 and x_1 are fed to the shared model at the same time. The output is then evaluated by the contrastive loss function (Fig. ?? (a)). Positive samples should have a small distance between each other as measured by the $D(\cdot)$ to decrease the loss

towards 0. Conversely, negative samples should have a distance beyond the threshold α .

Triplet Loss

Apart from the contrastive loss, this time three samples are required to compute the loss. The rationale is to supply additional context when forming the metric space. Siamese networks are usually implemented using shared model weights, but there are better approaches when the triplet loss is used. Conceptually speaking, the model could be implemented as shown in Fig. ?? (b). However, as we will discuss later, triplet mining strategies are required for the triplet loss to work properly. The model can be thus simplified from an architectural point of view at the cost of moving the complexity to the computation of the loss function.

Let N be the number of all possible valid triplets (x_a^i, x_p^i, x_n^i) for a given dataset. For any i -th triplet, let x_a^i be the *anchor* for a specific object (person, vehicle, etc.) with label $y(x_a^i)$, x_p^i be the positive sample of the same object with label $y(x_p^i)$, such that $x_a^i \neq x_p^i \wedge y(x_a^i) = y(x_p^i)$, and let x_n^i with label $y(x_n^i)$ be a sample of any other object, satisfying $y(x_a^i) \neq y(x_n^i)$, $\forall i = 1, \dots, N$. Let α be the margin value that is enforced between positive and negative pairs. Then, we want the relationship

$$D(f_\theta(x_a^i), f_\theta(x_p^i)) + \alpha < D(f_\theta(x_a^i), f_\theta(x_n^i)), \forall i = 1, \dots, N, \quad (3.7)$$

to hold true. The triplet loss function is therefore defined as

$$\mathcal{L}_{triplet}(\theta) = \sum_{i=1}^N [\alpha + D(f_\theta(x_a^i), f_\theta(x_p^i)) - D(f_\theta(x_a^i), f_\theta(x_n^i))]_+. \quad (3.8)$$

During the training using the triplet loss, the model should learn to push negative samples further away from the positive samples, ideally exceeding the margin α . Assuming a situation when the negative sample is mapped closer than a positive sample, the training should result in the desired situation of bringing the positive sample closer while pushing the negative one further (Fig. ??).

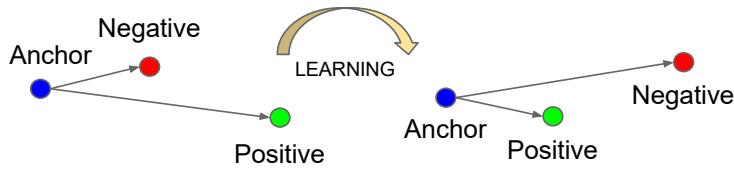


Fig. 3.4: The objective is to learn embeddings such that the anchor is closer to the positive example than it is to the negative example by some specified margin value. (source: [?])

triplet	constraint
<i>easy</i>	$D(f_\theta(x_a^i), f_\theta(x_p^i)) + \alpha < D(f_\theta(x_a^i), f_\theta(x_n^i))$
<i>semi-hard</i>	$D(f_\theta(x_a^i), f_\theta(x_p^i)) < D(f_\theta(x_a^i), f_\theta(x_n^i)) < D(f_\theta(x_a^i), f_\theta(x_p^i)) + \alpha$
<i>hard</i>	$D(f_\theta(x_a^i), f_\theta(x_n^i)) < D(f_\theta(x_a^i), f_\theta(x_p^i))$

Table 3.1: Definitions of various categories of triplets (regardless whether it is positive or negative) as imposed by their distance relationship.

3.3.4 Triplet Mining Strategies

Contrastive (Equation ??) and triplet (Equation ??) loss functions play an important role in training an embedding model. However, the way that pairs or triplets are selected is crucial, and as Hermans *et al.* [?] and Manmatha *et al.* [?] showed, it may significantly influence the training. Moreover, as the dataset gets larger, then the number of possible triplets grows cubically, rendering the use of all of them impractical. The majority of those triplets would be so-called *easy triplets*, thus hindering the learning process. To paraphrase the analogy from [?], showing the model that people with different clothes are not the same person after a certain point does not bring any new information. On the other hand, explicitly *mining* images of similar-looking yet different people with the same clothes (*hard negatives*) or of the same person with dramatically different poses (*hard positives*) vastly contributes to an understanding of the notion of the *same person*. As suggested, there are different kinds of triplets which are defined in Table ??, using a general (x_a^i, x_p^i, x_n^i) triplet for clarity. We encourage the reader to observe Fig. ??, too.

In order to attain an effective convergence during the training, it is necessary to select triplets that violate the triplet constraint in Equation ???. This means that given x_a^i , the goal is to select a *hard positive* x_p^i given by $\arg \max_{x_p^i} \{D(f_\theta(x_a^i), f_\theta(x_p^i))\}$ and a *hard negative* x_n^i as a result of $\arg \min_{x_n^i} \{D(f_\theta(x_a^i), f_\theta(x_n^i))\}$. Admittedly, it is often infeasible to compute the $\arg \min \{\cdot\}$ and $\arg \max \{\cdot\}$ over the entire training set. In this regard,

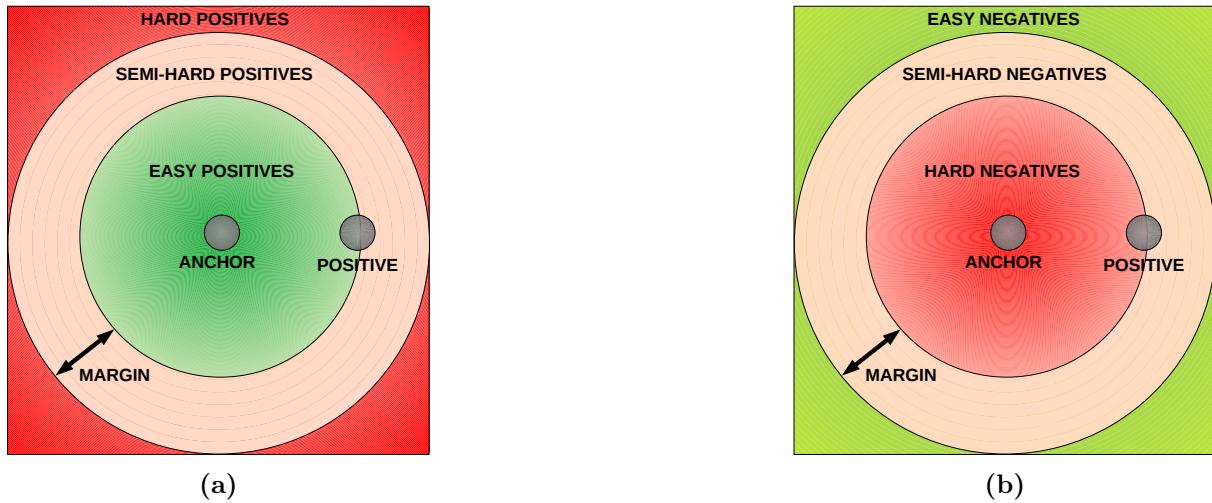


Fig. 3.5: Given a fixed anchor x_a^i and positive sample x_p^i as well as some positive margin value α , we discriminate between three different types of categories in terms of their level of *difficulty*. These categories vary in relation to positive **(a)** or negative **(b)** perspective.

there are two possible approaches to tackle this problem, either by selecting these hard triplets online or doing it offline [?].

Offline Triplet Mining

Given a training set, the task is to produce reasonable triplets off-line, for instance, at the epoch beginning. First, a list of N different valid triplets is randomly generated, then separated into $\lfloor N/B \rfloor$ batches of B triplets, followed by computation of $3N$ embeddings using the most recent model checkpoint. Then, hard or semi-hard triplets may be selected. Since this strategy has been shown on multiple occasions [?, ?, ?] as an inferior choice compared to the online triplet mining, we will not discuss this approach further.

Online Triplet Mining

Online mining is performed by selecting the hard positive/negative exemplars from within a mini-batch (Fig. ??). A condition that a minimum number of exemplars for any identity is present in each mini-batch has to be met. For example, Schroff *et al.* [?] used 40 different images of a single person (an identity) per mini-batch. Let P be the number of different objects/identities (*e.g.*, people, vehicles) and K be the number of different images for a concrete identity (*e.g.* different views of the same vehicle). There are two prominent approaches to online mining: *batch all* and *batch hard*.

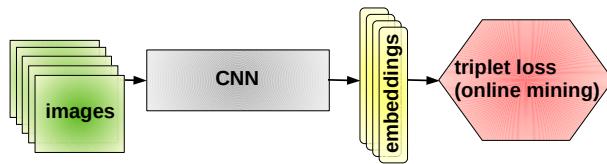


Fig. 3.6: The architecture of a triplet network where an online triplet loss function is used for training. In this architecture, no weight sharing is required as the triplet selection happens *online* solely in the loss function.

Online Triplet Mining: Batch All

This strategy aims for selecting all valid triplets and averaging the loss only on the hard and semi-hard triplets. Easy triplets, *i.e.*, those for which the loss function equals 0, are not taken into account. The reason is that averaging on them would result in a very small loss, since they would usually vastly outnumber the set of hard triplets [?]. This approach produces a total of $PK(K - 1)(PK - K)$ triplets (PK anchors, $K - 1$ positives per anchors, $PK - K$ negatives) incorporated in the loss function as

$$\mathcal{L}_{batchall}(\theta) = \sum_{i=1}^P \sum_{a=1}^K \sum_{\substack{p=1 \\ p \neq a}}^K \sum_{j=1}^P \sum_{\substack{n=1 \\ j \neq i}}^K \left[\alpha + D(f_\theta(x_a^i), f_\theta(x_p^i)) - D(f_\theta(x_a^i), f_\theta(x_n^j)) \right]_+ . \quad (3.9)$$

Online Triplet Mining: Batch Hard

In this strategy, the goal is to find the hardest positive and hardest negative for each anchor. The total number of triplets is PK . The selected triplets are the hardest among the given batch and can be considered moderate since they are the hardest within a small subset of the data. Therefore, the mining can be formulated as

$$\mathcal{L}_{batchhard}(\theta) = \sum_{i=1}^P \sum_{a=1}^K \left[\alpha + \max_{p=1, \dots, K} \{D(f_\theta(x_a^i), f_\theta(x_p^i))\} - \min_{\substack{j=1, \dots, P \\ n=1, \dots, K \\ j \neq i}} \{D(f_\theta(x_a^i), f_\theta(x_n^j))\} \right]_+ . \quad (3.10)$$

	Actual positive	Actual negative
Predicted positive	True Positive (TP)	False Positive (FP)
Predicted negative	False Negative (FN)	True Negative (TN)

Table 3.2: A confusion matrix in the standard form.

3.4 Evaluating Information Retrieval

In this section, we will use the standard terminology to denote whether the classification is correct (true) or not (false), for both positive and negative categories, as given by the confusion matrix (Table ??).

3.4.1 Evaluating Bounding Box Prediction

Intersection Over Union

Intersection over Union (IoU) measures the overlap between two BBOXes. Let $\mathbf{b}_1^T = [x_1, y_1, w_1, h_1]$ and $\mathbf{b}_2^T = [x_2, y_2, w_2, h_2]$ be two BBOXes described by vectors containing 4 elements. The respective elements are given by x, y coordinates of the top-left corner and the BBOX width and height. The intersection area between \mathbf{b}_1 and \mathbf{b}_2 is defined as

$$\begin{aligned} \mathbf{b}_1 \cap \mathbf{b}_2 = & \max \{0, \min \{x_1 + w_1, x_2 + w_2\} - \max \{x_1, x_2\} + 1\} \times \\ & \max \{0, \min \{y_1 + h_1, y_2 + h_2\} - \max \{y_1, y_2\} + 1\}, \end{aligned} \quad (3.11)$$

and the area of their union is given by

$$\mathbf{b}_1 \cup \mathbf{b}_2 = w_1 h_1 + w_2 h_2 - \mathbf{b}_1 \cap \mathbf{b}_2. \quad (3.12)$$

Then, the final IoU metric between \mathbf{b}_1 and \mathbf{b}_2 is computed as (Fig. ??)

$$\text{IoU}(\mathbf{b}_1, \mathbf{b}_2) = \frac{\mathbf{b}_1 \cap \mathbf{b}_2}{\mathbf{b}_1 \cup \mathbf{b}_2}, \quad (3.13)$$

where $0 \leq \text{IoU}(\mathbf{b}_1, \mathbf{b}_2) \leq 1$, such that value of 0 represents no intersection, while value of 1 represents a complete overlap. In terms of object detection or object tracking evaluation,

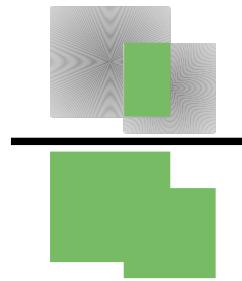


Fig. 3.7: Computation of the IoU metric between two BBOXes using of ratio of the area of overlap and the area of the union.

an IoU threshold, t , such that $0 \leq t \leq 1$, can be associated with this metric, denoting the decision boundary whether the prediction is a TP or a TN.

3.4.2 Mean Average Precision

Mean Average Precision (mAP) is commonly used for evaluation of tracking algorithms, document searching systems, object detection, object ReID, and many others. Generally speaking, it measures the success rate of an information retrieval algorithm.

Object Re-Identification

A common use case in the context of object ReID is to use mAP to assess the search results for a particular query using Euclidean distance or cosine similarity as a metric. Oftentimes the model is trained with the intent to use one of these trivial metrics. Furthermore, this approach is often paired with *top-k* accuracy, typically *top-1*, *top-2* and *top-5*.

In a typical ReID evaluation setup, there is a query set and a gallery set. For each object in the query set the aim is to retrieve a similar identity from the gallery set. The computation of the Average Precision (AP) for a query image q is thus defined as

$$\text{AP}(q) = \frac{1}{N_{gt}(q)} \sum_k P(k) \times \delta_k, \quad (3.14)$$

where $P(k)$ represents precision at rank k , $N_{gt}(q)$ is the total number of true retrievals for the query q . The indicator δ_k is equal to 1 when the matching of query image q to a test image is correct at rank r , such that $1 \leq r \leq k$. The mAP is then calculated as

average over all query images, concretely

$$\text{mAP} = \frac{1}{Q} \sum_q \text{AP}(q), \quad (3.15)$$

where Q is the total number of query images, as described in [?]. Equation ?? essentially tells us is that, for a given query q , we calculate its corresponding AP (defined in Equation ??), and then take the mean of the all these AP scores, quantifying how well our model responds to queries on average.

Object Detection

Object detectors seek to identify the presence of objects in images. The evaluation metric of such a model has to take the BBOX prediction into account, as there can be just a partial overlap of the predicted BBOX with the ground truth one. The previously defined mAP is also used for object detection [?].

In a ranked retrieval context, appropriate sets of retrieved documents are naturally given by the $top-k$ retrieved documents and for each such set, the precision-recall curve can be plotted. With this in mind, recall is defined as the proportion of all positive examples ranked above a given rank. Precision is the proportion of all examples above that rank which are from the positive class. In [?], the AP is computed for 11 equally spaced discrete recall levels, specifically $[0.0, 0.1, 0.2, \dots, 1.0]$, using

$$\text{AP} = \frac{1}{11} \sum_{r \in \{0.0, 0.1, \dots, 1.0\}} p_{\text{interp}}(r), \quad (3.16)$$

where the precision at each recall level r is interpolated by taking the maximum precision measured for a method for which the corresponding recall exceeds r . Precision interpolation is used to remove the *zig-zag* pattern by evaluating

$$p_{\text{interp}}(r) = \max_{\tilde{r}: \tilde{r} \geq r} p(\tilde{r}), \quad (3.17)$$

with $p(\tilde{r})$ representing the measures precision at a specific recall level \tilde{r} [?].

3.5 Evaluating Visual Multiple Object Tracking

When it comes to evaluating MOT performance, there is still no consensus on how to approach the evaluation and subsequent comparison of multi-object trackers. There is one established metric, however, called Classification of Events, Activities and Relationships (CLEAR) metric [?], that we will employ to quantitatively assess the performance of a MOT system. The reasons are:

- This metric is still considered a reasonably effective and intuitive metric to use, despite multiple proposals for improvements [?].
- Numerous works in object tracking, especially tracking of people, report statistics from the MOT challenges that historically have utilized this metric.
- There exist libraries (*e.g.*, [?]) allowing an evaluation of a MOT tracker inference.

Bernardin *et al.* [?], the authors of the CLEAR metric, designed few criteria that performance metrics should meet. Therefore, a useful metric:

1. allows assessing the tracker's precision regarding how well it is capable of determining the exact object location,
2. reflects the tracker's ability to track objects consistently, *i.e.*, to correctly trace object trajectories such that one and only one trajectory is established per object,
3. has as few free parameters as possible,
4. is clear and easy to interpret while emphasizing an intuitive human understanding of the tracking process,
5. is general enough so that comparison of different types of trackers is possible,
6. contains expressive values rich in information yet not abundant in quantity.

Let t denote a time for a specific frame. For each frame t , the multi-object tracker produces a set of hypotheses $\{h_1, h_2, \dots, h_m\}$ for a set of visible objects $\{o_1, o_2, \dots, o_n\}$. The evaluation procedure can be briefly described in the following pseudocode.

For each time frame t :

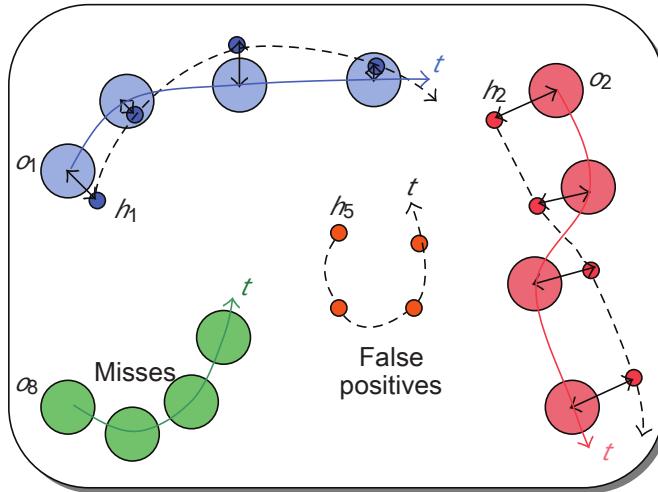


Fig. 3.8: With a demonstration of a correct tracker inference at the top, the CLEAR metric distinguishes between three fundamental types of errors, misses (false negatives), false positives and ID switches, shown in this order respectively. (source: [?])

1. Establish the best possible correspondence between hypotheses h_i and objects o_j , where $i = 1, 2, \dots, m$ and $j = 1, 2, \dots, n$.
2. For each determined correspondence between object and hypothesis:
 - (a) quantify the error in estimation of the object's position.
3. Perform accumulation of all errors (Fig. ??) in the found correspondences:
 - (a) count false negatives (misses), *i.e.*, objects without assigned hypothesis,
 - (b) count ala false positives, *i.e.*, hypotheses for which there was no object,
 - (c) count mismatch errors (swaps of object IDs), *i.e.*, situations in which the hypothesis for a given object changed compared to the previous frame.

3.5.1 Establishing Correspondences

The correspondence between a hypothesis h_i and an object o_j should not be made unless their distance (denoted as $d_{i,j}$) is within a specific threshold T . The measure of distance has to be defined for each task, but the IoU distance or Euclidean distance of BBOX centroids are most commonly used. From now on, we define object-hypothesis correspondence to be valid as long as $d_{i,j} < T$.

The value of T is critical and greatly influences the outcome. Evaluating tracking performance bears the burden of having parameters that are difficult to generalize and

the process of setting their values is often accompanied by experimentation. For example, conceptually speaking, there is, by all means, a boundary (the threshold T) beyond which we can no longer speak of an error in position estimation, but we should rather claim that the tracker has drifted away and is tracking a completely different object.

3.5.2 Tracking Consistency

To properly examine the tracker in terms of how consistent it is at tracking objects, one has to detect conflicting predictions for the given object over time. Bernardin *et al.* [?] remarked that such procedures need to decide what the “best” mapping is. For instance, assuming an object o_j and a hypothesis h_i , the “optimal” matching may be based on the initial correspondence made for o_j or the most frequent correspondence made throughout the whole sequence. If any violation is encountered, it is then treated as a discrepancy.

However, there are several issues. Consider scenarios depicted in Fig. ???. The authors raised their concerns regarding the objectivity of such evaluation and proposed a slightly different method. They only count mismatch errors once at the time frame where the change occurs and consider the remaining intermediate correspondences as correct.

Let $M_t = \{(h_i, o_j)\}$ be the set of mappings made up to time t , such that $M_0 = \{\cdot\}$. Once a new correspondence is made at the next step at time $t+1$ between the hypothesis h_k and the object o_j that conflicts the already established identity by the pair (h_i, o_j) in M_t , this contradiction is then counted as a mismatch error and (h_i, o_j) is replaced by (h_k, o_j) in M_{t+1} . Consequently, mapping that is constructed this way enhances decision-making when facing multiple competing hypotheses for the same object. The implicit assumption is that the previously assigned hypothesis is more likely to be correct than the new one, even if the distance metric alone would indicate otherwise (Fig. ??).

3.5.3 Mapping Procedure

Let $M_0 = \{\cdot\}$. For each time frame t :

1. Verify if mappings (h_i, o_j) in M_{t-1} are still valid. A pair is deemed valid as long as the hypothesis h_i exists at time t , the object o_j is still visible while the distance between the two does not exceed T . If these conditions hold, establish a correspondence.
2. If there are objects for which no correspondence has been made so far, then a suitable

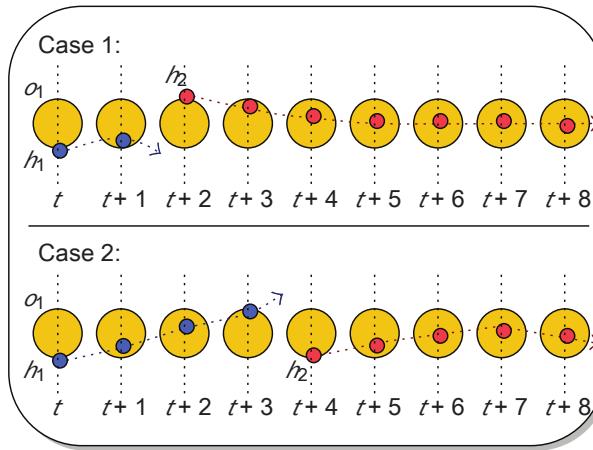


Fig. 3.9: Illustration of the inherent “unfairness” when relying on sequence-level “best” object-hypothesis mapping induced by the most frequent correspondence. As shown in the case 1, the correct hypothesis is the h_2 , and thus only 2 errors are incurred for the first mismatch. The case 2 is practically identical, the h_2 also represents the most common assignment. However, 4 errors are accumulated for the alleged mismatch for h_1 . (source: [?])

matching hypothesis is searched for. This step involves one-to-one matching for pairs the distance of which does not exceed the threshold T . The matching procedure is formulated as a minimum cost assignment problem. In case there happens to be a correspondence that contradicts a mapping $[h_i, o_j]$ as part of M_{t-1} , then replace the previous pair $[h_i, o_j]$ with $[h_k, o_j]$ and treat such an occurrence as a mismatch error. For simplicity, let mme_t be the number of the mismatch errors for the frame t .

3. The two previous steps guarantee that a complete set of matching pairs has been generated for the current time t . At this point, we may start calculating values that will be utilized later for computing the final metrics. So, let c_t be the number of matches found for time t . For each such match, compute the distance between the object o_j and the corresponding hypothesis, denoted by d_{ti} .
4. Every hypothesis that is not part of any pair up to this point is reckoned as false positive. Likewise, all the remaining objects are marked as misses. Thus, let fp_t and m_t be the number of false positives and misses, respectively. For future reference, let us define g_t as the number of ground-truth objects visible at time t .

3.5.4 Performance Metrics

In light of the previously described procedure, here we present the two most relevant performance metrics by which the tracking performance can be intuitively expressed,

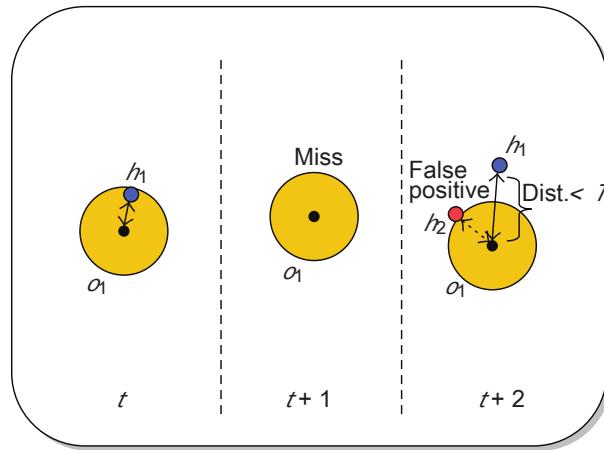


Fig. 3.10: Track reinitialization. At time t , the identity of the object o_1 is accounted for by the hypothesis h_1 . At time $t + 1$, the object disappears and the track is temporarily lost. At time $t + 2$, the tracker is responsible for reinstating the object identity. During evaluation, the underlying assumption is that the previous hypothesis should be the correct one, even if the new hypothesis is closer according to the used distance function. (*source: [?]*)

namely the “tracking precision” and “tracking accuracy”.

The role of tracking precision is to measure the alignment between the predicted object position and the ground-truth position only for the positive sample. For that reason, precision is not influenced by the ability (or lack thereof) of the tracker to detect objects. Therefore, the Multiple Object Tracking Precision (MOTP) metric can be defined as

$$\text{MOTP} = \frac{\sum_{\forall t} \sum_{\forall i} d_{ti}}{\sum_{\forall t} c_t}, \quad (3.18)$$

representing the total error in the estimated position for the pairs where the object-hypothesis relationship was correctly determined averaged over the total number of such matches made.

Conversely, the accuracy metric attempts to reflect the number of mistakes the tracker made in terms of misses, false positives, object mismatches. Given this description, the Multiple Object Tracking Accuracy (MOTA) metric can be expressed as

$$\text{MOTA} = 1 - \frac{\sum_{\forall t} (m_t + fp_t + mme_t)}{\sum_{\forall t} g_t}, \quad (3.19)$$

the possible values of which lie within the interval $[-\infty, 1]$.

We would like to emphasize that the errors have to be first summed up across all the frames before computing the ratios rather than evaluating the ratio locally. Independent computation of the given ratios would lead to non-intuitive outcome (Fig. ??).

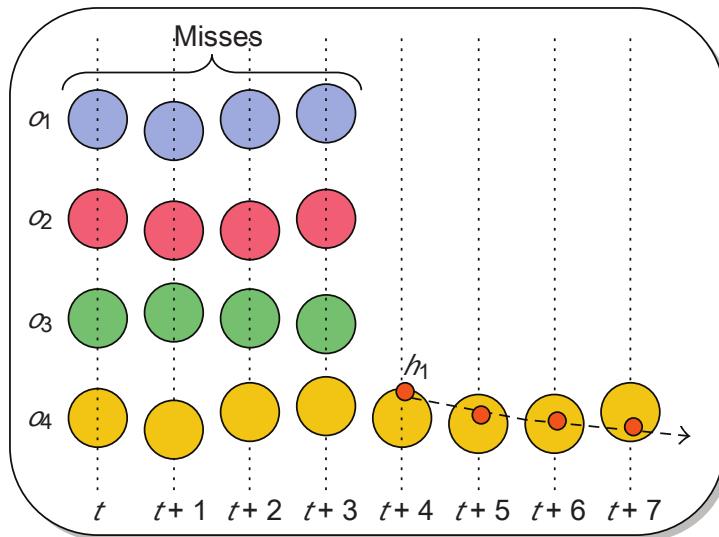


Fig. 3.11: Computing of error ratios needs to be performed on a global level, rather than on a local, frame level. Assume a sequence consisting of 8 frames. Moreover, assume that objects o_1, \dots, o_4 are visible on the frames from t_1 to t_4 , but none of them is being tracked. The situation changes at frame t_4 where only the object o_4 is being tracked properly by the hypothesis h_1 . As a result, in frames t_1, \dots, t_4 , the resulting miss rate is 100%, whereas in frames t_5, \dots, t_8 it is exactly 0%. Applying arithmetic average to these values yields a global miss rate of $\frac{1}{8} (4 \cdot 100 + 4 \cdot 0) = \frac{1}{2}$, or, 50%. Conversely, performing summation prior to quantifying the final global ratio produces far more intuitive result of 16 out of 20 misses, or the miss rate of 80%. (source: [?])

3.5.5 Other Performance Metrics

Besides the two primary performance metrics discussed in the previous section, there are also partial metrics that are worth evaluating to get a better grasp of the tracker's performance. For our purposes, we computed metrics outlined in Table ??.

3.6 Single Object Tracking

3.6.1 Initial Deep Learning-Based Solutions

At the time of publishing the work of Held *et al.* [?], most generic object trackers required online training from scratch, without taking advantage of available datasets to at least provide a starting point by initial offline training. This was the incentive behind the development of the famous Generic Object Tracking Using Regression Networks (GOTURN) [?]. This approach used to be SOTA in single-object tracking, but nowadays it is considered obsolete. A major issue is that the object has to be located initially, and occlusion handling is not performed as well as management of abrupt changes in position.

Metric Name	Description
no. of frames	Total number of frames.
no. of matches	Total number matches.
no. of switches	Total number of track ID switches.
no. of false positives	Total number of false positives (false alarms).
no. of misses	Total number of misses (false negatives).
no. of detections	Total number of detected objects including matches and switches.
no. of objects	Total number of unique object appearances over all frames.
no. of predictions	Total number of unique prediction appearances over all frames.
no. of fragmentations	Total number of switches from tracked to not tracked.
no. of mostly tracked	Number of objects tracked for at least 80% of lifespan.
no. of partially tracked	Number of objects tracked with lifespan from 20% to 80%.
precision	Number of detected objects over sum of detected and false positives.
recall	Number of detections over number of objects.
IDF1	ID measures: global min-cost F1 score.
MOTP	Multiple object tracker precision.
MOTA	Multiple object tracker accuracy.

Table 3.3: Other important CLEAR metrics that we adopted for evaluation of our experiments with various MOT approaches.

So it is common for the object to drift away. Nevertheless, it stands to reason that the notion of leveraging data for offline training has pervaded the VOT community ever since. Nowadays, it is scarce to find a tracker that is trained online purely from scratch.

Given an initial state in a form of a BBOX belonging to the first frame (a search region), the network then crops a new region in the next frame and tries to find the location of the target object within this region. It practically performs a comparison of the current search region given the predicted target location from the previous frame (Fig. ??). A key concept to highlight is that GOTURN addresses the tracking as a box regression problem.

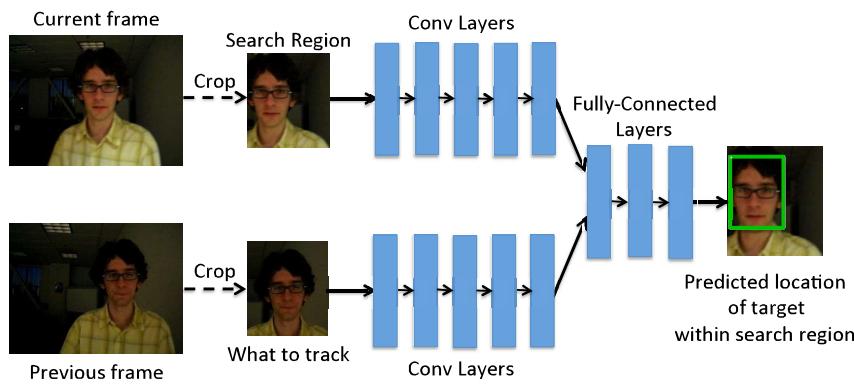


Fig. 3.12: The architecture of GOTURN showing the input as a search region from the current frame where the target from the previous frame is to be searched for. The network essentially learns to localize the target within the cropped region. (*source: [?]*)

3.6.2 Fully Convolutional Tracking

Transfer learning, *e.g.*, exploiting an already pre-trained CNN model to extract visual features, often comes with one drawback: the model accepts only a fixed input size. Although newer architectures can handle variable input size, this trait is more prevalent in object detection and segmentation than in the basic task of image classification. A common approach is to resize the image to the required shape, but this may significantly distort important features. Using fully connected layers demands known dimensions in advance, which is complicated to acquire when dealing with the input of diverse shapes. Convolutional layers are invariant to input size, therefore an avoidance of fully connected layers may provide an answer. An efficient solution to replace fully connected layers utilizes 1×1 convolutions was notably propagated in Network In Network model [?]. 1×1 filters were also used in the Inception architecture for dimensionality reduction and at the same time to increase the dimensionality of feature maps [?].

The CNNs provide valuable spatial clues about the image content. Thus, interclass variations are thoroughly captured in the top layers, and intraclass variations conversely in the bottom layers (Fig. ??). This concept led the authors of [?] to propose a fully convolutional visual object tracker that exploits different layers of the pre-trained VGG network [?]. As a result, the model responsible for extracting visual features is no longer treated as a black box. An in-depth study was conducted on the properties of CNN features of the offline pre-trained model for the task of classification of ImageNet dataset [?]. It was found out, as suggested above, that characterization from different perspectives is provided by convolutional layers at different levels.

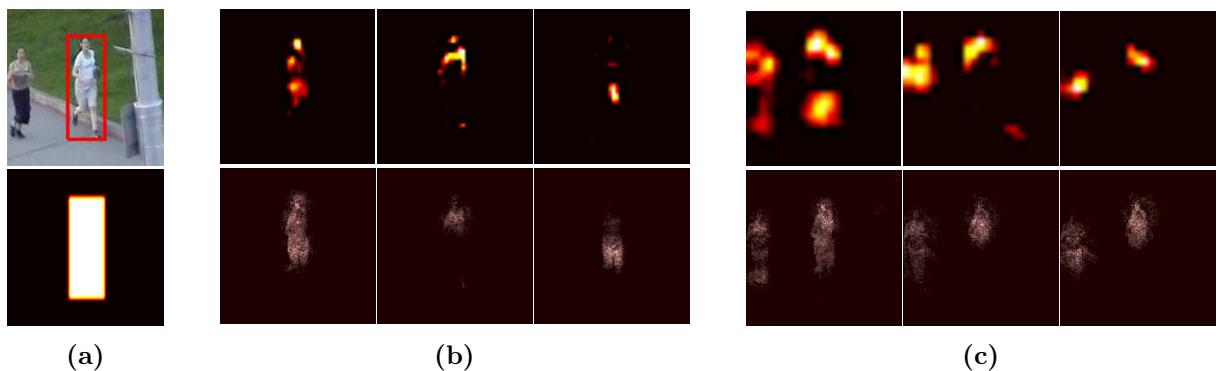


Fig. 3.13: (a) Input image with an associated ground truth mask. (b) Visualization of feature maps from convolutional layers placed at the bottom of the model, capturing differences between foreground and background of a particular object instance. (c) As opposed to the previous group of images, a more holistic, abstract view on the object category itself is provided by feature maps from top convolutional layers. The top row in the (b) and (c) represents feature maps, whereas the bottom row represents the corresponding saliency map with spatial information of the category. (*source: [?]*)

Authors of [?] put together a list of three observations that summarize properties of the fully convolutional nature of a tracker proposed by them.

- Despite a large receptive field of CNN feature maps, few of them are activated and they are sparsely distributed and localized.
- The majority of the feature maps can be considered noisy or irrelevant when discriminating a specific target object (foreground) from the background.
- Different layers encode different types of features (related to the intraclass or inter-class variations discussed at the beginning).

3.6.3 Tracking Using Siamese Networks

Although CNNs condense valuable visual information into low dimensional space, it is still not sufficient in many situations during object tracking. The object representation from convolutional layers trained on image classification is not robust enough for dramatic visual changes and occlusion of varying intensity. As we discussed in Section ?? dedicated to metric spaces and embeddings, an object representation supporting ReID requires different types of models, *i.e.*, a *Siamese neural network* (Section ??). We have already mentioned our intention of utilizing custom metric space for tracking, and [?] were among the first ones to successfully demonstrate it.

Authors of [?] approved of the idea that visual feature extraction using CNNs is pertinent to the robustness of the tracking algorithm, yet they advocated to train the visual model to a more general task of similarity learning rather than just classification. This observation and its further implementation was the main contribution of their work, achieving SOTA performance back then. Broadly speaking, they trained a *fully convolutional* Siamese network to locate an *exemplar* (also referred to as a *template* or *target* in the literature, but for clarity, we will stick to *exemplar* only) image within a larger *search* image (Fig. ??). The model got the name Siamese Fully Convolutional Network (SiamFC). We mentioned this to make the comparison easier because a lot of follow-up works have been done, producing models such as Semantic-Appearance Siamese Network (SA-Siam) [?], Siamese Region Proposal Network (SiamRPN) [?], Siamese Network with Segmentation Mask (SiamMask) [?], Siamese Network with Segmentation Mask and Ellipse Fitting (SiamMask-E) [?], and so forth.

Let γ be a transformation that extracts visual features from the input, and g be the function that combines two representations produced by the function γ . Siamese networks apply an identical transformation γ to both inputs, search image x and exemplar image z , and combine the result as

$$f(x, z) = g(\gamma(x), \gamma(z)). \quad (3.20)$$

As such, when trivial distance or similarity measure is computed by the function g , then γ can be deemed as the already introduced embedding. The output of computing the cross-correlation operation is the response (confidence) map. During the training, however, the ground-truth response map is centered at the target object and the cosine function is used to create a 2D penalty map. The reason is to focus the response to the center under the assumption that the object is always focused in the center of the current search region. In practice, this assumption holds most of the time, that's why adoption of the so-called *cosine window* is prevalent in Siamese trackers (Fig. ??).

A team of authors in [?] made the following observation: features learned in an image classification task (denoted as semantic features) complement features learned in a similarity matching task (denoted as appearance features). They also suitably commented that the key to designing a high-performance tracker is to utilize expressive features that are simultaneously discriminative and generalized. In light of this, they developed a model

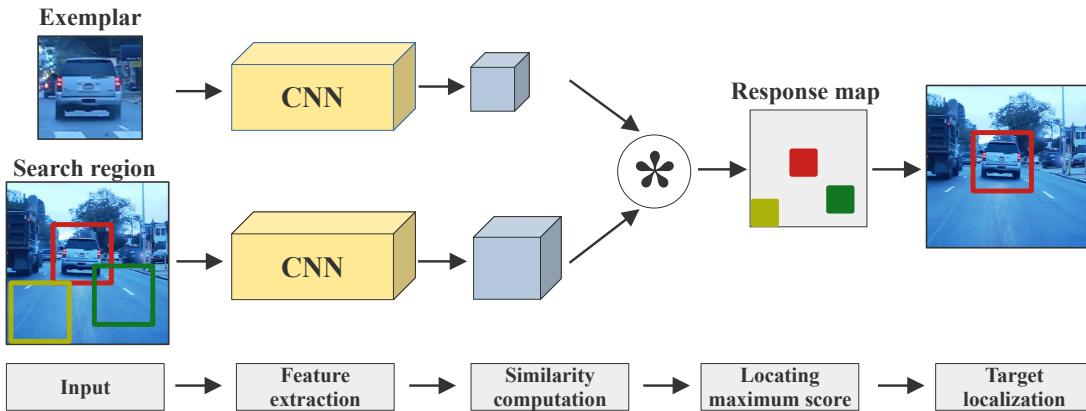


Fig. 3.14: The fully convolutional Siamese architecture produces a scalar-valued score map. The similarity function is computed for all sub-windows within the search image and stored in a 2D score map, rather than just a pure 1D embedding vector. This computation requires only one evaluation. In this image, the red and blue pixels in the output score map represent similarity values for the two sub-windows on the input.

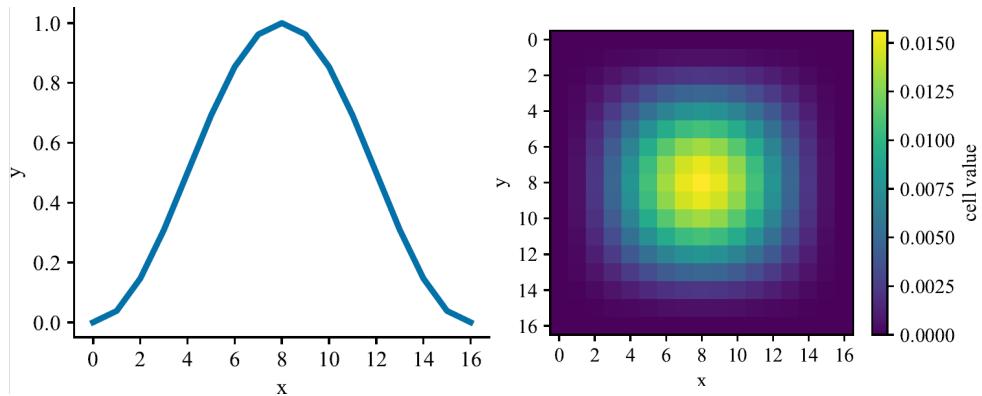


Fig. 3.15: A visualization of 1D (left) and 2D (right) cosine window under the assumption that the response map is of size 17×17 . Such a mask conveys the idea of putting the highest weight to the center with nonlinear, even reduction when moving away from the center.

consisting of a semantic and an appearance branch, with each branch being represented by a standard similarity-learning Siamese network (as in SiamFC [?]). However, an important distinction is that these two branches were trained separately, making them effectively heterogeneous to avoid any sharing of information. They reported that both branches were less powerful when trained jointly than when trained separately. The reasoning was that each branch provides different features produced at different levels of abstraction, yet they complement each other. The merge of their respective outputs happens only during the testing time. Nowadays, joint training is prevalent, especially due to its effectiveness. Given the advantage of hindsight, there are more important aspects of Siamese trackers to address to reap even greater benefits in terms of accuracy, for example, feature fusion.

The SA-Siam receives an input as a pair of image patches cropped from the initial

(exemplar) frame and the current (search) frame. Let z , z^s and X be the image of exemplar, exemplar including the surrounding context and the search region, respectively. Dimensions of x^s and X are identical, $W_s \times H_s \times 3$. Dimensions of the exemplar z located in the exact center of the region of z^s are $W_t \times H_t \times 3$, such that $W_t < W_s$ and $H_t < H_s$. The appearance branch (Appearance Network (A-Net)) takes (z, X) as input and essentially clones the entire SiamFC network. Let $f_a(\cdot)$ denote the visual features extracted by the A-Net. Then, the response map of this branch is given by

$$h_a(z, X) = \text{corr}(f_a(z), f_a(X)), \quad (3.21)$$

where $\text{corr}(\cdot)$ is the correlation operation. The training of this branch is rather straightforward thanks to the abundance of training pairs (z_i, X_i) with their related response map Y_i , for $i = 1, \dots, N$, with N being the no. of chosen training samples. Analogically, the semantic branch (Semantic Network (S-Net)) assumes as input a pair (z^s, X) . Contrary to the A-Net, this model is pre-trained for the image classification task and its weights are frozen during the training phase. These last two convolutional layers of this model are of primary interest as their features provide abstraction at distinct levels. However, spatial resolutions are not alike. Let $f_s(\cdot)$ be the concatenated multilevel features. For the correlation operation ($\text{corr}(\cdot)$) to be usable, a special fusion module is introduced, implemented by a simple 1×1 convolution layer. The fusion operation is applied to features within the same layer, and this fused feature vector will be referred to as $g(f_s(X))$.

Semantic features of a higher level are robust to appearance variation. This contributes to the generalization ability of the tracker but exacerbates its discriminative abilities. To circumvent this, the attention module is presented. The reasoning is that individual feature channels have varying importance for object tracking as far as different exemplars are concerned. The goal is then to assign a degree of importance (weight) to each channel for each exemplar. Still, the exemplar information is not sufficient, so the context must be supplied, too. The proposed attention module thus processes the feature map of z^s instead of just z . The attention module operates channel-wise and incurs negligible computational overhead as it's only active during the target processing on the first frame. Later on, the weight coefficient is used to scale each feature map according to its importance.

When training the S-Net branch, only the fusion and the attention modules are updated. No fine-tuning techniques are taken advantage of, regardless of the potential

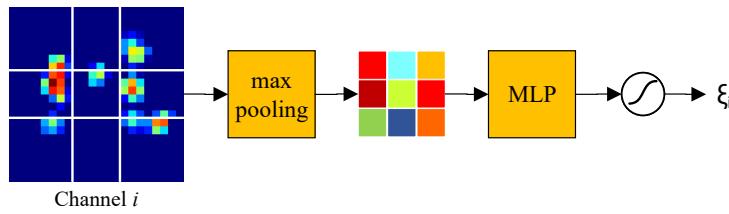


Fig. 3.16: The attention module of the S-Net network. (source: [?])

improvement of the semantic branch alone. Authors informed about such experiments, and they resulted in diminished overall performance thanks to A-Net and S-Net becoming less heterogeneous. The inference phase involves computation of the overall heat map for which a weighted average of the two produced heat maps.

The series of Siamese-based architectures for tracking continued with the idea of using the RPN [?] (see Section ?? for the same concept applied in object detection). Under the flag of end-to-end training, the SiamRPN model consists of a Siamese subnetwork for feature extraction (again, a duplicate of the SiamFC [?]) and RPN as another subnetwork encompassing both classification and regression branch (Fig. ??). The notable contribution is that the proposed framework is formulated as a local one-shot detection task in the inference phase (the first work to make such a step). The template branch encodes the object appearance information for further foreground/background discrimination. Analogically, the BBOX from the first frame is the only exemplar for one-shot detection in the inference phase.

The region proposal subnetwork contains a pair-wise correlation section as well as a supervision section. Let k denote the number of anchors. Then, the model has to output $2k$ channels for the classification and $4k$ channels for the regression. Following the established notation, the Siamese subnetwork produces feature maps $\gamma(z)$ and $\gamma(x)$. The pair-wise correlation splits $\gamma(z)$ into $[\gamma(z)]_{cls}$ and $[\gamma(z)]_{reg}$ while increasing the no. of channels (Fig. ??). Conversely, $\gamma(x)$ is also split into $[\gamma(x)]_{cls}$ and $[\gamma(x)]_{reg}$, but the no. of channels remains unchanged. The correlation, when computed on both branches, is given by

$$\begin{aligned} A_{w \times h \times 2k}^{cls} &= [\gamma(x)]_{cls} \star [\gamma(z)]_{cls}, \\ A_{w \times h \times 4k}^{reg} &= [\gamma(x)]_{reg} \star [\gamma(z)]_{reg}, \end{aligned} \quad (3.22)$$

where the template feature maps $[\gamma(z)]_{cls}$ and $[\gamma(z)]_{reg}$ stand in place of kernels in the convolution operation signified by the \star character.

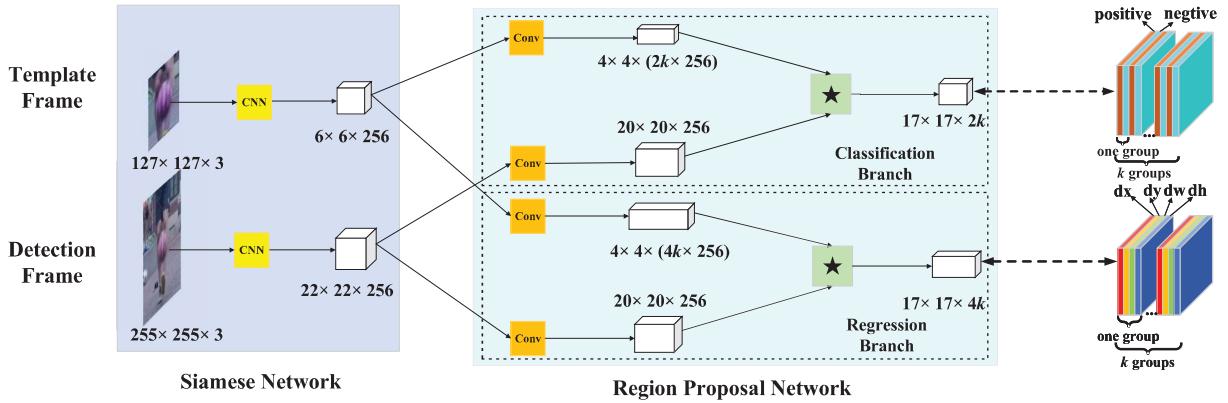


Fig. 3.17: The pipeline starts with the original SiamFC network followed by the RPN which has two branches: classification and regression. The output of the two branches is obtained using a pair-wise correlation. Foreground/background classification and the box regression are given by the $17 \times 17 \times 2k$ and $17 \times 17 \times 4k$ feature maps, respectively. (*source: [?]*)

The noteworthy formulation of tracking as one-shot detection was proposed as follows. In general terms, the goal is to minimize the average loss \mathcal{L} of a predictor function $\psi(x; W)$ by finding its parameters W . When computed over a dataset of N samples x_i with corresponding labels y_i , $\forall i = 1, \dots, N$, it is given by

$$\arg \min_W \left\{ \frac{1}{N} \sum_{i=1}^N \mathcal{L}(\psi(x_i; W), y_i) \right\}. \quad (3.23)$$

One-shot learning aims to learn W when only a single exemplar z is available, tackling a major challenge of *learning to learn* [?]. If we consider a meta-learning feed-forward function ω that maps $(z_i; W')$ to W , then the problem can be stated as

$$\arg \min_{W'} \left\{ \frac{1}{N} \sum_{i=1}^N \mathcal{L}(\psi(x_i; \omega(z_i, W')), y_i) \right\}. \quad (3.24)$$

In this setting, this objective function can be re-written in terms of the Siamese subnetwork feature extraction γ and region proposal subnetwork Ψ as

$$\arg \min_W \left\{ \frac{1}{N} \sum_{i=1}^N \mathcal{L}(\Psi(\gamma(x_i; W); \gamma(z_i; W)), y_i) \right\}. \quad (3.25)$$

The template branch provides training parameters to predict the kernel for the detection task. The template branch embeds necessary category information into the kernel that is subsequently utilized for detection.

Later on, a fork of publications emerged with an endeavor to improve the tracking

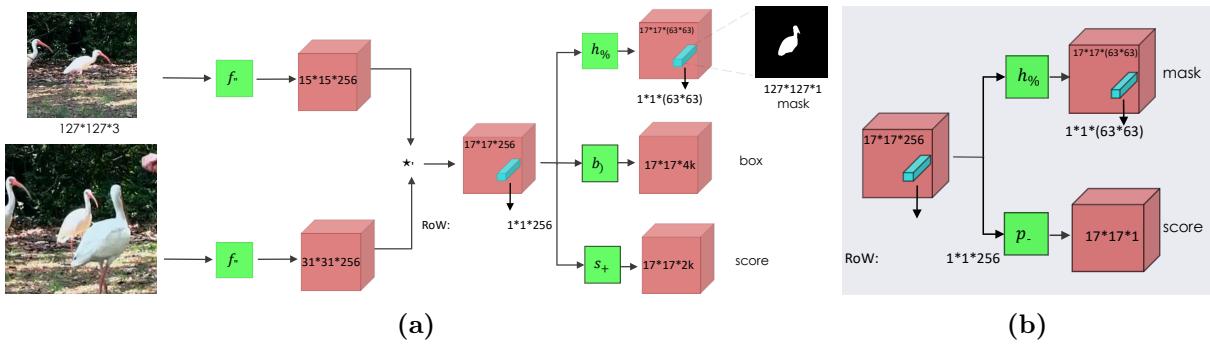


Fig. 3.18: Illustration of the two SiamMask variants: **(a)** a three-branch architecture (full), **(b)** a two-branch architecture (head), both with depth-wise correlation layers. (*source: [?]*)

performance by estimating not only a regular axis-aligned BBOX but a rotated box, too. Put into perspective, the rotated BBOX, as opposed to an ordinary, axis-aligned, contains the minimal amount of background pixels [?]. Thus, datasets with rotated BBOXes provide tighter enclosed rectangles.

Inspiration from object segmentation yielded another approach where the tracking process was assisted with additional semi-supervised object segmentation [?]. The relevant contribution is the augmentation of the training loss with a binary segmentation task. Further, once trained, the model (dubbed SiamMask) relies exclusively upon a single BBOX initialization and operates online while producing rotated BBOXes instead of axis-aligned ones together with class-agnostic object segmentation masks.

Again, the SiamFC [?] served as the fundamental building block. However, a notable modification consisted of the use of a depth-wise cross-correlation layer instead of a simple cross-correlation layer. The latter compresses all the information into a single channel, impeding the potential to encode richer information about the object. As a reminder, the original model used $6 \times 6 \times 128$ and $22 \times 22 \times 128$ tensors to produce a $17 \times 17 \times 1$ response map (Fig. ??). Here, a multi-channel response maps are utilized (Fig. ??).

An incremental improvement of SiamMask model came when [?] proposed a novel, efficient algorithm for the estimation of the BBOX rotation when the object segmentation mask is given. Particularly, a mask produced by the SiamMask model, as this work builds on top of [?], under the derived name SiamMask-E. In addition, their approach can be used to generate a rotated box ground truth from any segmentation datasets to train a rotation angle regression model. To estimate the rotation angle, they adopted the least-squared scheme as part of the ellipse fitting algorithm.

The idea to employ fully convolutional networks seems to pertain to the modern com-

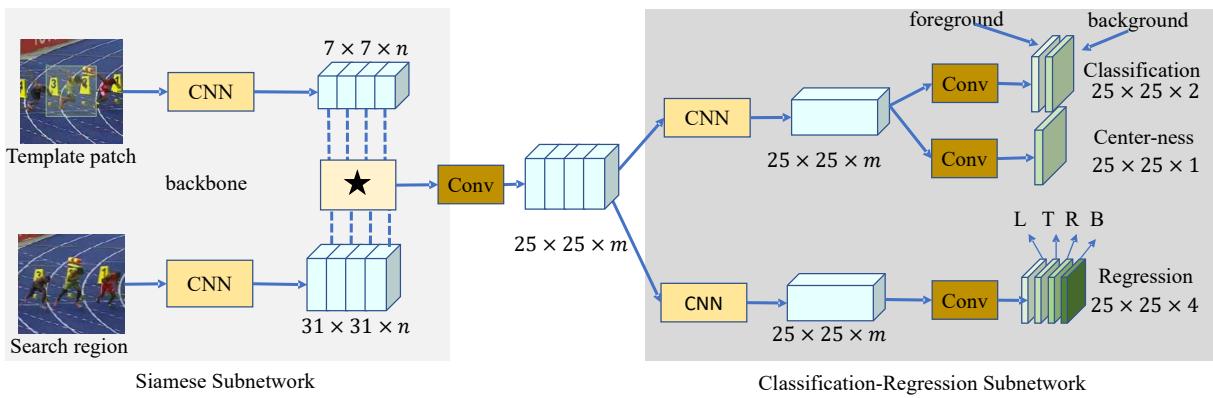


Fig. 3.19: SiamCAR architecture. The left side consists of the original SiamFC [?] model, with a simple amendment of using depth-wise correlation for multi-channel response map extraction. The right side depicts the subnetworks for foreground/background classification and BBOX regression. (source: [?])

puter vision community. Besides a simpler model, the fully convolutional design often leads to a reduced number of hyperparameters. One such an architecture (a descendant of the famous SiamFC [?] model) has been recently proposed, named Siamese Classification and Regression Network (SiamCAR) [?]. This approach relies on the decomposition of the task of VOT into two subproblems: classification for pixel category and regression for object BBOX at the given pixel. The leading concept of the article is that this tracker operates in an end-to-end, per-pixel manner. The authors managed to avoid the use of anchors as well as region proposals, hence reducing the need for human intervention. The use of the two aforementioned traits commonly leads to sensitivity to dimensions and aspect ratios of the anchor boxes, which requires expertise in hyperparameter tuning.

An indispensable part of localization is low-level features like edges, corners, and so on, whereas high-level features strengthen the representational power from the semantic point of view, which is crucial for discrimination. Authors fused low-level and high-level features from the last 3 residual blocks of the Residual Neural Network (ResNet)-50 backbone, forming a unity after concatenation.

An important observation was made that locations further away from the object center may aggravate the predicted box as they can be considered low-quality. To diminish the effect of such locations, another branch alongside the classification branch to suppress the outliers is introduced, based on the concept of *centerness*, borrowed from the [?]. This branch outputs a feature map where each point indicates the *centerness* score for the corresponding location. We can say that *centerness* is a very general concept, and practically it represents a weighting mechanism to penalize areas within the Region of

Interset (ROI) that most likely do not contain the target object.

Conclusions Made In the Survey Paper

We would like to remark that this branch of trackers formed the basis of our research. Its importance reached such a high level that we even composed an up-to-date comprehensive survey paper [?] solely focused on Siamese trackers and their fundamentals. Due to space limitations, in this short section, we summarize the most important conclusions and observations that we made in the aforementioned paper.

In the referred survey, we aimed to identify and elaborate on the most significant challenges the Siamese trackers face. The objective was to answer what design decisions the authors had made and what problems they had attempted to address. That treatise could be thought of as an in-depth analysis of the core principles on which Siamese trackers operate together with a discussion of the underlying motivation. In addition, we also provided an up-to-date qualitative and quantitative comparison of the prominent Siamese trackers on established benchmarks, since the last survey that involved a thorough discussion of Siamese trackers was published in 2018 [?]. Last but not least, we discussed the current trends in developing Siamese trackers at the time of writing the article as well.

We have to emphasize that Siamese trackers are a research direction in VOT with great potential. In practical terms, they belong to the fastest trackers with the “accuracy-to-speed” ratio being their primary strength. Contrary to the initial expectation, we realized that fast trackers were also among the most accurate ones (with some existing exceptions). Simply put, high processing speed is an inherent property of Siamese trackers. Nevertheless, there are existing drawbacks that require research attention. The presence of distractors (similar interference) is in our paper often mentioned as one of the leading causes of problems for this type of trackers. Our quantitative evaluations indicate that trackers where the presence of semantic background is explicitly treated often yield the top performance. Siamese metric learning is powerful enough to encompass numerous visual variations, but in case there are distractors present, then additional steps conditionally executed seem to contribute positively. To name a few, there are the explicit distractor-awareness [?], custom sampling strategies for foreground/background discrimination [?], or conditional object re-detection [?]. Besides, there are a plethora of examples where RPN was used for object proposals even for Single-Object Tracking (SOT) in Siamese trackers.

We venture to claim that the top-performing trackers exploited the above-mentioned RPN head, *e.g.*, [?, ?, ?]. A comprehensive survey concerning deep visual object tracking by Marvasti-Zadeh *et al.* [?] also reached a similar conclusion.

The utilization of cross-correlation has a great share of the leading performance in terms of its effectiveness. But the original single-channel formulation from [?] has been improved into a multi-channel, depthwise cross-correlation that has been in use up to date. It was argued that a single channel did not capture sufficient information [?], thus multi-channel cross-correlation layers were used instead [?]. On top of that, since multiple channels are present, we observed an emerging trend in using various attention mechanisms to aid the feature selection [?].

Speaking of cross-correlation in terms of its core principle of performing a “learned template matching” using the exemplar and the search region, it raises the question of whether and how the exemplar template should change during the training. Several works have remarked that incorporating memory or template updating strategies could potentially enhance the tracker performance, *e.g.*, [?, ?]. It seems that relying solely upon the exemplar image from the initial frame may have detrimental effects as the object undergoes severe visual deformations, so the tracker may eventually lose its track.

Even though our discussed survey focused on SOT, there are emerging works where Siamese architectures were integrated into a MOT pipeline.

3.7 Multiple Object Tracking

Our research originally targeted SOT, especially Siamese trackers. The plan to incorporate multiple objects remained only as a hypothesis to explore later. However, thanks to our comprehensive survey on Siamese tracking [?], we gained enough background knowledge to quickly absorb the newly emerging body of literature on a specific branch of multi-object trackers that exploit Siamese architectures. Our research is focused on Siamese neural networks, whereas the MOT is dominated by approaches that utilize detections + linking based on solutions exploiting a wide range of methods, from simple Munkre’s algorithm [?] through complicated graph formulations [?] to even graph-based convolutional neural networks [?]. Even though there are works that claim the use of Siamese neural networks in MOT, *e.g.* [?], their utilization serves for the ReID within the tracking-by-detection

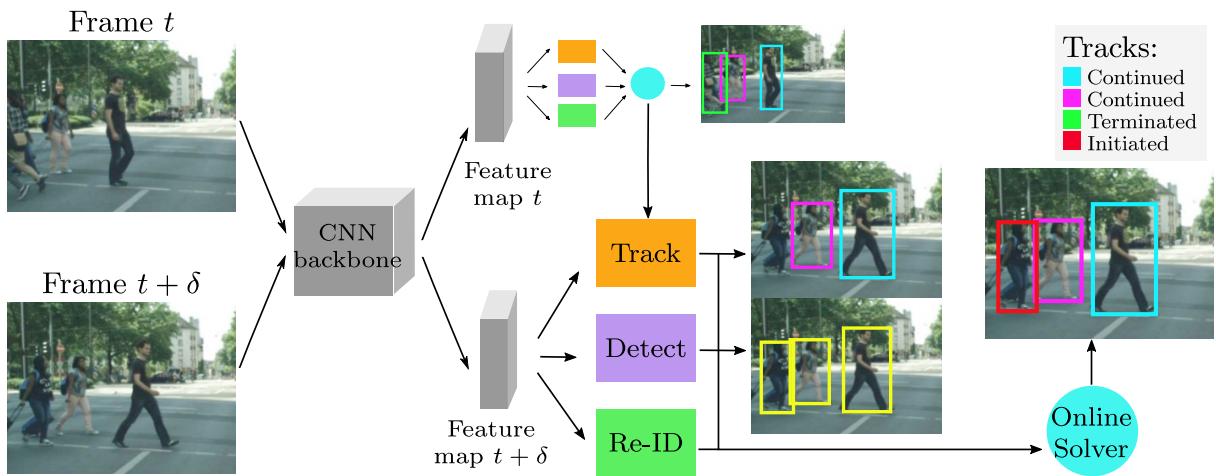


Fig. 3.20: Demonstration of how unification of the detection, tracking and ReID within a single architecture can be achieved. (*source: [?]*)

philosophy, for which Siamese networks are widely adopted. However, by Siamese tracking, we explicitly mean the type of trackers described in Section ???. We identified that Siamese-based MOT is a freshly rising subfield of trackers.

3.7.1 Siamese-based Multiple Object Tracking

Shuai *et al.* [?] proposed a Siamese-based framework that can simultaneously handle object tracking, detection, and ReID (Fig. ??). The unification of all these aspects into a single pipeline is a significant advantage. In addition, the formulation allows the use of any Siamese tracker. Although this tracking system follows an inference pipeline similar to other tracking-by-detection systems, the distinction is that it does so based on features generated by a single network.

A very effective extension aptly dubbed as Siamese Multi Tracker (SiamMT) of the often-mentioned came in [?] where the SiamFC tracker was utilized n exemplars to produce n response maps and, therefore, to perform tracking of n objects simultaneously (Fig. ??). The incentive to develop this tracker was to address the problem with running a costly detector for every frame to produce detections upon which another performance-demanding linking stage is usually executed. This framework was the first to demonstrate the qualities of a purely deep learning-based, end-to-end tracking pipeline capable of tracking multiple arbitrary objects at once.

The endeavor to exploit Siamese neural networks to assess the degree of similarity between two objects has spurred a plethora of proposals combining various mechanisms.

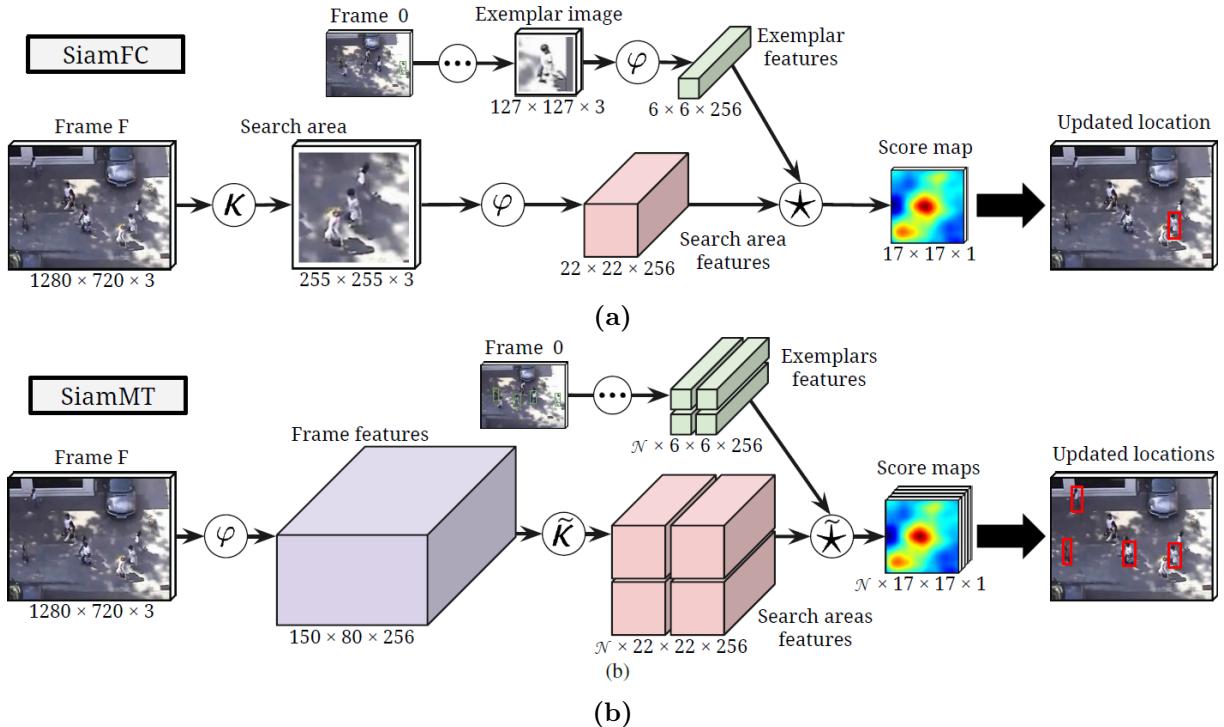


Fig. 3.21: The inference phase of the (a) SiamFC architecture to see the difference between the (b) SiamMT successor. The SiamMT framework first extracts features of the entire frame via the backbone φ . Subsequently, the obtained features are cropped and resized using the \tilde{K} operator, utilizing ROI-align operations. Finally, all these features are combined in the traditional cross-correlation way (although slightly adjusted to handle more objects) to produce a multi-object response map. (source: [?])

Lee *et al.* [?] combined Siamese similarity learning with Feature Pyramid Networks (FPNs) (Section ??). This tracker still follows the path of the tracking-by-detection paradigm, in which the similarity metric between the current detections and existing tracks plays an essential role. In this work, criticism was raised concerning the plain Siamese architectures for not being sufficient for tracking owing to their structural simplicity and lack of motion information. To address the structural simplicity, a Feature Pyramid Siamese Network (FPSN) was proposed. Then, to overcome the lack of motion information, additional spatiotemporal motion features were added to the FPSN module.

As a matter of fact, our research ended up working with Siamese Multi-Object Tracker (SiamMOT) [?] (Section ??) architecture, which we will introduce in great detail later on. It is a multi-object tracker that encompasses some of the best approaches we have discussed so far into an end-to-end framework, such as Siamese tracker (multi-channel cross-correlation), RPN head, *centerness*, feature fusion, and much more.

3.8 Feature Extraction and Feature Fusion

As we have observed in our survey of Siamese trackers [?], incremental improvements in feature extraction were often the major contribution of numerous works. With this in mind, we consider feature extraction a necessary part of any deep learning model design.

3.8.1 Residual Neural Networks

He *et al.* [?] aptly remarked that deeper neural networks are more difficult to train. In this work, a residual learning framework to facilitate easier training of neural networks that were significantly deeper than their previously used counterparts was proposed. The explicit reformulation of the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions, led to a breakthrough in the utilization of deep neural networks.

The foundation of ResNets is the adoption of skip connections that represent shortcuts to jump over certain layers. Typically, such models are implemented using double or even triple layer skips containing nonlinearities (*e.g.*, Rectified Linear Unit (ReLU)) and batch normalization [?] in between. The primary reason for adding skip connections was to avoid vanishing gradient problems. As demonstrated in Fig. ??, the degradation problem manifests itself out in deeper networks when their accuracy shows signs of saturation followed by a rapid decline, but not as a result of overfitting.

Let $H(\mathbf{x})$ denote the desired underlying mapping. The stacked nonlinear layers are then expected to fit a mapping $F(\mathbf{x}) = H(\mathbf{x}) - \mathbf{x}$. The original mapping is reformulated as $H(\mathbf{x}) = F(\mathbf{x}) + \mathbf{x}$. The initial hypothesis, which turned out to be correct, was that it is easier to optimize the residual mapping instead of the original, unreferenced mapping.

3.8.2 Feature Pyramid Networks

Feature Pyramid Network (FPN) [?] is an extension to existing backbones used for feature extraction serving various tasks ranging from image classification, object detection, object tracking or even image segmentation. Its greatest strength is the combination of low-resolution, semantically strong features with high-resolution, semantically weak but discriminative features via a top-down pathway and lateral connections.

Fig. ?? compares competing methods of feature aggregation by their core principles.

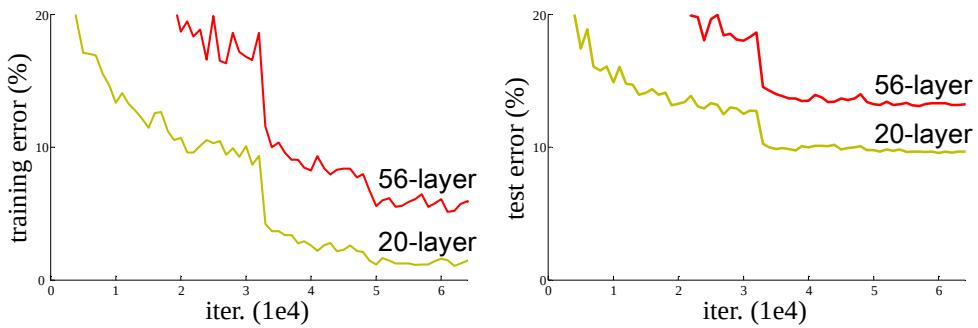


Fig. 3.22: A motivation behind the ResNets. The training error and the test error are greater for the deeper model than for the shallower model. Therefore, the inevitable conclusion is that in order to learn better networks, it takes more than just stacking more layers. (*source: [?]*)

Regarding the FPN itself, observe the two pathways in Fig. ?? (d). The bottom-up pathway represents a feed-forward computation of the backbone (*e.g.*, a basic CNN), where one pyramid level corresponds to one stage. The output of the last layer of each stage will serve the purpose of enriching the feature maps when processing the top-down pathway by the use of lateral connections. The top-down pathway consists of upsampling operations followed by an application of 1×1 convolutions to align tensor channels dimensions and then element-wise addition of features.

3.8.3 Deep Layer Aggregation

A successor of the previously discussed FPN is the Deep Layer Aggregation (DLA) [?]. This architectural extension emphasizes the importance of feature aggregation across multiple levels to merge information from different stages of input processing (Fig. ??). Experimentally, this technique shows significant improvements in both memory usage and performance over frequently employed baselines such as ResNet [?] or Densely Connected Convolutional Network (DenseNet) [?]. In contrast with the skip connections, the DLA introduces more depth and sharing. There are two main different approaches to DLA, namely Iterative Deep Aggregation (IDA) and Hierarchical Deep Aggregation (HDA). The two mentioned approaches above are independent as well as compatible enough to facilitate combining the two for even richer feature aggregation. A great advantage is that these structures are expressed through an architectural framework, which is, more importantly, independent of the choice of backbone, thus preserving the compatibility with current and future networks.

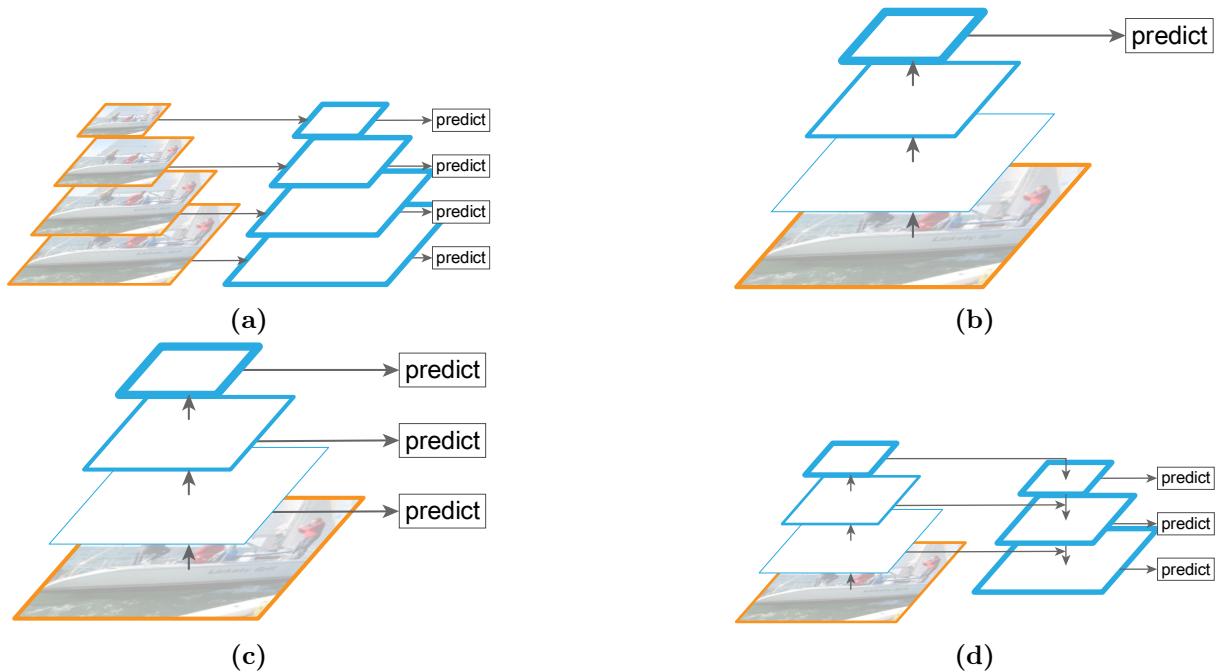


Fig. 3.23: A comparison of four traditional approaches to feature aggregation. (a) Computing features on distinct image scales (computationally expensive); (b) the use of single scale features only (fast, but not robust); (c) Reusing pyramidal feature hierarchy (fast and robust); (d) the proposed FPN - pyramidal feature aggregation in both directions (practically fast as previous methods but considerably more accurate). (*source: [?]*)

Iterative Deep Aggregation

IDA aims at resolution and scale fusion. The process starts at the smallest scale and then iteratively merges larger (deeper) scales, which can be described as

$$I(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \dots, \mathbf{x}_n) = \begin{cases} \mathbf{x}_1 & \text{if } n = 1 \\ I(A(\mathbf{x}_1, \mathbf{x}_2), \mathbf{x}_3, \dots, \mathbf{x}_n) & \text{otherwise} \end{cases}, \quad (3.26)$$

where A is the aggregation node.

Hierarchical Deep Aggregation

This process of aggregation exploits a tree-like structure that combines layers spanning multiple levels of a feature hierarchy. The HDA with aggregation function T_n with n representing the depth can be formulated as

$$T_n(\mathbf{x}) = A(R_{n-1}^n(\mathbf{x}), R_{n-2}^n(\mathbf{x}), \dots, R_1^n(\mathbf{x}), L_1^n(\mathbf{x}), L_2^n(\mathbf{x})), \quad (3.27)$$

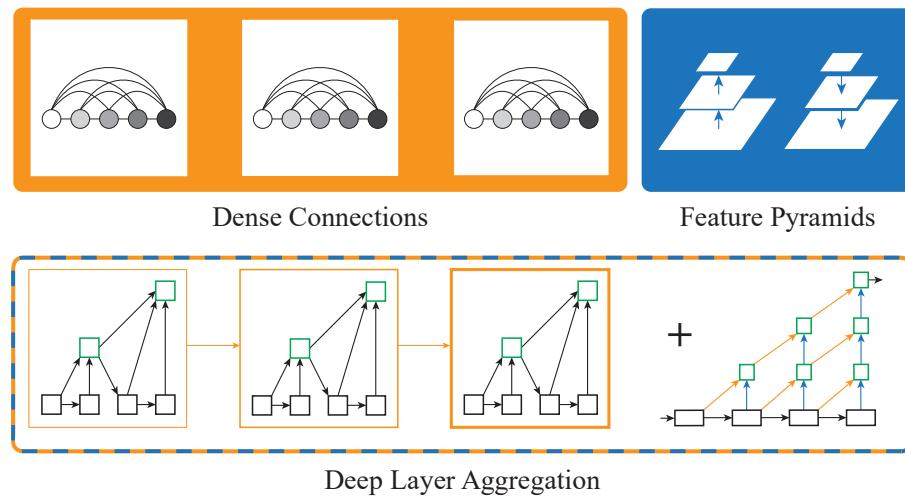


Fig. 3.24: A demonstration of unification of semantic and spatial information. The DLA architecture extends densely connected networks, *i.e.*, DenseNets, and FPNs. This extension builds on the idea of skip connections for enhanced feature fusion. (*source: [?]*)

where A is the aggregation node. The functions R and L are defined as

$$L_1^n(\mathbf{x}) = B(R_1^n(\mathbf{x})), \quad L_2^n(\mathbf{x}) = B(L_1^n(\mathbf{x})) \quad (3.28)$$

and

$$R_m^n(\mathbf{x}) = \begin{cases} T_m(\mathbf{x}) & \text{if } m = n - 1 \\ T_m(R_{m+1}^n(\mathbf{x})) & \text{otherwise} \end{cases}, \quad (3.29)$$

where B represents some convolutional block.

Chapter 4

Overview of Relevant Datasets

4.1 Object Detection Datasets

4.1.1 MS-COCO

The MS-COCO dataset [?] (web: [?]) was created for the purpose of object segmentation. However, if a model solves a much more complicated problem of object segmentation, pure object detection is then just one of many steps. To this end, this dataset is often adopted for training object detectors. It is a widely used dataset for image classification, object detection, and semantic segmentation. It is considered a benchmark dataset in different academic and industrial research areas. The images in the dataset are everyday objects captured from everyday scenes. This adds some “context” to the objects captured in the scenes (Fig. ??).

4.1.2 KITTI Object Detection

The KITTI Object Detection dataset [?] (web: [?]) can be adopted for training models for object detection and object orientation estimation. This benchmark holds 7 481 training images and 7 518 test images, comprising a total of 80 256 labeled objects. This work is of significant value for our research because of a firm ground for building an object detector aimed at traffic applications. Classes of objects are "car", "van", "truck", "pedestrian", "person sitting", "cyclist", "tram", "miscellaneous" or "don't care". Besides the rich content of traffic-related objects, additional information such as object and camera real-world positions are available, too. As we have mentioned the goal of dealing with object



Fig. 4.1: The MS-COCO dataset provides 80 classes (81 if background is taken into account) of common, everyday objects, which brings an additional contextual information. (*source: [?]*)



Fig. 4.2: The KITTI Object Detection dataset provides traffic-related, unconstrained scenarios with full details about the scene: object BBOXes, camera and object positions as well as indicators of occlusion level. (*source: [?]*)

occlusion, this dataset provides an occlusion tag that represents the level of occlusion, from full visibility to large occlusion (Fig. ??).

4.2 Object Re-identification Datasets

4.2.1 PKU VehicleID

The VehicleID dataset [?] (web: [?]) contains data captured during daytime by multiple real-world surveillance cameras distributed in a small city. There are 26 267 vehicles (221 763 images in total). Each image is attached with an ID label corresponding to its identity in the real world (Fig. ??). In addition, there are manually labeled 10 319 vehicles (90 196 images in total) of their vehicle model information (*i.e.* “MINI-cooper”, “Audi A6L”, “BWM 1 Series”, etc.). This dataset is usable thanks to the information about the model. We posit that at some point a hypothesis could be tested whether incorporating car model information into the machine learning model would improve its robustness. Nevertheless, only the front or rear view of the vehicle is available. This disadvantage is lessened by a reasonable number of training images. This dataset is also available only upon request and only for research purposes.



Fig. 4.3: Few samples from the VehicleID dataset. Each vehicle has at least two images in the dataset, but only its front and rear view were obtained. (*source: [?]*)

4.2.2 VeRI-776

A large-scale benchmark dataset named VeRI-776 (web: [?]) for vehicle ReID in the real-world urban surveillance scenario [?] (Fig. ??). In our opinion, this dataset is one of the best available, and it already has been explored and served the purpose of training ReID models. The featured properties of this include the following important properties for training robust ReID models:

- It contains over 50 000 images of 776 vehicles captured by 20 cameras covering an 1 km^2 area in 24 hours.
- The images were captured in a real-world unconstrained surveillance scene and labeled with varied attributes, *e.g.*, BBOXes, types, colors, and brands.
- Each vehicle is captured by at least 2 up to 18 cameras in different viewpoints, illuminations, resolutions, and occlusions.
- Data samples are also labeled with license plates and other spatio-temporal information, such as the BBOXes of plates with corresponding strings, the timestamps of vehicles, and the distances between neighboring cameras.

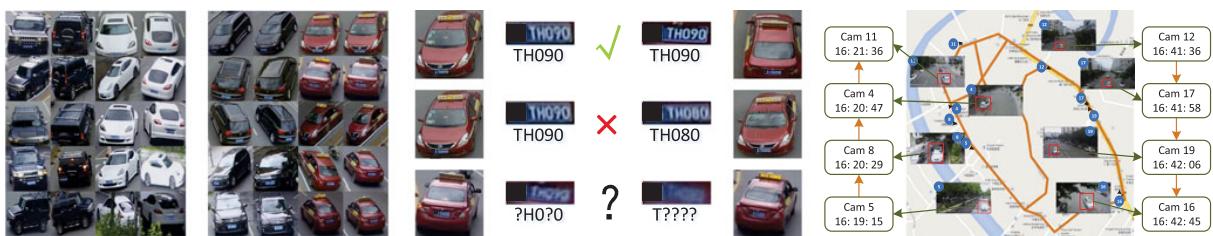


Fig. 4.4: The properties of the VeRI-776 dataset. Individual vehicles offer rich within-class differences in distinct viewpoints. At the same time, different but similar vehicles may have trivial inter-class differences. (*source: [?]*)

4.3 Visual Object Tracking Datasets

4.3.1 OTB 2013 and 2015

The Object Tracking Benchmark (OTB) dataset [?] is widely adopted for performance evaluation of a SOT algorithms. Each sequence is annotated frame-by-frame with BBOXes and 11 different attributes that are used for various challenge evaluations. The objective is to evaluate general object tracking, thus this dataset can be often encountered in SOT, especially in older works. Each sequence represents a single object to be tracked. The initial version, OTB-2013 contains 51 sequences whereas the upgraded version, OTB-2015 provides exactly 100 sequences, including the ones from the previous version.

4.3.2 GOT10k

One of the newest benchmarks is the GOT-10k dataset [?] (web: [?]). This dataset belongs to the top ones freely available in terms of quality. The number of video segments this dataset contains exceeds 10 000. These videos cover real-world moving objects represented by 1.5 million manually labeled BBOXes. Besides a plethora of real-world classes the number of which surpasses 560, there are also more than 80 classes of motion patterns present throughout the dataset (Fig. ??). As we have already mentioned, the principle of one-shot learning has been advertised in the tracking community several times. This benchmark encourages the development of generic purposed trackers by following the one-shot rule. It is also worth emphasizing the zero overlap between the train and test sets in terms of object classes. The comprehensiveness of this work is further reinforced by providing extra labeling information, such as object visibility ratios.

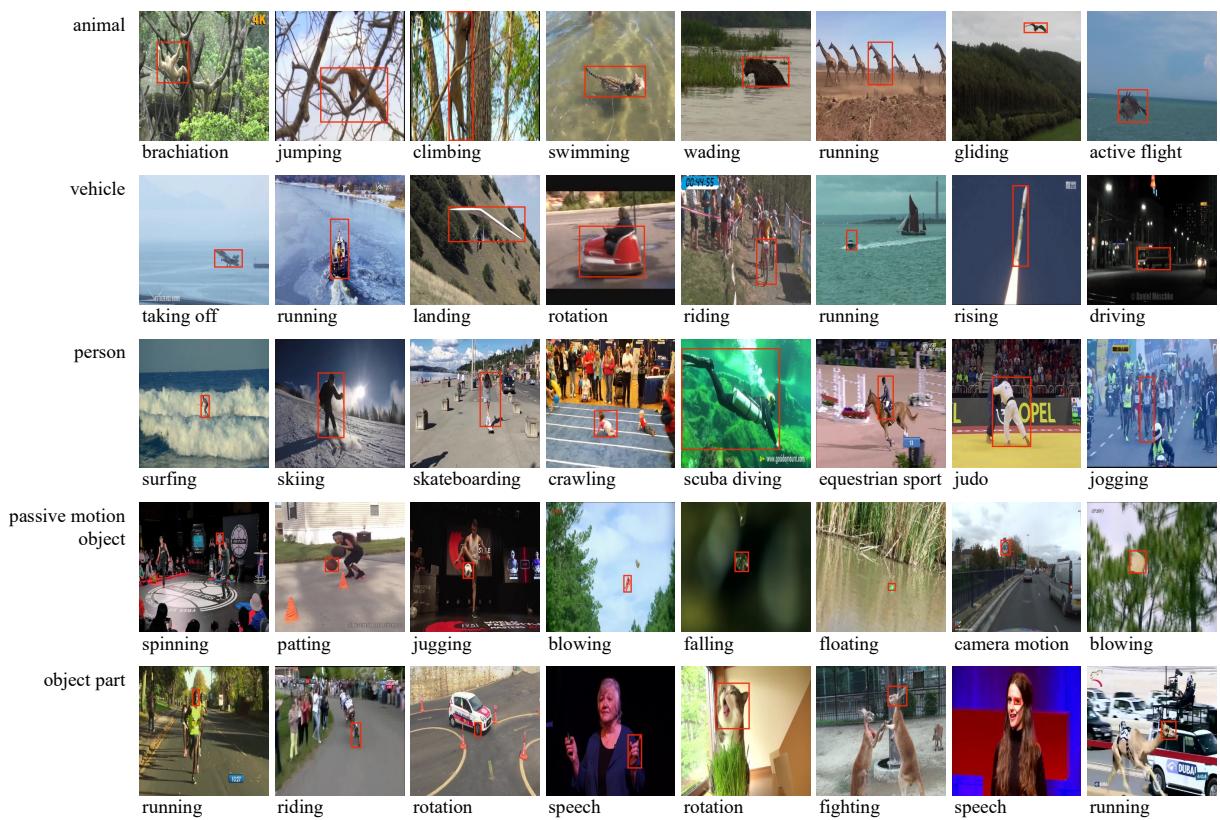


Fig. 4.5: Representative videos collected and annotated in the GOT-10k dataset. (*source: [?]*)

4.3.3 VOT 2015-2019

This prominent dataset (web: [?]) is a benchmark in VOT [?]. It has a long history of development and annual challenges for the best visual tracker. All VOT datasets are available through the VOT toolkit. The pipeline for evaluating a tracker is automated to facilitate ease of use and a framework for researchers to allow an objective comparison with others. The dataset provides multiple video sequences where a single object is present under various conditions. For example, people running, a fish swimming behind corals, a vehicle driving in a city, and so forth. Each object is annotated by a single BBOX per each frame in which it appears. There are various versions of this benchmark, with incremental updates every year to existing challenges or with the introduction of new challenges altogether.

4.3.4 KITTI Object Tracking

This object tracking benchmark [?] (web: [?]) consists of 21 training sequences and 29 test sequences. Even though there have been labeled 8 different classes, only the classes



Fig. 4.6: A sample of the only two classes ("car" and "pedestrian") that are evaluated in the benchmark using the KITTI Object Tracking dataset. (*source: [?]*)

"car" and "Pedestrian" are evaluated in this benchmark, as only for those classes enough instances for a comprehensive evaluation have been labeled. Considering our potential traffic application, this fact does not represent a disadvantage. The goal of the object tracking task in this benchmark is to estimate object tracklets for the classes "car" and "pedestrian". Only 2D 0-based, axis-aligned BBOXes in each image are evaluated.

4.3.5 UA-DETRAC

The most important benchmark dataset for our work is UA-DETRAC [?] (web: [?]). To the best of our knowledge, this dataset most favorably suits the needs of all surveyed datasets available. The primary reason is that it provides a plethora of traffic situations recorded using a static camera (Fig. ??). This setup appropriately reflects the requirements of our goal, which is the analysis of traffic scenes using object tracking algorithms. This work provides high-quality human-generated annotations with a lot of additional information about the captured vehicles, such as the intensity of their occlusion.

UA-DETRAC is considered a challenging real-world multi-object detection and multi-object tracking benchmark. The dataset consists of 10 hours of videos captured at 24 different locations in China. The videos are recorded at 25 FPS, with resolution of 960×540 pixels. There are more than 140 000 frames and 8 250 vehicles that are manually annotated, leading to a total of 1.21 million labeled BBOXes of objects.

Since this dataset is of paramount importance to our research, here we provide more details about the structure and properties of the contained data compared to other datasets described in our work. The dataset consists of 100 videos, where 60 of them are dedicated to training, while the remaining 40 are used for testing. Ground-truth annotations are provided in both variations. This is not always the case, as several benchmarks do not disclose annotations for the test dataset, *e.g.*, KITTI [?].

The dataset authors provide extensive information about the vehicle, including its



Fig. 4.7: A sample from the UA-DETRAC dataset. The whole dataset consists of diverse traffic situations captured using a static camera viewed from various angles. (*source: [?]*)

Min.	Max.	Mean	Stdev.	Median
1	49	9.21	6.60	21

Table 4.1: Basic statistics for the per-frame vehicle density in the UA-DETRAC dataset.

speed in frames per second, color, orientation, and occlusion. Due to space restrictions, we limit our elaboration on how the data was obtained only to the fields pertinent to our usage. In the beginning, there is a section describing some ignored regions (Fig. ??). The authors decided to omit regions with very dense traffic. Nevertheless, the dataset contains a plethora of scenes where the number of cars is very high. More specifically, Table ?? describes some basic statistical properties of the distribution of the number of cars throughout the dataset. The data were obtained by collecting the number of annotated cars for each frame. Training and testing data were merged for simplicity.

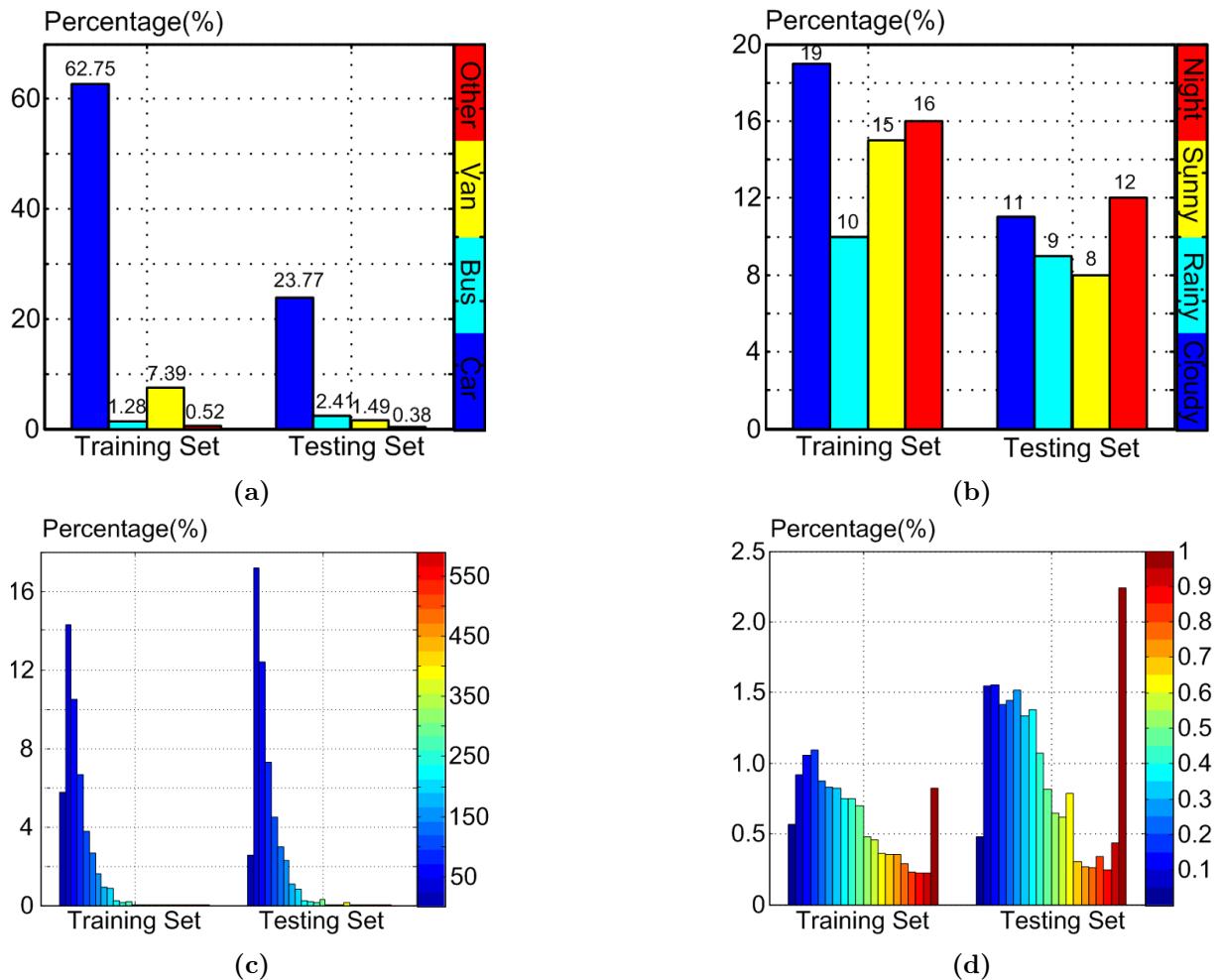


Fig. 4.8: Summary statistics of the UA-DETRAC dataset. **(a)** shows the distribution of vehicle categories, one of *car*, *bus*, *van* or *other*; **(b)** shows the varying weather conditions belonging to either *night*, *sunny*, *rainy* or *cloudy*; **(c)** depicts the change in scale given by the square root of the BBOX pixel area; and **(d)** reflects the occlusion ratio throughout the dataset computed as the fraction of the vehicle BBOX being occluded . (*source: [?]*)

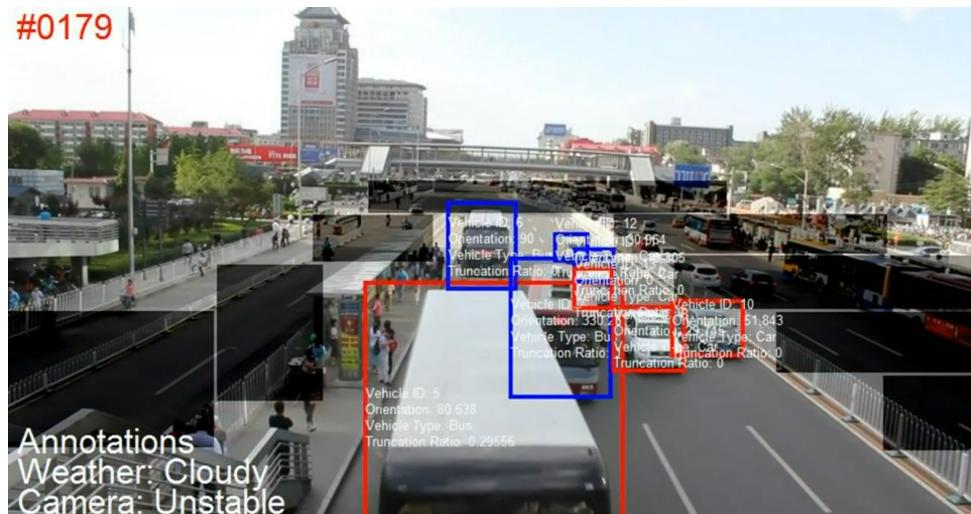


Fig. 4.9: A demonstration of a possible distribution of ignored regions in the UA-DETRAC dataset. (*source: [?]*)

Chapter 5

Developed Homography Ranking Method

This chapter is dedicated to one of our experiments that were not completely related to the VOT itself, yet we achieved an original scientific contribution that could potentially be applied to object tracking, especially traffic analysis. Even though we did not continue with the homography-based object tracking due to limitations of available datasets, still we would like to elaborate on our developed approach. The proposed method was fully described as well as scrupulously tested under difficult conditions. The write-up of the whole research process was eventually published in a journal [?].

5.1 Introduction

Computer vision often deals with diverse image transformations to improve the outcome of the subsequent post-processing phase. The perspective transformation was deemed of particular interest to our goal of traffic analysis. Specifically, removal of perspective distortion. To this end, the so-called homography mapping is often exploited.

Broadly speaking, homography is a perspective projection of a plane from one camera view into a different camera view. The perspective projection maps points from a 3D world onto a 2D image plane along lines that emanate from a single point [?, ?]. Such a projection is contained within a 3×3 invertible transformation matrix called the homography matrix (or just homography) with 8 Degrees of Freedom (DoF). A general

homography matrix can be defined as follows

$$\mathbf{H} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \quad (5.1)$$

This transformation may facilitate mapping between two views of the same plane. Concretely, a single vector $\mathbf{u}^T = [u_x, u_y, 1]$, which represents a warped keypoint in homogeneous coordinates, is mapped onto the rectified keypoint $\tilde{\mathbf{u}}^T = [\tilde{u}_x, \tilde{u}_y, 1]$, by the homography \mathbf{H} using the transformation $s\tilde{\mathbf{u}} \approx \mathbf{H}\mathbf{u}$, where s represents the scale factor. In our case, the goal was to rectify the image so that it looks as if the camera was in an orthogonal position with respect to the desired plane in the world.

Homography is frequently adopted for text document rectification to generate a fronto-parallel view [?, ?], image stitching [?, ?], extracting metric information from 2D images [?], pose estimation [?], and for various traffic-related applications, *e.g.*, ground-plane detection [?], and bird's-eye view projection [?].

We aimed at exploring the possibility of employing homography for VOT. The primary incentive was the fact that as long as a static camera is used and a few assumptions that we will discuss later hold, the scene may be easily stripped off the effect of the perspective distortion. Consequently, the use case of tracking vehicles visually using a static camera while exploiting a fronto-parallel view over the road seemed like a plausible extension with possible advantages for traffic analysis. Furthermore, the combination of homography and object tracking is present in the literature, *e.g.* [?, ?, ?]. Bose *et al.* [?] presented a fully automated technique for both affine and metric rectification of a given ground plane by simply tracking moving objects. The derivation of the necessary constraints for projective transformation between the image and the ground plane was obtained by observing objects that moved at constant velocity in the world for some part of their trajectory. We conjectured that the extra information about the scene geometry that we may achieve using rectification could aid in making the tracking more accurate. Visual trackers are often supported by motion models such as Kalman filter [?], so the rationale was to estimate the motion model in an orthogonal projection, rather than a perspectively distorted one.

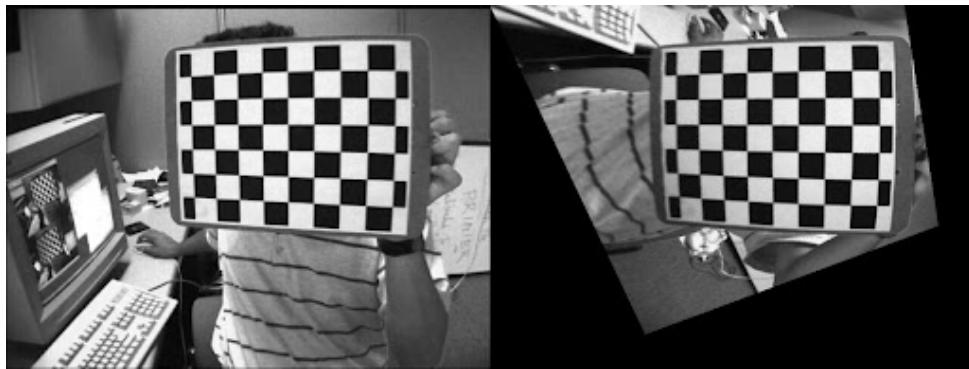


Fig. 5.1: An example of a chessboard marker present in a scene that may be used to establish a point correspondence that would serve for a homography transformation. The rectified, fronto-parallel view demonstrates the desired effect that points present on the “ground” plane are properly projected, whereas other points suffer from substantial distortion. (*source: [?]*)



Fig. 5.2: An example of a virtual square marker present on a road that may be used to establish a point correspondence, and thus the homography transformation, too. The obtained view allows for many applications such as speed and size measurements that would otherwise be a lot more problematic in a perspective-deformed view. (*source: [?]*)

A common approach to estimate the homography is to use a set of at least four 2D point correspondences [?]. The points that are used for establishing the 2D point correspondences will be referred to as keypoints. These keypoints may belong to a marker which is an object with a known shape that is either naturally occurring or artificially positioned in the scene. A regular, easy-to-detect pattern (*e.g.*, a chessboard) is commonly utilized [?] (Fig. ??). A single marker is identified in the image by multiple independent keypoints that have a direct correspondence to its real shape, thus making a group of point correspondences. For the sake of traffic analysis, the marker may be represented by virtually any points on the image as long as certain conditions are met (Fig. ??). However, the point correspondences established this way are often subjected to noise, thus errors may be introduced in the homography estimation. Although 4 keypoints are satisfactory, often a greater number of keypoints is used, allowing to use optimization to minimize a

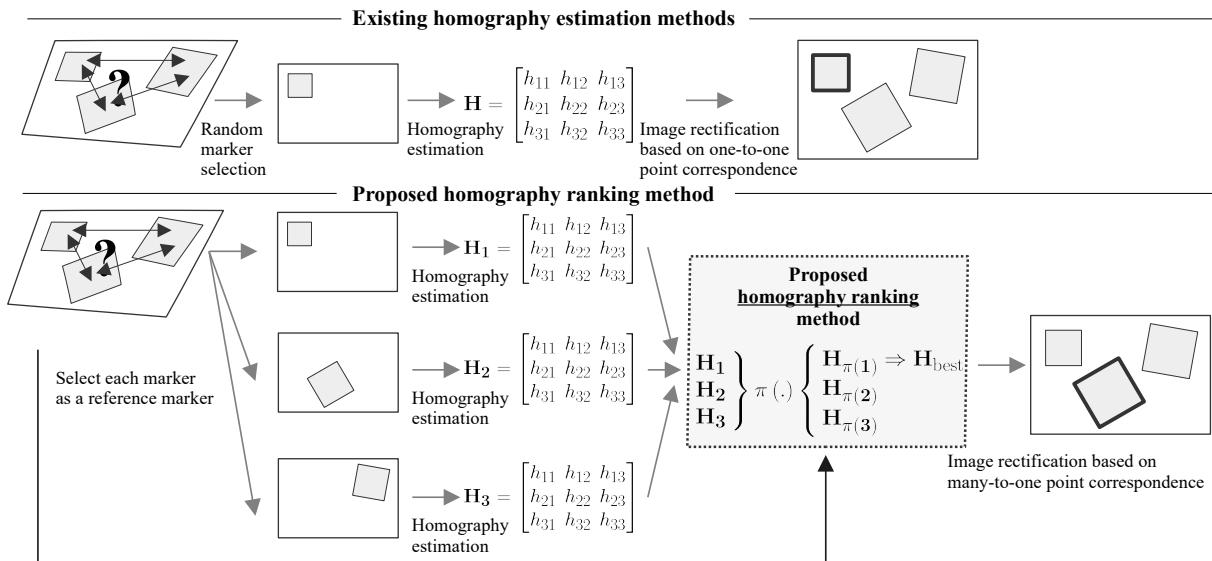


Fig. 5.3: A fundamental difference between existing homography estimation methods and our proposed method for homography ranking. If there are multiple markers while the information about their relative positions in the world is absent, the existing approaches can only estimate isolated homographies without the ability to select the best one. To address this issue, our method easily serves as an extension to existing approaches by exploiting multiple markers to rank the isolated homographies from the “best” to the “worst”.

suitable cost function [?, ?]. Subsequently, an outlier removal becomes an important step in the processing pipeline, for which effective and robust algorithms such as RANSAC [?] are usually employed.

A real-world application of generating a bird’s-eye view over a road (Fig. ??) from a video recording when we could not use a large marker to cover a sufficient portion of the road (Fig. ??) motivated this entire project. We observed that, under our conditions, the homography estimation based on a single small marker was inaccurate. Therefore, there was an attempt to utilize multiple small markers and measure their relative positions. However, as is often the case in practice, their position measurements were highly noisy at best. Thus, we had to bypass the position measurements altogether, which led us to adopt the proposed method, instead. It is crucial to emphasize that our method can also be adopted in a situation when the marker placed at various positions on the same planar surface can be seen at different frames using a static camera. Stacking the captured frames onto each other would effectively yield an artificially generated view of multiple markers.

Assume a presence of a sole marker in the scene (Fig. ??). Moreover, assume the view of the marker is perspectively distorted. If we know its real shape makes at our disposal, then it is possible to compute the homography. However, when multiple copies of the same marker are visible, but their positions in the world are unknown, the detailed information

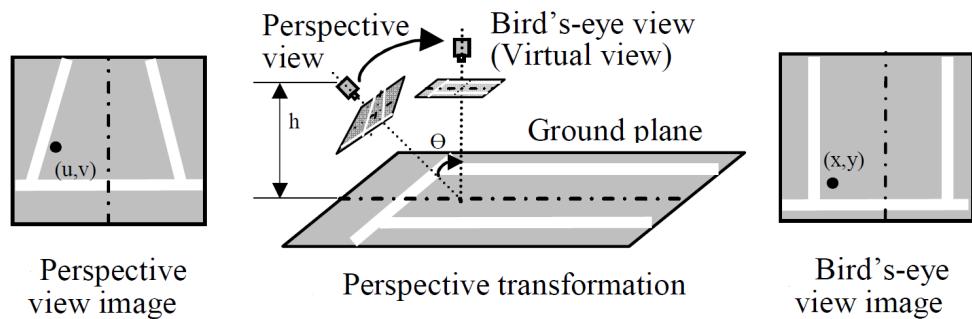


Fig. 5.4: A demonstration of the process of rectification when obtaining a fronto-parallel view over the road using homography. (source: [?])

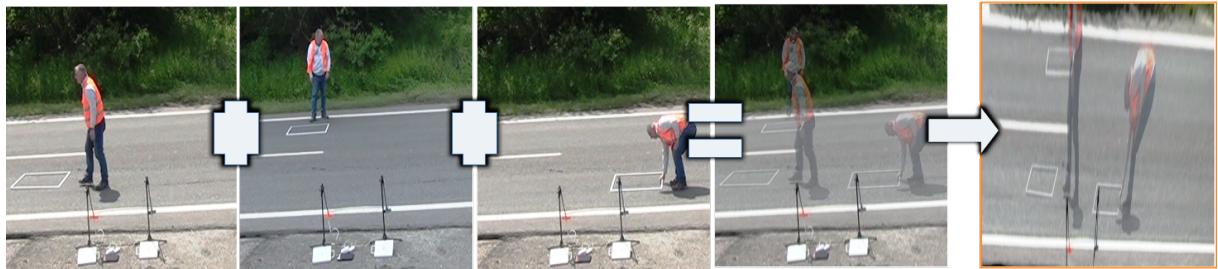


Fig. 5.5: A motivating real-world example. We can see different frames captured during a video recording that show various positions of the same marker. The picture after the “equality” sign is a merge of the previous frames for better illustration. Due to the use of a static camera, we may treat the positions of the given marker on individual frames as if they were captured simultaneously. However, the question remains unanswered. Given multiple markers in the absence of their position information, which one is the best to choose for rectification?

about the shape is not enough to incorporate all the keypoints in the estimation. In the absence of position information, existing approaches for homography estimation based on point correspondences do not work because the projection has to preserve the proportional positions. As a result, estimating the homography while not knowing the ground-truth layout of the keypoints up to an arbitrary scale does not guarantee, and often does not even lead, to the correct result.

Under the constraints discussed above, the existing methods can only generate an isolated homography for each marker based on the one-to-one point correspondence (Fig. ??). Each homography may be affected by different sources of noise, *e.g.*, low resolution, blur, or keypoint detection. Thus, the outcome of rectification may vary up to a great extent. In addition, many practical applications often use a marker that just covers a small portion of the image, increasing susceptibility to noise as a result. The trivial solution would be to use a bigger marker that covers the majority of the estimated plane’s area. But such a solution is often cumbersome. It is simply not possible to “merge” multiple isolated homographies together.

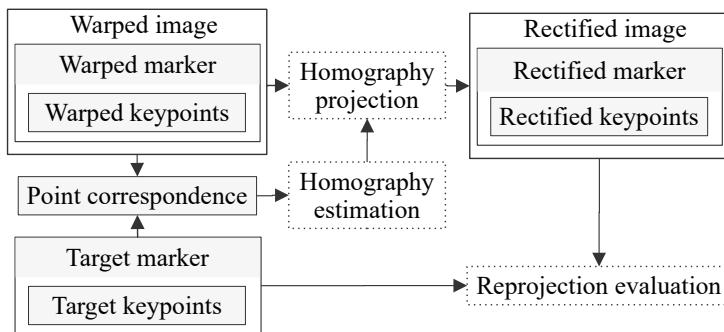


Fig. 5.6: Visualization of relationships within our established terminology. This diagram also depicts the hierarchical dependence between individual terms. In addition, the dotted elements represent processes with arrows denoting their input and output.

5.2 Preliminaries

We define a marker as an object with a known, easy-to-detect shape. Such an object can be either naturally occurring or artificially placed on the planar surface of the scene we want to remove perspective distortion from. The marker contains keypoints, which is a set of distinct, independent, visual feature points (for instance, corners). The chosen keypoints visible in the perspectively deformed image are called the warped keypoints. The set of the rectified keypoints is represented in the desired image (not subjected to perspective distortion) and is produced from the warped keypoints using the homography projection. Last but not least, the point correspondence is a relationship between the warped and the target keypoints and it is necessary for homography estimation. In an ideal case, the rectified keypoints match the target keypoints in terms of their pixel positions (Fig. ??).

Unless stated otherwise, a **similarity transformation** denotes a limited affine transformation with 4 DoF which encompasses translation, rotation and uniform scaling (Equation ??). Specifically, let \mathcal{K}_1 and \mathcal{K}_2 be sets of feature keypoints belonging to objects O_1 and O_2 . We refer to the objects O_1 and O_2 as **similar** if there exists a similarity transformation ψ , such that $\mathcal{K}_1 = \psi(\mathcal{K}_2)$ and $\mathcal{K}_2 = \psi^{-1}(\mathcal{K}_1)$. For instance, O_1 and O_2 may represent rectangles of different sizes whilst having a equal aspect ratio.

Let m denote the number of markers and k represent the number of keypoints belonging to each marker in consideration. We describe each i -th marker using a $3 \times k$ matrix $\mathbf{W}^{(i)}$

that stores the warped keypoints as

$$\mathbf{W}^{(i)} = \begin{bmatrix} x_1^{(i)} & x_2^{(i)} & \dots & x_k^{(i)} \\ y_1^{(i)} & y_2^{(i)} & \dots & y_k^{(i)} \\ 1 & 1 & \dots & 1 \end{bmatrix}, i = 1, \dots, m. \quad (5.2)$$

Analogically, we describe the target keypoints using a $3 \times k$ matrix \mathbf{T} . Owing to the many-to-one point correspondence, only one specification is sufficient. Just beware that the ordering of keypoints had to match the warped keypoints defined above, so

$$\mathbf{T} = \begin{bmatrix} \tilde{x}_1 & \tilde{x}_2 & \dots & \tilde{x}_k \\ \tilde{y}_1 & \tilde{y}_2 & \dots & \tilde{y}_k \\ 1 & 1 & \dots & 1 \end{bmatrix}, \quad (5.3)$$

with the point correspondence relationship formulated as

$$x_j^{(i)} \simeq \tilde{x}_j, y_j^{(i)} \simeq \tilde{y}_j, i = 1, \dots, m, j = 1, \dots, k. \quad (5.4)$$

5.3 Developed Method

Our work aimed to devise a systematic approach to select the “best” homography according to the proposed score function. The assumption was that there was no prior knowledge about the quality of individual markers.

Here is the description of the proposed method. Each homography is induced by a single independent marker. The input to our method is multiple sets (*i.e.*, groups) of point correspondences between the warped and the ground-truth (ideal) markers. Therefore, each marker is represented by a unique set of keypoints. The use case of our method is to rank multiple homographies and select the best performing one with respect to the tailor-made score function. Consequently, we require a homography matrix for each marker (a

set of point correspondences) on the input. The great advantage comes from the fact that to compute these matrices, any state-of-the-art method can be utilized as a black box. The benefit is that it is capable of ranking the referred homographies without the knowledge of absolute or relative positions of markers in the world (Fig. ??). However, we have to emphasize that we did not propose any method to simultaneously estimate multiple homographies. We only build upon the existing homography matrices.

Due to our assumption of not knowing the arrangement of markers in the scene, there is no way to create one virtual, compound marker that contains all the keypoints. If we could, then we would employ RANSAC [?] or any other sophisticated algorithm to select the best subset of keypoints to estimate the homography. In that scenario, our approach would be useless. We only have information about the relative position of the marker’s keypoints at our disposal, not the markers themselves. As a result, the point correspondence is globally indeterminate. We can only establish a local point correspondence between a single marker and its corresponding ground-truth shape. For the best performance, to obtain the isolated homographies, we suggest the user chooses the most robust method available.

The homography estimation between existing point correspondences is a standard problem we heavily rely on. As already highlighted, we did not contribute to this problem in terms of improving the homography estimation itself. We only provided a way to rank the resulting homographies. We developed a way to, under certain circumstances, choose the “best” homography from multiple existing ones. Therefore, our method could not even be compared to RANSAC, because we tackle a different problem. There are three following assumptions the proposed method is based upon:

1. The markers are geometrically similar, which means that they are allowed to differ only in translation, rotation, and uniform scale in the real world.
2. The shape of at least one of the used markers is known beforehand.
3. These markers are positioned on the same planar surface visible in the scene.

One important caveat is that our method handles only transformation from a distorted to the undistorted view of the target plane.

We exploited the properties of homography and similarity transformations and expressed them in a single score function, which stands at the core of our contribution.

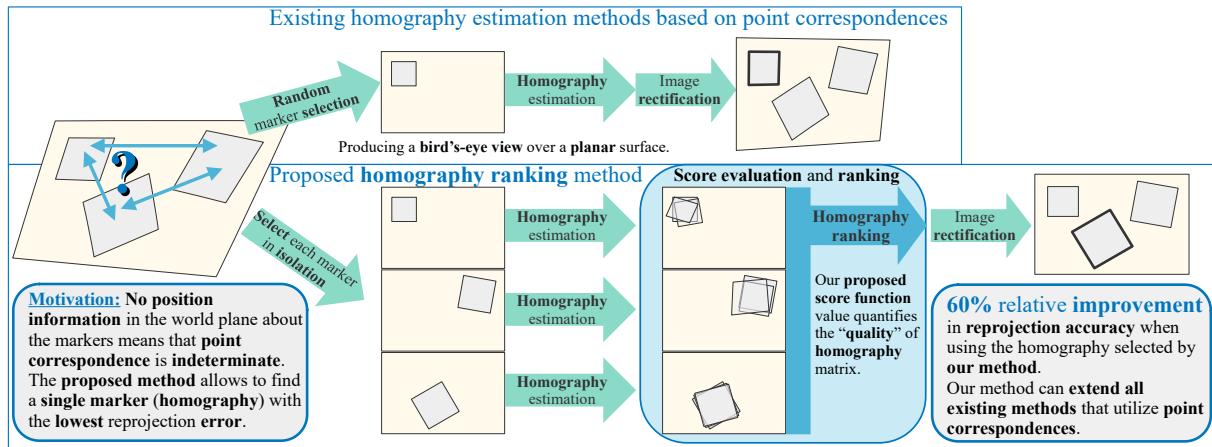


Fig. 5.7: The graphical abstract from our paper. The basic idea is that existing approaches may only estimate an isolated homography for each marker and cannot determine which homography achieves the best reprojection over the entire image. Therefore, we proposed a method to rank isolated homographies obtained from multiple distinct markers to select the best homography. This method extends existing approaches, provided that the point correspondences are available and the markers differ only by similarity transformation after rectification.

The score function value is exploited as a proxy for homography ranking according to their reprojection error over the entire image using only markers' keypoints. It is only an estimate. The usual use case would be to select the homography with the lowest score, *i.e.*, the highest-ranked matrix, to perform the image rectification with the expectation of obtaining the most accurate reprojection.

Our method utilizes multiple similar markers (Fig. ??). The input is point correspondences and homographies estimated for each marker. Each marker becomes the reference marker only once during the course of the algorithm. All the remaining markers serve as auxiliary markers. The reference marker's homography is used to perform the perspective transformation to rectify all the visible markers. To rank which reference markers' homography yields the best reprojection, we exploit auxiliary markers. Auxiliary markers are subsequently mapped onto the target marker using similarity transformations (Equation ??). The transformed keypoints are converted to homogeneous coordinates and the reprojection error is measured as the mean Euclidean distance between the rectified and the target keypoints (Equation ??). The objective is to minimize the computed quantity.

Let r be the index of the reference marker. The 3×3 matrices describing similarity

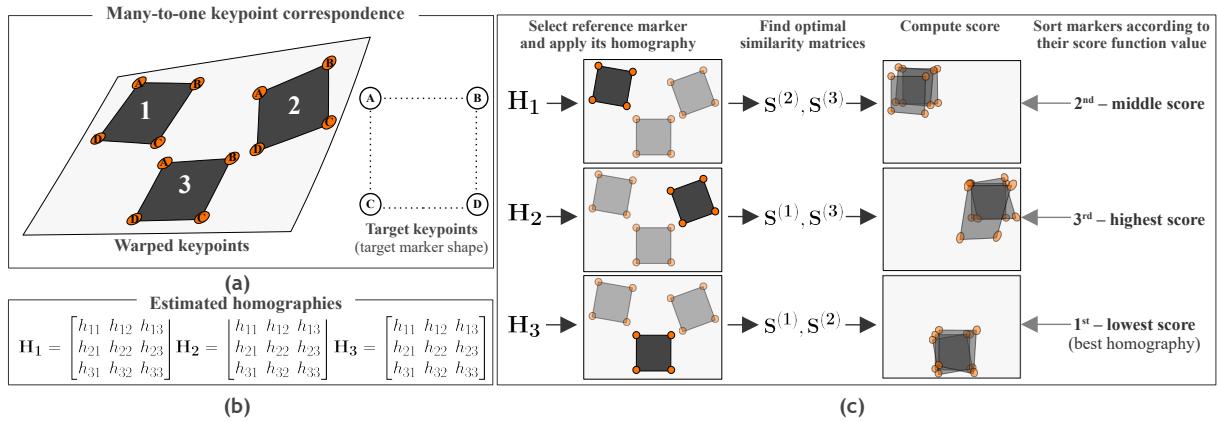


Fig. 5.8: A system diagram of our method. (a) The input consists of a many-to-one point correspondence specified by multiple similar markers together with the information about the ground-truth shape (up to an arbitrary positive scale) of the target marker. (b) The assumption is that the isolated homographies related to each marker are ready on the input as well. (c) The algorithm processes each marker by applying its corresponding homography matrix to the image to produce a rectified image. Subsequently, it computes optimal similarity matrices using auxiliary markers. These transformations are required for the computation of the score function. The obtained score values then serve for comparison when ranking the homographies. The homography that ends up ranked first is considered (predicted) to be the “best” candidate for achieving the minimal reprojection error over the whole image.

transformations are contained in a set $\mathcal{S} = \{\mathbf{S}^{(i)} \mid i = 1, \dots, m\}$, such that

$$\mathbf{S}^{(i)} = \begin{cases} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} & \text{if } i = r \\ \begin{bmatrix} \mathbf{R}_{2 \times 2}^{(i)} & \mathbf{T}_{2 \times 1}^{(i)} \\ \mathbf{0}_{1 \times 2} & 1 \end{bmatrix} & \text{if } i \neq r \end{cases}, \quad (5.5)$$

for $i = 1, \dots, m$, where

$$\mathbf{R}_{2 \times 2}^{(i)} = \begin{bmatrix} s^{(i)} \cdot \cos(\theta^{(i)}) & -s^{(i)} \cdot \sin(\theta^{(i)}) \\ s^{(i)} \cdot \sin(\theta^{(i)}) & s^{(i)} \cdot \cos(\theta^{(i)}) \end{bmatrix}, \quad \mathbf{T}_{2 \times 1}^{(i)} = \begin{bmatrix} t_x^{(i)} \\ t_y^{(i)} \end{bmatrix}. \quad (5.6)$$

This transformation (besides the identity) involves 4 DoF: a single rotation angle $\theta^{(i)}$, two x and y translation coefficients $t_x^{(i)}$, $t_y^{(i)}$, and a scale coefficient $s^{(i)}$. A full affine transformation (6 DoF) would incorporate horizontal and vertical scales, shear and rotation, and x , y offsets [?]. The application of homography that rectifies an image generates a

frontal plane that is related to the ground-truth plane by a similarity transformation [?, ?]. Thus, we do not include the shear and we only support uniform scaling. The mathematical justification can be found in the appendix section of our paper [?].

As all the markers share the same planar surface, a valid homography corresponding to any of them by definition to provide a valid perspective projection. However, all perspective projections are subjected to different noise. The endeavor then is to quantify which homography estimation could provide the best perspective projection for the whole plane in the image. To do so, we propose a score function based on the aforementioned constraints. The score function computes a score for individual homographies in along with the estimated similarity matrices corresponding to auxiliary markers as

$$\mathcal{F}(\mathbf{H}, \mathcal{S}) = \frac{1}{m} \sum_{i=1}^m \|h(\mathbf{S}^{(i)} \mathbf{H} \mathbf{W}^{(i)}) - \mathbf{T}\|_F, \quad (5.7)$$

where $\|\cdot\|_F$ denotes the Frobenius norm. The function $h(\cdot)$ converts points to homogeneous coordinates as

$$h \left(\begin{bmatrix} x_1 & x_2 & \dots & x_k \\ y_1 & y_2 & \dots & y_k \\ z_1 & z_2 & \dots & z_k \end{bmatrix} \right) = \begin{bmatrix} x_1/z_1 & x_2/z_2 & \dots & x_k/z_k \\ y_1/z_1 & y_2/z_2 & \dots & y_k/z_k \\ 1 & 1 & \dots & 1 \end{bmatrix}. \quad (5.8)$$

In what follows, we describe the proposed Algorithm ?? for homography ranking. Assume a set of warped markers described by the warped keypoints and a single target marker represented by the target keypoints. There is a many-to-one point correspondence linking these objects. Besides, assume that homographies have been estimated for each marker in isolation. Our algorithm ranks the input set of all pairs $(\mathbf{W}^{(i)}, \mathbf{T})$, $i = 1, \dots, m$ in ascending order by how well each i -th marker preserves the target shape of all the markers in the image after removing the perspective distortion. The score function defined in Equation ?? is used to measure this objective. The algorithm evaluates all markers as candidates for the reference marker. In each iteration, it computes optimal similarity matrices belonging to the auxiliary markers in the rectified plane, *i.e.*, after applying the perspective projection induced by the current homography. The aim is to find a homography with a minimal score. The algorithmic complexity is quadratic in the number

Algorithm 2 Homography ranking algorithm.

```

1:  $\bar{\mathbf{H}} \leftarrow \text{array}[m]$                                  $\triangleright$  empty array for the homography matrices
2:  $\mathbf{s} \leftarrow \text{array}[m]$                                  $\triangleright$  array of scores computed before the ranking (sorting)
3: for  $i \leftarrow 1, \dots, m$  do                                 $\triangleright$  for each reference marker
4:    $\bar{\mathbf{H}}[i] \leftarrow \text{HOMOGRAPHY}(\mathbf{W}^{(i)}, \mathbf{T})$      $\triangleright$  retrieve or estimate perspective transform.
5:    $\bar{\mathbf{S}}^{(i)} \leftarrow \mathbf{I}_{3 \times 3}$                        $\triangleright$  identity matrix to stand for a similarity transformation
6:    $\bar{\mathcal{S}} \leftarrow \{\bar{\mathbf{S}}^{(i)}\}$                            $\triangleright$  set of similarity matrices
7:   for all  $j : \{1, \dots, m\} - \{i\}$  do                 $\triangleright$  for each auxiliary marker
8:      $\bar{\mathbf{S}}^{(j)} \leftarrow \text{SIMILARITY}(\bar{\mathbf{H}}[i] \cdot \mathbf{W}^{(j)}, \mathbf{T})$      $\triangleright$  estimate similarity transformation
9:      $\bar{\mathcal{S}} \leftarrow \bar{\mathcal{S}} \cup \bar{\mathbf{S}}^{(j)}$                           $\triangleright$  store the similarity matrix
10:  end for
11:   $\mathbf{s}[i] \leftarrow \mathcal{F}(\bar{\mathbf{H}}[i], \bar{\mathcal{S}})$                    $\triangleright$  evaluate score function (Equation ??)
12: end for
13:  $\omega \leftarrow \text{ARGSORT}(\mathbf{s})$   $\triangleright$  indirect sort, only obtain indices of “would-be” sorted elements
14: return  $\bar{\mathbf{H}}, \omega$                                  $\triangleright$  return homographies and their respective ranking positions

```

of markers, thus $\Theta(m(m-1) + m\log_2(m)) \simeq \Theta(m^2)$. It is important to remark that the two functions used to compute the homography and similarity matrices in the pseudocode may stand for arbitrary methods that produce the required transformations.

5.4 Experiments and Discussion

The evaluation of the proposed homography ranking algorithm involved various conditions. We tested cases that included diverse similarity transformations applied to original markers as well as noisy point correspondence, *e.g.*, errors in marker detection since these are the expected problems in real-world scenarios.

Fig. ?? demonstrates how the reprojection error varies with respect to the marker position. It can be observed that the marker position can be approximately estimated by looking at the heatmap which represents the pixel-wise reprojection error over the image. However, the important property is that not all markers are subjected to the same pattern of error variation. This is the core observation that motivated our solution in the first place. The objective is to select the marker that minimizes the pixel-wise reprojection error within the region of the image that is as broad as possible. That is why we evaluate our method by computing the reprojection error over each pixel, not just the keypoints. The rationale is that subsequent image postprocessing would greatly benefit from having the area of the image as large as possible that is reprojected properly.

All the test scenarios indicated the following trend. On average, the homography

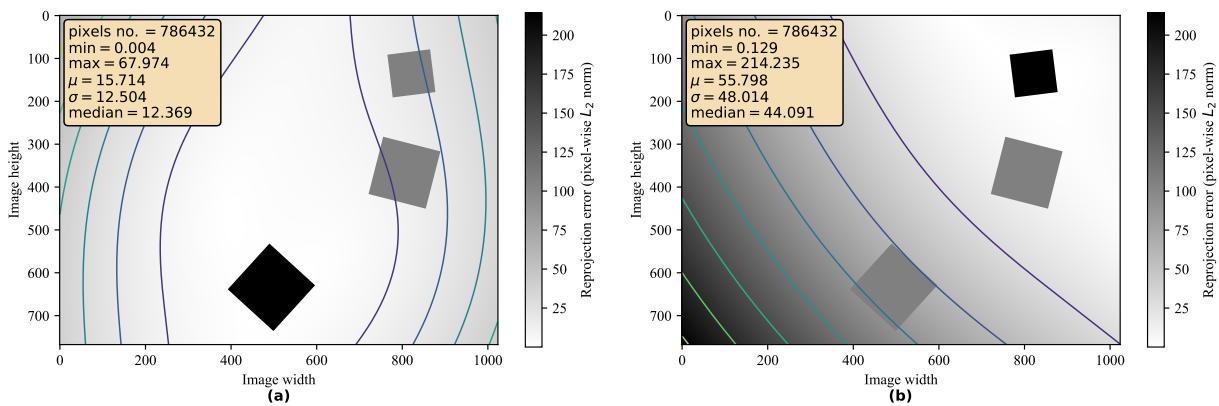


Fig. 5.9: Distribution of pixel-wise reprojection error. The heat map along with the corresponding contours demonstrate the varying distance between the ground truth and rectified pixel position after removing the perspective distortion. The bold square represents the reference marker. We show the result of (a) the “best” marker and (b) the “worst” marker. This test scenario includes all similarity transformations as well as noise in point correspondence.

with the highest score improved the relative performance to the baseline performance the most (both median and mean above 60%). The lowest-ranked homography often led to significantly worse performance (median and mean around -90%). These values varied moderately across different setups.

Implementation Details

Our proposed algorithm can be utilized to extend any homography estimation method that exploits point correspondences. To demonstrate, we adopted time-tested implementations from the OpenCV 4.4.0 library [?]. Each homography was estimated by the `findHomography()` function which internally employs DLT [?] algorithm for $k = 4$ and RANSAC [?] algorithm for $k > 4$, where k is the size of the point correspondences set. At the same time, each optimal similarity transformation between two 2D point sets was estimated by the `estimateAffinePartial2D()`, which also utilizes RANSAC for robustness. We used the default parameters whenever possible.

5.4.1 Dataset Creation

Our synthetic dataset was created to simulate the presence of markers in the scene subjected to perspective distortion to facilitate a pixel-wise comparison of the reprojection error. This dataset covered multiple setups named as **test scenarios**. For each test scenario, we generated t different samples which we call **test instances**. We set $t = 1\,000$. Table ?? contains description of the generated test scenarios. To create test instances

(within test scenarios), we employed the procedures described below (Fig. ??). Our dataset easily allows complete reproducibility of the reported results thanks to the synthetic nature of our data. The source code for running the experiments is freely available on our GitHub repository [?].

Image Initialization

Each test instance was initialized as a blank 1024×768 image. This image served for m randomly generated copies of the same shape (marker) placed in a 3×3 grid, where $0 < m \leq 9$. We used a uniform border with 20% size of the corresponding side to prevent the generated shapes from reaching outside of the image. We experimented with a different number of markers. From the set of 3×3 possible anchors, we chose m randomly onto which we placed the generated markers. We also studied the effect of 3, 5, 7, and 9 out of 9 possible markers, given that all the similarity transformations and noise were applied. Regarding marker shapes, we tested squares or convex, equilateral polygons, with a tight BBOX of size 100×100 pixels (covering approximately 1.3% of the image). However, other similar shapes could be used as well. Their centroids were evenly distributed over the image whereas the grid cells served as anchors. We adopted pseudo-random generators based on a uniform probability distribution. The described settings represented the default configuration. Later on, we applied further transformations to the generated markers and the image.

Similarity Transformation

To justify our use case, we demonstrated the effect of similarity transformations before perspectively distorting the image. The translation and rotation would demonstrate that markers could be positioned arbitrarily in a real environment provided they shared the same planar surface. The change in scale showed that markers could be of different sizes. The similarity transformation was simulated by applying random rotation from the interval $[0, 360)$ degrees with origin in the marker center. Then, we generated a random coordinate shift from interval $[-20, 20]$ pixels for translation in x and y direction. However, an identical translation had to be applied to the entire marker to prevent distortion. Subsequently, uniform scaling was performed with the origin in the marker center with a scale factor randomly generated from interval $[0.8, 1.5]$. Due to this range, a ratio of the

marker to image area ranged from 1.0% to 1.9%.

Perspective Distortion

The most important transformation was the change in perspective. To this end, we simulated a 3D rotation of an image around its center to represent a change in perspective on the plane that contained several markers. We rotated the image around its center in x , y , and z axis by a random angle from interval $[-20, 20]$ degrees to accomplish a change in perspective. The original keypoints were transformed along with the entire image, producing the warped keypoints.

Noisy Point Correspondence

To simulate a noisy point correspondence, we applied a random noise (translation) to each x and y coordinate of the warped keypoints from the interval $[-2, 2]$ pixels. At this stage, each keypoint was modified in isolation to achieve the distortive effect. Thanks to the perspective deformation, the generated random shift represented different levels of noise depending on how much the image had been warped. This step imitated errors in the marker detection, leading to noisy point correspondence.

5.4.2 Evaluation Methodology

Error Computation

The accuracy of the developed method was evaluated by measuring the reprojection error using the Euclidean distance between the original and the rectified pixel positions. To obtain an error over the entire image, we computed the error for each pixel. Specifically, let w and h be the width and height of the image, respectively. The 3D rotation of a point in the image around the image center that produces perspective distortion is represented by $\varphi(\cdot)$. Let $\mathbf{g}_{i,j}^T = [j, i, 1]$ be the original (ground-truth) pixel position at the i -th row and j -th column, and let $\mathbf{w}_{i,j} = \varphi(\mathbf{g}_{i,j})$ be the analogically defined warped pixel position, for $i = 1, \dots, h, j = 1, \dots, w$. We then compute the 2D reprojection error grid (a $h \times w$

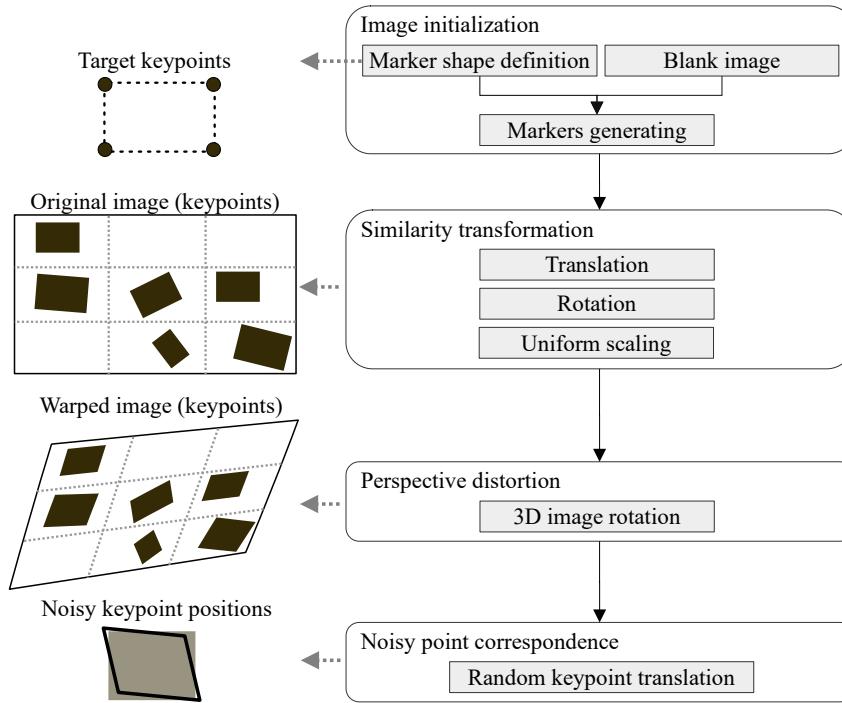


Fig. 5.10: Description of how each one of t test instances in a specific test scenario is created. The input is a blank $w \times h$ image over which m markers are initialized in a uniform grid, which produces the original marker keypoints. Depending on the test scenario, a particular subset of similarity transformations is applied to the entire image. Subsequently, warped keypoints are modified by random noise to simulate noisy point correspondence.

matrix) for the given homography \mathbf{H} as

$$\xi_{wh} = \begin{bmatrix} e(\mathbf{w}_{1,1}, \mathbf{g}_{1,1}) & \dots & e(\mathbf{w}_{1,w}, \mathbf{g}_{1,w}) \\ \dots & \dots & \dots \\ e(\mathbf{w}_{h,1}, \mathbf{g}_{h,1}) & \dots & e(\mathbf{w}_{h,w}, \mathbf{g}_{h,w}) \end{bmatrix}, \quad (5.9)$$

where

$$e(\mathbf{w}, \mathbf{g}) = \|\mathbf{H}\mathbf{w} - \mathbf{g}\|_2. \quad (5.10)$$

To simply express the reprojection error as a single number for the whole image, we adopted an arithmetic mean of all the values in the error grid above, so

$$\xi_{\text{reproj}} = \frac{1}{wh} \sum_{i=1}^h \sum_{j=1}^w e(\mathbf{w}_{i,j}, \mathbf{g}_{i,j}). \quad (5.11)$$

Evaluation Algorithm

On the input, there are m markers (Section ??) and thus an m -to-1 point correspondence. Each marker, by definition, provides a unique homography. Therefore, the aim is to quantify the relative improvement in the reprojection error over the baseline when the k -th ranked homography is used for rectification. Even though we are primarily concerned only with the single, top-performing homography, we evaluate the entire ranking to demonstrate its stable behavior.

We evaluated our homography ranking in terms of reprojection error improvements against the existing approaches based on the isolated homography estimation represented by implementation from the OpenCV [?] library. Since our method provides a ranking, we compare our performance against a random marker selection based on uniform probability distribution. We refer to this performance as the “baseline”; an unbiased marker selection. In practice, the user would rely on “educated guess” when predicting which marker could potentially be the best one to use. To obtain the aforementioned baseline, we evaluated the reprojection error ?? for each marker in isolation and computed the arithmetic mean of these values. When we executed our proposed algorithm, we got the full ordering of markers by their score value computed using the proposed criterion ???. We expected that if the first marker were used to rectify the image, then the reprojection error would be minimal (and lower than the baseline error). If any subsequent marker in the given order were used instead, the reprojection error would increase.

We computed the relative improvement in % for each k -th homography according to the baseline performance. Each test scenario was evaluated one by one. For each test instance, we obtained a k -dimensional vector where its elements represented a percentual improvement at each k -th position. We represented our data as a $t \times k$ matrix, where t was the number of test instances. We treated each column independently to compute the statistics. The details of our evaluation algorithm are described in Algorithm ???. For simplicity, we show an evaluation of just a single instance.

5.4.3 Experimental Results

Fig. ?? shows how the reprojection error varies with respect to the marker position. We can see that the marker position can be deduced by looking at the heatmap represent-

Algorithm 3 Homography ranking evaluation algorithm.

```

1:  $\bar{\mathbf{H}}, \omega \leftarrow \text{RANKHOMOGRAPHIES}()$             $\triangleright$  apply homography ranking (Algorithm ??)
2:  $e_b \leftarrow 0$                                           $\triangleright$  baseline error, initially zero due to summation
3:  $\mathbf{e} \leftarrow \text{array}[m]$                                  $\triangleright$  empty array to store reprojection errors
4:  $\mathbf{p} \leftarrow \text{array}[m]$                                  $\triangleright$  empty array to store relative improvements
5: for  $i \leftarrow 1, \dots, m$  do                                $\triangleright$  for each marker
6:    $\mathbf{e}[i] \leftarrow \xi_{\text{reproj}}$                           $\triangleright$  compute reprojection error (Equation ??)
7:    $e_b \leftarrow e_b + \mathbf{e}[i]$                              $\triangleright$  update baseline error
8: end for
9:  $e_b \leftarrow e_b/m$                                       $\triangleright$  compute the mean reprojection error
10: for  $i \leftarrow 1, \dots, m$  do                              $\triangleright$  for each marker
11:    $k \leftarrow \omega[i]$                                      $\triangleright$  position of  $i$ -th best homography
12:    $\mathbf{p}[i] \leftarrow (e_b - \mathbf{e}[k]) / e_b$            $\triangleright$  compute the relative improvement
13: end for
14: return  $\mathbf{p}$                                           $\triangleright$  return the array of relative improvements

```

ing the pixel-wise reprojection error over the image. The transformation achieves the best accuracy in the marker neighborhood and steadily decreases for more distant pixels. However, not all markers are subjected to the same pattern of error variation. This observation was the core motivation for our solution. We aim to choose the marker that minimizes the pixel-wise reprojection error within the region of the image that is as broad as possible. That is why we evaluate our method by computing the reprojection error over each pixel, not just the keypoints.

All tested scenarios depict similar trends as shown on the plots in Fig. ??, Fig. ??, Fig. ?? and in Fig. ???. The box plots extend from the lower to upper quartile values, with the thin and thick lines representing the median and mean, respectively. The plots discussed further show relative improvements over the baseline OpenCV [?] method. We evaluated relative improvements for the sake of interpretability. For better comprehension, we suggest to see Table ???. It contains individual test scenarios and their corresponding top performances in percents. Conversely, the reprojection error in absolute terms is difficult to interpret without additional context. Nevertheless, to highlight the differences in reprojection errors we also provide absolute values in Table ???. The presence of noise shifted the errors by multiple magnitudes, but still preserved the pattern of distribution.

Influence of Similarity Transformations

In this test scenario, we tested in isolation each allowed similarity transformation, *i.e.*, translation, rotation, and uniform scaling. Fig. ?? demonstrates that the relative im-

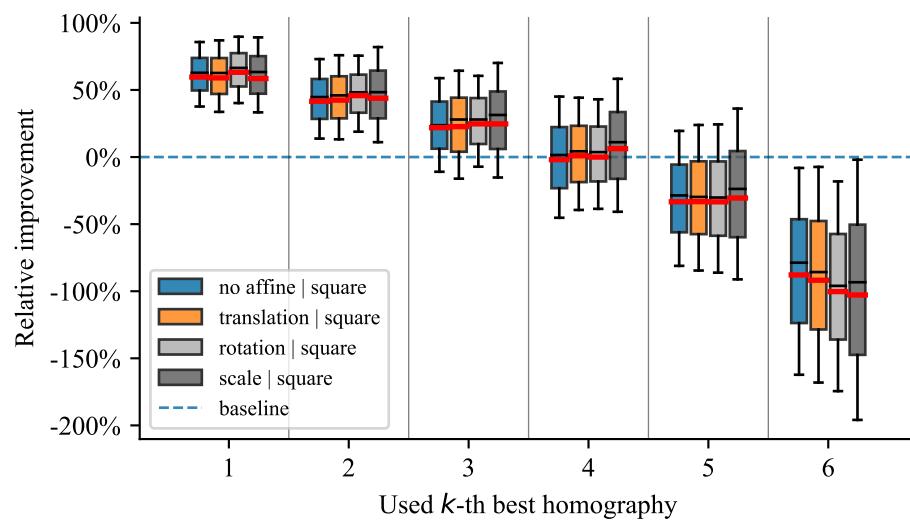


Fig. 5.11: Influence of similarity transformation on the reprojection error.

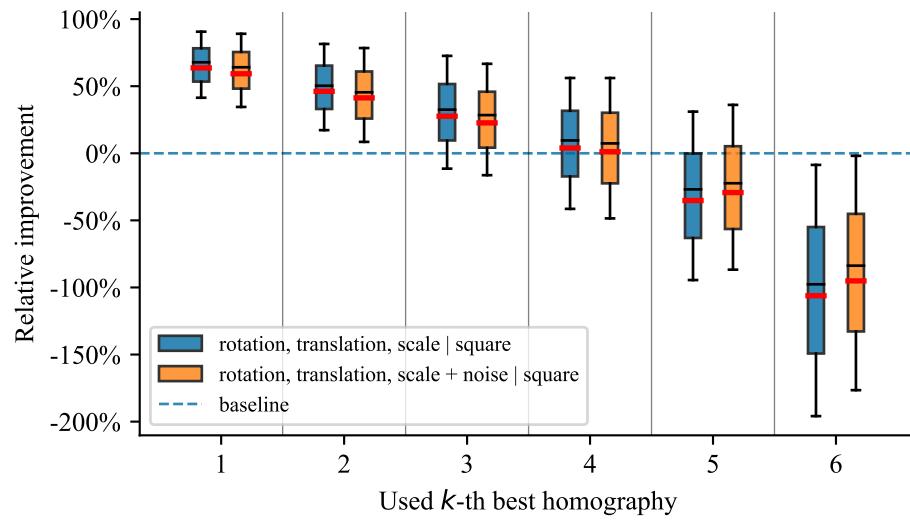


Fig. 5.12: Influence of noise applied to the warped keypoints representing a noisy point correspondence.

provement was circa equal in all situations. Besides, we show that the proposed method is practically invariant to similarity transformations allowing the markers to be in arbitrary positions in a plane. When all similarity transformations were utilized, our method performed even better, showing its stability and robustness.

Influence of Noise

In Fig. ??, we can see the effect of noisy point correspondence that simulated inaccurate keypoint detection. The ranking method preserved the trend of the relative improvement in presence of noise. Absolute reprojection error demonstrated that unless noise was present, the errors varied on sub-pixel levels, so they were practically zero.

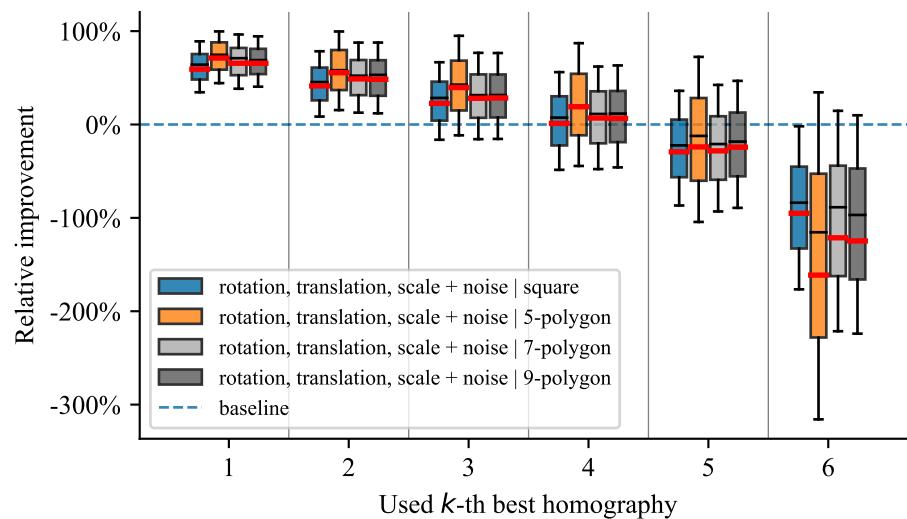


Fig. 5.13: Results for different marker shapes.

Influence of Variable Shapes

We expected that the relative improvement of our method should be invariant to variable shapes as long as they were similar. Fig. ?? demonstrates that with an increasing number of keypoints our method consistently preserved its capabilities. Introducing more complicated shapes than just rectangles did not exacerbate the outcome of the algorithm.

Influence of Number of Markers

We tested a variable number of markers to demonstrate that our method preserved its improvement. Fig. ?? shows that the greater the set of markers, the better the relative improvement. Even when we used just three markers, the proposed method achieved a 46.91% median relative improvement. While it is beneficial to use a larger number of markers, we believe that the improvement we can obtain from an increasing number of markers has a logarithmic trend. On the extreme side, if we used only one marker, there would be no improvement since there would be only one homography to choose from.

5.5 Conclusion

We proposed a method that builds on top of existing approaches for homography estimation that utilize existing point correspondences. The method is a systematic ranking of a set of homography matrices while exploiting the proposed score function to establish the order. Each homography in such a set belongs to a specific marker.

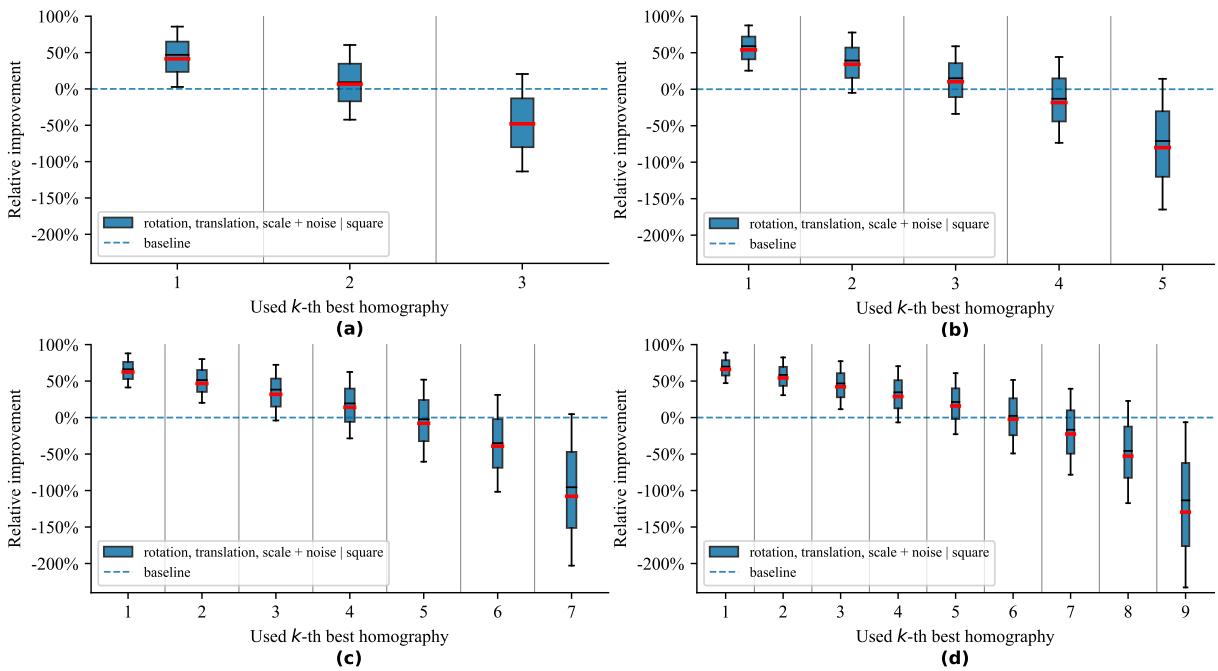


Fig. 5.14: Influence of different number of markers on reprojection error. We experimented with (a) three, (b) five, (c) seven, and (d) nine markers.

We consistently demonstrated that the proposed solution is robust in presence of noise in the point correspondences. These correspondences can be either algorithmically found using feature-matching algorithms or annotated manually, but one has to keep in mind that even human annotations are often inaccurate. We also showed the robustness of our method to a varying number of markers and a change in shape.

Generally speaking, all the improvements at individual ranking positions steadily decreased, reaching 0% improvement at around $\frac{2}{3} m$, where m is the number of markers. A practically applicable statement would be the following: “the first half of ranked homographies yields a better reprojection compared to the baseline on average.”. The baseline performance was given by an average OpenCV [?] reprojection error under the assumption of no prior preference of specific markers, hence the random marker selection.

An advantage of our algorithm is that it is invariant to the underlying homography estimation method. It can, therefore, serve as an extension to all existing or future approaches that handle point correspondences, either as part of run time or a post-processing stage. Moreover, it is computationally efficient, as it scales well with a quadratic complexity $\Theta(m^2)$ in the number of markers, which is usually a single-digit number.

The proposed homography ranking found a real-world application within our solution for the university-related *Interreg SK-CZ* project where we tackled the problem of tracking

Table 5.1: Description of test scenarios in our synthetic dataset with corresponding settings and results for the top-ranked homography. One row represents one test scenario. Four visually separated groups (from top to bottom) are related to experiments shown in Fig. ?? - ??.

shape #	trans.	rot.	scale	noise	relative improvement			absolute improvement		
					median	mean	stdev	median	mean	stdev
square 6	no	no	no	no	62.80%	59.63%	19.64%	0.0003	0.0003	0.0001
square 6	yes	no	no	no	62.65%	59.00%	19.72%	0.0003	0.0003	0.0001
square 6	no	yes	no	no	66.42%	63.17%	19.11%	0.0004	0.0004	0.0002
square 6	no	no	yes	no	63.38%	58.51%	23.97%	0.0002	0.0003	0.0002
square 6	yes	yes	yes	no	67.82%	63.66%	20.30%	0.0004	0.0004	0.0002
square 6	yes	yes	yes	yes	64.11%	59.26%	22.12%	22.0781	24.3177	15.0085
5-poly 6	yes	yes	yes	yes	74.67%	71.19%	21.98%	69.5553	336.2653	685.7427
7-poly 6	yes	yes	yes	yes	71.02%	65.63%	22.99%	46.7939	135.6574	395.7526
9-poly 6	yes	yes	yes	yes	68.97%	65.57%	21.98%	44.9763	115.1219	309.2720
square 3	yes	yes	yes	yes	46.91%	41.36%	31.58%	14.7750	18.1155	20.6746
square 5	yes	yes	yes	yes	59.03%	53.91%	24.56%	19.7629	22.5333	16.0080
square 7	yes	yes	yes	yes	66.19%	62.41%	19.98%	23.8768	27.1364	32.2853
square 9	yes	yes	yes	yes	69.86%	66.09%	18.18%	25.6645	26.6838	11.6975

vehicles for the purpose of speed and dimension estimation. Homography mapping was a necessary part of our approach. However, we did not continue with this branch of research due to the lack of available datasets that we would require for a deep learning-based object tracking solution involving perspective projections.

Chapter 6

Developed Approaches to Visual Object Tracking

In this chapter, we dive into our contributions to the field of Siamese-based VOT. Every section in this entire chapter follows more-or-less the same pattern. It consists of four sections in total, and every section begins with motivation, then elaborates on some particular solution, and finishes off by providing experimental evaluation followed by a discussion. We start by commenting on why we think the most important tracker, the SiamMOT framework, was chosen as a base for our experimentation, and then proceed to a general description of this approach (Section ??). Subsequently, we propose our first enhancement to this tracker based on top of external neural network used for object ReID (Section ??). In the following section, we describe our experiments aimed at incorporating learning metric embeddings into the end-to-end training pipeline of the SiamMOT tracker itself (Section ??). We close the chapter by examining the inclusion of attention mechanism and a deformable convolution operation, once again, by altering the end-to-end tracking pipeline (Section ??).

6.1 Siamese Multi-Object Tracking Framework

6.1.1 Motivation For Model and Dataset Selection

Our initial motivation to adopt the SiamMOT [?] tracker as the base for our custom implementation was that it exploited the best practices honed within the SOT community

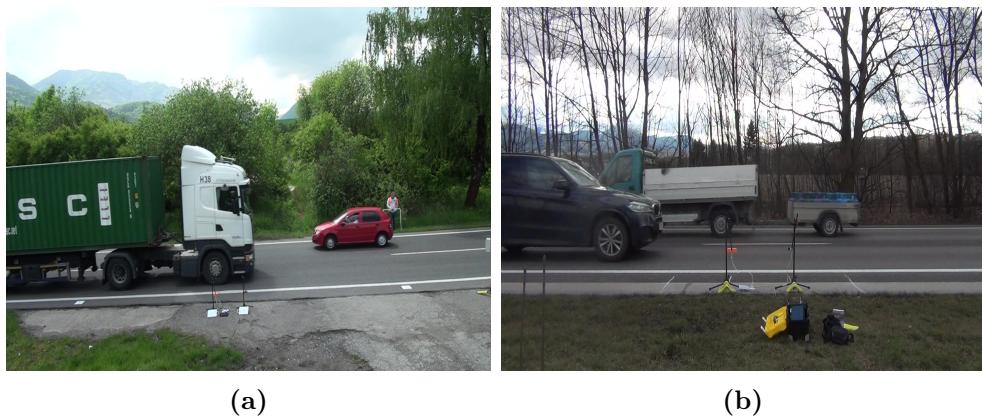


Fig. 6.1: Road scenes we encountered when working on the *Interreg SK-CZ* project.

and that it had been published very recently, reaping a SOTA performance back then. Besides, we saw several potential parts for improvement, namely its ability to inherently handle trivial cases of short-term occlusion during the identity association stage. Even though we were above all concerned with Siamese SOT, we also knew that traffic analysis is inherently preoccupied with multiple objects. But still, we tried to avoid MOT at the beginning as it introduced additional problems to overcome.

However, training, as well as evaluation of the developed tracker, requires a lot of annotated data. Admittedly, we devoted a great deal of time to finding datasets for tracking single vehicles, but to no avail. Nevertheless, the *Interreg SK-CZ* project provided a lot of data that involved traffic scenes. We mentioned this project as part of our discussion related to homography ranking (Chapter ?? on page ??). The major drawback is that the data encompass only a limited view of the road (Fig. ??).

Consequently, we set out to search for different datasets, preferably with a target use of benchmarking tracking algorithms. And we found the UA-DETRAC dataset (Section ?? on page ??), which we consider one of the best available datasets. The data fit the task of traffic analysis using a static camera under various conditions very well. However, one needs to officially ask for the data annotations. We were unfortunate enough that the public web dedicated to this benchmark was broken for a long time and we had to personally negotiate with the authors to provide unofficial access to the data.

6.1.2 General Description

The authors of [?] tracker focused on improving online MOT. As far as their methodology was concerned, they employed a region-based approach [?] in conjunction with a Siamese

multi-object tracking network, hence the name SiamMOT. Broadly speaking, this architecture employs a Siamese tracker for motion estimation between two frames. We would like to note that all the principles so far discussed regarding Siamese trackers apply here. However, as already suggested, the adoption of RPN enables this framework to have more information available. Not only there is the motion prediction from the Siamese tracker, but there are also detections produced by a Faster R-CNN object detector [?] that is integrated within the whole architecture. Subsequently, an online solver is utilized to merge the predictions obtained from the tracker and detector heads.

We will dissect this framework in great detail since we have studied it scrupulously. We performed multiple experiments, many of which did not yield expected improvements. Nevertheless, the practical part of our work was focused on contributing to the open-source repository dedicated to this project developed by several Amazon researchers [?]. We followed a standard path of how contributing to open source projects should be done in a transparent and, more importantly, compatible fashion. We initialized a fresh fork of this project on our personal GitHub account [?] to preserve as much compatibility with the original software as possible and not to strip ourselves of the opportunity to easily receive potential updates from the original repository.

During our development, we often engaged in discussions related to this project incentivized by other researchers who were also working on the very same codebase and trying to either only apply this work to their specific use case or even extend the model. Our detailed knowledge of this model acquired through deliberate work on this project often helped several other programmers who dealt with various issues. From the programming standpoint, our work involved a considerable amount of programming, even though the base architecture was provided and was fully functional from the start, even though the initial setup lasted for several days as well, mainly due to libraries support and compatibility issues. We would like to emphasize that the project consisted of several thousand lines of source code programmed purely in Python programming language. Concerning the deep learning aspect, the PyTorch library [?] was primarily used. It is a widely known library aimed at building deep neural network models while exploiting automatic differentiation.

6.1.3 Model Architecture

The two key aspects of the SiamMOT architecture are Faster R-CNN [?] object detector and Siamese tracker. The salient element of the Faster R-CNN is the RPN. Simply put, SiamMOT adds a region-based Siamese tracker along the standard 2-stage object detection pipeline in order to model instance-level motion.

As depicted in Fig. ??, the input consists of two frames, namely \mathbf{I}^t and $\mathbf{I}^{t+\delta}$, accompanied by a set of detected object instances $\mathbf{R}^t = \{R_1^t, R_2^t, \dots, R_i^t, \dots\}$ at time t . During inference, the detection head produces a set of detected object instances $\mathbf{R}^{t+\delta}$ whilst the tracker's task is to propagate the detections \mathbf{R}^t to time $t + \delta$, and yielding the tracker output denoted as $\tilde{\mathbf{R}}^{t+\delta}$. Please note that it is not the output of the entire tracker, only of the Siamese tracker itself. These instances have to be further processed.

This framework relies on a motion model that tracks each detected object instance from time t to $t + \delta$. A specific BBOX R_i^t at time t is propagated to its future counterpart $\tilde{R}_i^{t+\delta}$ at time $t + \delta$. This process is then completed by a spatial matching phase the objective of which is the association of the tracker output $\tilde{R}_i^{t+\delta}$ with detections $R_i^{t+\delta}$ at time $t + \delta$ such that detected instances are linked from t to $t + \delta$.

Assume there is a specific object instance i detected at time t . Then, the Siamese tracker searches for this particular instance at frame $\mathbf{I}^{t+\delta}$ while exploiting a contextual window spanning a fixed neighborhood of the object's location (*i.e.*, R_i^t) at frame \mathbf{I}^t . In order to define this step more formally, consider the following dependency:

$$\left(v_i^{t+\delta}, \tilde{R}_i^{t+\delta} \right) = \mathcal{T} \left(\mathbf{f}_{R_i}^t, \mathbf{f}_{S_i}^{t+\delta}; \Theta \right), \quad (6.1)$$

where \mathcal{T} is a module (head) represented by the Siamese tracker with learnable parameters Θ . In light of the already stated efficiency of this framework in terms reusing information as much as possible, the module \mathcal{T} is trained on shared feature maps extracted from the backbone using ROI-align operations. As a short reminder, a basic Siamese tracker uses an exemplar image encoded as a kernel to search for the occurrence of the corresponding object in a future frames over a specific search region that should be, by definition, greater than the exemplar region. Thus, the feature map $\mathbf{f}_{R_i}^t$ is extracted over the region R_i^t contained in the frame \mathbf{I}^t . Analogically, the feature map $\mathbf{f}_{S_i}^{t+\delta}$ is extracted over the search region $S_i^{t+\delta}$ delineated in the frame $\mathbf{I}^{t+\delta}$. The region $S_i^{t+\delta}$ is computed by simple

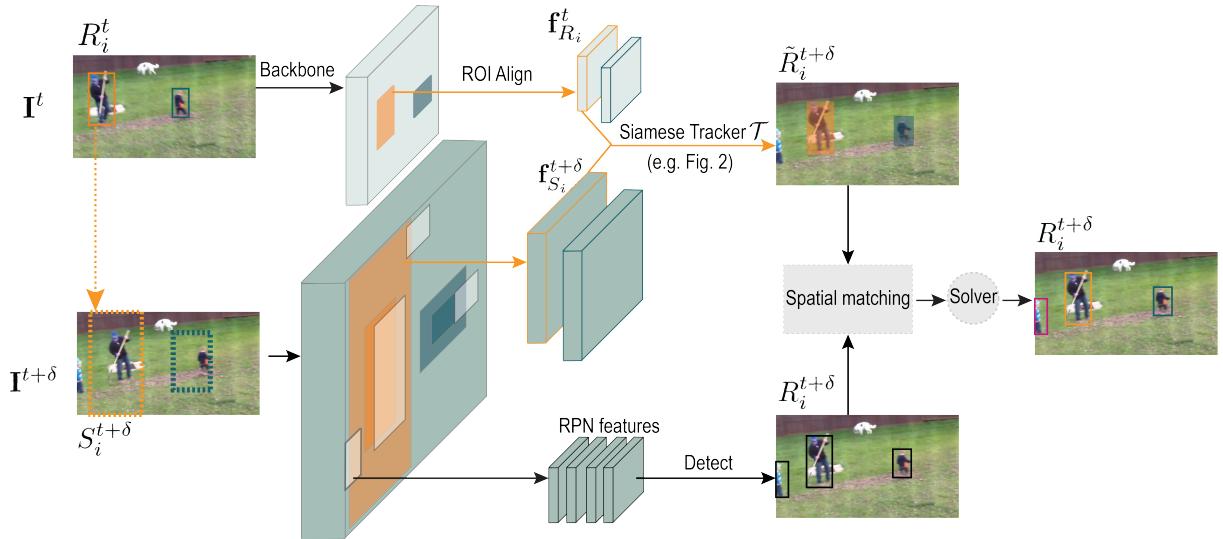


Fig. 6.2: The SiamMOT tracker that detects and associates object instances simultaneously. The Siamese tracker in the top branch serves the purpose of predicting the motion of objects, thus facilitating the temporal linking of objects in an online fashion. Simply put, the Siamese tracker module can be thought of as a single object tracker with all the pros and cons we have discussed so far. On the other hand, a 2-stage object detection is performed as part of the bottom branch. These two branches are then merged using a solver that spatially and temporally attempts to match tracker and detector predictions to produce the tracker output. Note that the feature map corresponding to the frame I^t is shrunk to $1/2$ of its actual size to fit the figure. Backbone features are identical in terms of tensor shapes for both inputs. (source: [?])

expansion of the region R_i^t by a factor r , such that $r > 1$, while preserving the location of the geometric center, as illustrated in Fig. ?? by the dashed BBOX. Last but not least, $v_i^{t+\delta}$ represents the visibility confidence for the detected instance i at time $t + \delta$. This visibility score reflects the tracker’s prediction confidence, and so the value $v_i^{t+\delta}$ should be high if the instance is visible in $S_i^{t+\delta}$, otherwise the value should be low. On top of this formulation that is reminiscent of single object tracking, in the MOT context Equation ?? is applied multiple times, *i.e.*, for each object detected in frame t , signified by $R_i^t \in \mathbf{R}^t$. However, from implementation’s perspective, all these operations can run in parallel and thus the backbone features are computed only once.

The authors conjectured that motion modeling is of paramount importance for online MOT. Given our experience in this field so far, we agree. The motion modeling is practically responsible for association between \mathbf{R}^t and $\mathbf{R}^{t+\delta}$. Despite its efficacy, there are still issues to be addressed. The association fails due to the following reasons:

1. if $\tilde{\mathbf{R}}^{t+\delta}$ does not match to the correct object instance in $\mathbf{R}^{t+\delta}$,
2. or if $v_i^{t+\delta}$ is low (below a specific threshold) for a visible object at time $t + \delta$.

To tackle the problems outlined above, the Siamese tracker exploits various SOTA techniques developed in the single-object Siamese tracking community. We affirmatively approve of the authors' decisions, since we deliberately elaborated on multiple aspects that this tracker heavily relies upon in our survey on Siamese tracking [?].

As far as the Siamese part of the SiamMOT is concerned, the authors dubbed their technique as “explicit motion modeling”. They also worked with “implicit motion modeling”, but that branch of experiments was neither sufficiently expanded in the paper nor it is of particular importance to our research due to its inferior performance.

Explicit Motion Modeling

The most fundamental aspect of Siamese trackers is the cross-correlation operator (Section ?? on page ??) to generate a 2D response map. In SiamMOT, this operation correlates each location of the search feature map (belonging to the search region) $\mathbf{f}_{S_i}^{t+\delta}$ with the exemplar feature map $\mathbf{f}_{R_i}^t$ to produce a response map

$$\mathbf{r}_i = \mathbf{f}_{S_i}^{t+\delta} \star \mathbf{f}_{R_i}^t. \quad (6.2)$$

Therefore, each map r_i captures a different aspect of similarity at every pixel.

Inspired by the Fully Convolutional One-stage Object Detector (FCOS) [?] visual object detector, this tracker adopts fully convolutional network ψ to facilitate instance detection using the response map \mathbf{r}_i . Besides, the so-called *centerness* is also utilized in this architecture (discussed later). The network ψ enables a prediction of a dense visibility confidence map \mathbf{v}_i . Every pixel of \mathbf{v}_i is used as an indicator of the likelihood that this pixel falls within the location of the target object. Besides, a dense location map \mathbf{p}_i is also predicted with the goal of encoding offsets from that particular location to the top-left and bottom right BBOX corners. Consequently, the instance region at (x, y) can be derived by the transformation

$$\mathcal{R}(\mathbf{p}(x, y)) = [x - l, y - t, x + r, y + b], \quad (6.3)$$

where $\mathbf{p}(x, y) = [l, t, r, b]$, *i.e.*, individual corner offsets. This map is then decoded as

$$\begin{aligned}\tilde{R}_i^{t+\delta} &= \mathcal{R}(\mathbf{p}_i(x^*, y^*)) \\ v_i^{t+\delta} &= \mathbf{v}_i(x^*, y^*) \\ \text{s. t. } (x^*, y^*) &= \arg \max_{x, y} (\mathbf{v}_i \odot \boldsymbol{\eta}_i),\end{aligned}\quad (6.4)$$

in which \odot symbolizes the element-wise multiplication, $\boldsymbol{\eta}_i$ incurs a non-negative penalty score throughout the entire candidate region computed as

$$\boldsymbol{\eta}_i(x, y) = \lambda \mathcal{C} + (1 - \lambda) \mathcal{S}(\mathcal{R}(\mathbf{p}(x, y)), R_i^t). \quad (6.5)$$

Here, the letter λ , such that $0 \leq \lambda \leq 1$, is a weighting coefficient, \mathcal{C} is the cosine-window function with respect to the geometric center of the previous target location given by R_i^t , and \mathcal{S} is a Gaussian function that is supposed to penalize the height-to-width ratio changes between candidate region $\mathbf{p}(x, y)$ and R_i^t . The penalty map aims to discourage abrupt changes in target location between individual frames during the course of tracking. This technique is widely adopted in Siamese trackers (Fig. ?? on page ??).

Loss Function

The loss function for this model consists of multiple parts and for its completeness it requires a triplet $(R_i^t, S_i^{t+\delta}, R_i^{t+\delta})$. The following function is minimized during the training:

$$\begin{aligned}\mathcal{L} = \sum_{\forall(x,y)} l_{focal}(\mathbf{v}_i(x, y), \mathbf{v}_i^*(x, y)) + \\ \sum_{\forall(x,y)} \mathbb{1}[\mathbf{v}_i^*(x, y) = 1] (w(x, y) \cdot l_{reg}(\mathbf{p}_i(x, y), \mathbf{p}_i^*(x, y)))\end{aligned}. \quad (6.6)$$

In the expression above, the pairs (x, y) enumerate all valid positions within the $S_i^{t+\delta}$ region. The loss function dedicated to regression task, *i.e.*, l_{reg} , is formulated as the IoU loss for regression [?, ?]. To address the class-imbalance problem in an effective way, the focal loss for classification [?] given by the term l_{focal} is employed, too. All ground-truth

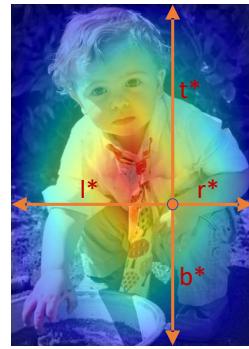


Fig. 6.3: Red, blue, and other colors denote 1, 0 and the values between them, respectively. Centerness is calculated as given by Equation ?? and decays from 1 to 0 as the location deviates from the center of the object. Best viewed in color. (*source: [?]*)

values are marked by the * character. So,

$$\mathbf{v}_i^*(x, y) = \begin{cases} 1 & \text{if } (x, y) \text{ is within } R_i^{*,t+\delta} \\ 0 & \text{otherwise} \end{cases}, \quad (6.7)$$

and

$$\mathbf{p}_i^*(x, y) = [x - x_0^*, y - y_0^*, x_1^* - x, y_1^* - y], \quad (6.8)$$

where (x_0^*, y_0^*) and (x_1^*, y_1^*) correspond to the top-left and bottom-right coordinates of the ground-truth BBOX $R_i^{t+\delta}$, respectively. Following the line of inspiration from the FCOS [?] tracker, the regression loss l_{reg} is additionally modulated by computing the *centerness* for every location (Fig. ??). This concept describes the amount of deviation of the pixel's location from the object center. The reason for adding this score was to suppress locations that are further away from the object's center, because they produced low-quality BBOX predictions. This score is then used as a weighting. The *centerness* coefficient $w(x, y)$ is calculated for each pixel with respect to the target instance $R_i^{t+\delta}$ as

$$w(x, y) = \sqrt{\frac{\min(x - x_0, x_1 - x)}{\max(x - x_0, x_1 - x)} \cdot \frac{\min(y - y_0, y_1 - y)}{\max(y - y_0, y_1 - y)}}. \quad (6.9)$$

6.1.4 Training and Inference Phases

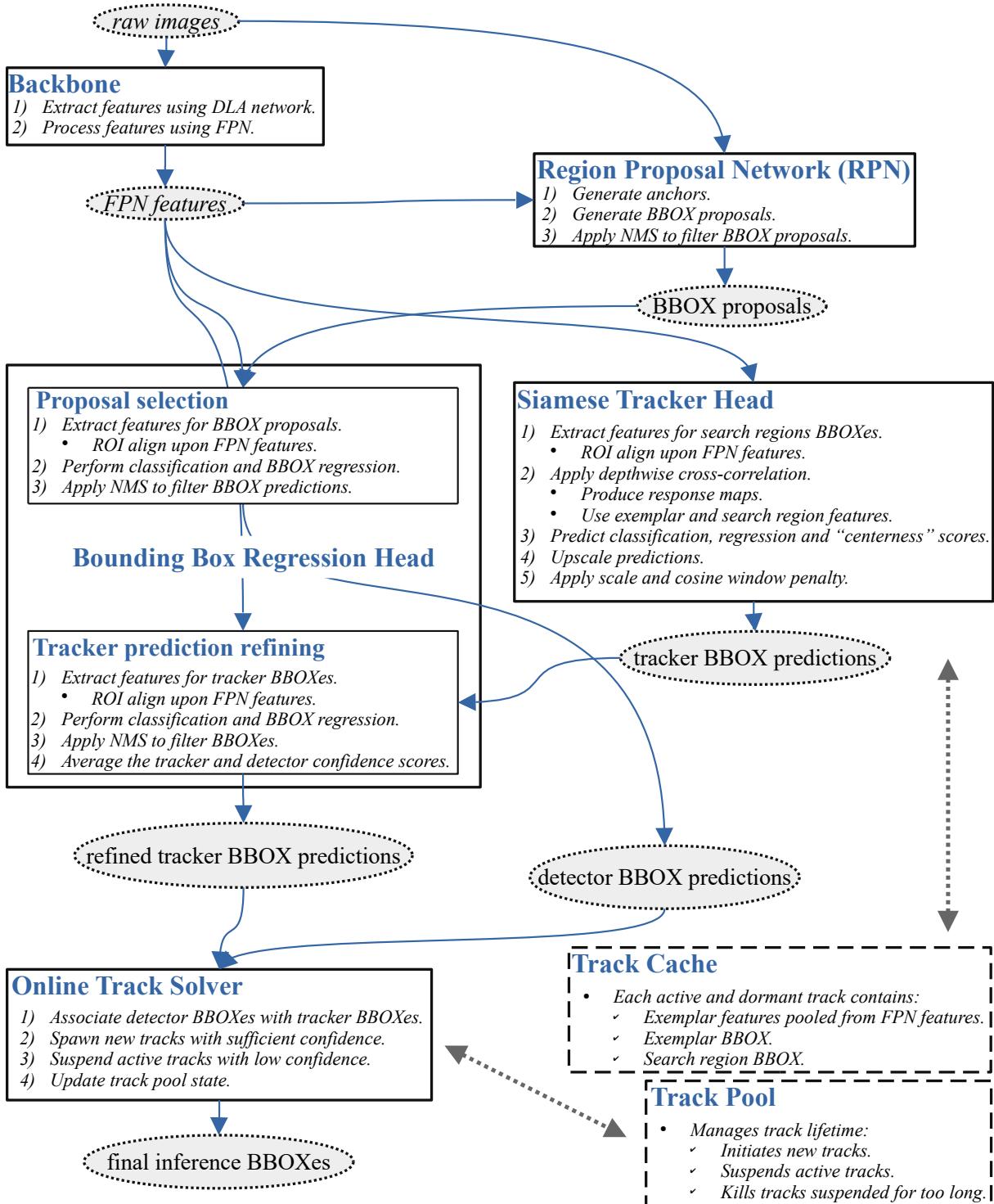


Fig. 6.4: The inference pipeline in the SiamMOT tracker. Backbone features that are a result of intricate DLA and FPN processing are fed into the detector and the tracker. The predictions from the tracker are further refined using the detector head. The two heads function on top of backbone features through the ROI align operations. During the inference, the “online solver” works only with the final BBOXes. It utilizes a simple caching mechanism to store the features belonging to active or dormant objects. The “track pool” module manages the track lifetime.

The SiamMOT model can be trained in an end-to-end fashion, which is one of its great advantages in terms of usability. The general loss function can be formulated as

$$\mathcal{L} = l_{rpn} + l_{detect} + l_{motion}, \quad (6.10)$$

where the l_{rpn} as well as the l_{detect} are standard RPN [?] and detection-subnetwork [?] losses, respectively. The l_{motion} loss is used to train the Siamese tracker.

For better understanding, we suggest the reader follows the diagram in Fig. ???. At inference, the well-established IoU-based NMS operation (Section ?? on page ??) is exploited to process the outputs of the detection and tracker subnetwork independently. The subsequent phase aimed at spatial matching is used to merge detections with the tracker output. This stage also involves the already mentioned IoU-based NMS operation.

The spatial matching and the object identity association happen within the non-trainable module that is activated only during the inference, further referred to as the *online solver*. The purpose is to propagate existing object identities to the future frames given the predictions made by the object detector and object tracker independently. In Fig. ??, we provide a graphical illustration of the original algorithm. The solver algorithm is governed by the following rules listed below. Let v_i^t be the visibility confidence, then

1. the object's trajectory is continued as long as its visibility confidence is above a specific threshold α , otherwise, this trajectory becomes dormant,
2. a new trajectory is spawned in case there is a non-matched detection (during the spatial-matching process) and its visibility confidence is above a threshold β ,
3. a dormant trajectory is terminated, *i.e.*, the object ID will never be used again if its visibility confidence is below α for τ consecutive frames.

This model tackles short-term occlusion. It is one of the key incentives that led us to consider this architecture for experiments, which is the model's ability to handle occlusion. In the beginning, we guessed that this phase could be improved upon due to its inherent simplicity. However, regardless of how rudimentary their approach may seem, it implicitly addresses plenty of frequent cases that appear in object tracking in a competent way. Our endeavors later described often involved a minor improvement in rare situations with simultaneous minor detriments to common situations. Therefore, the outcome of performing worse than the baseline on average was usually inescapable.

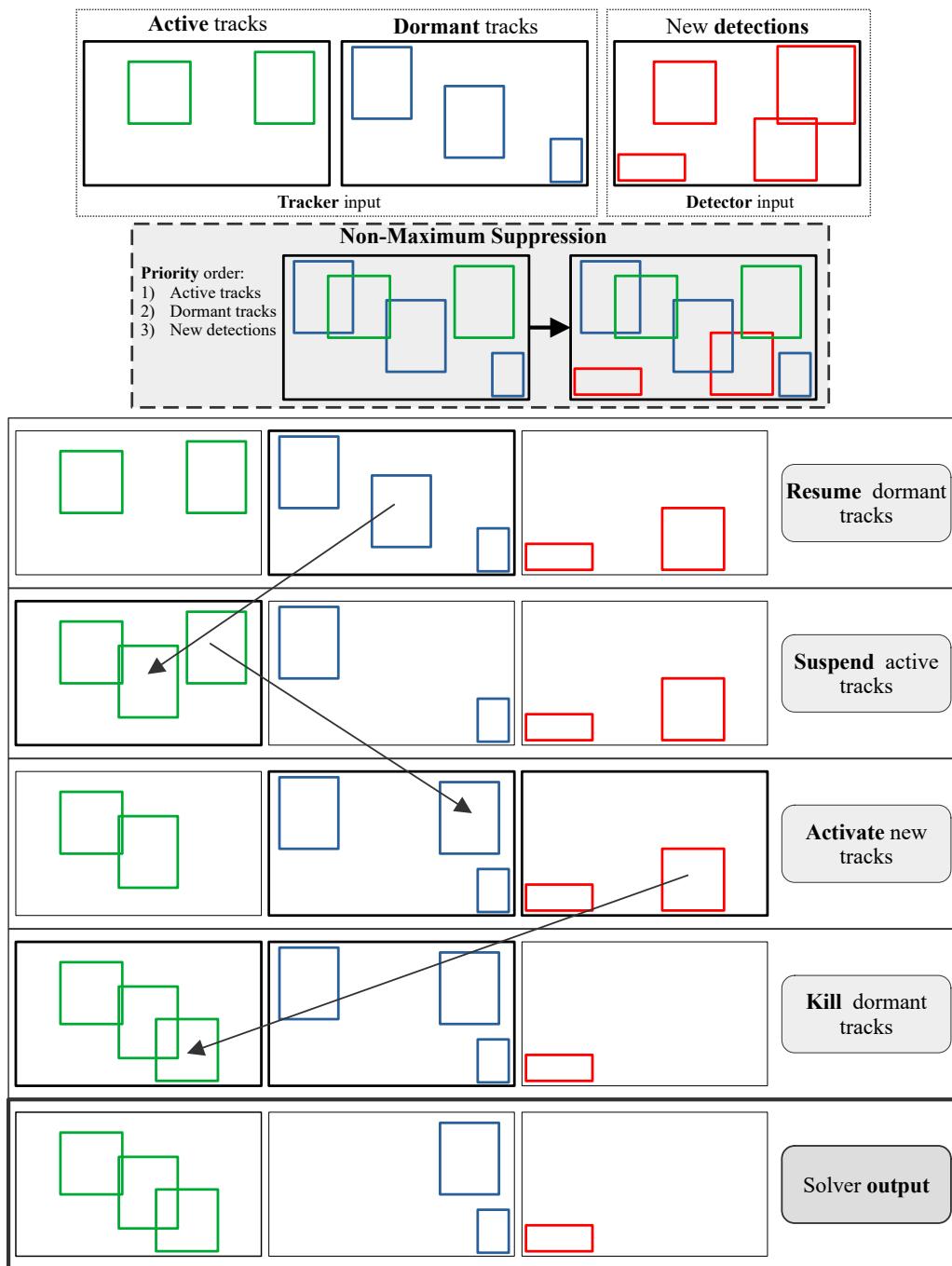


Fig. 6.5: Here we present the process which the tracker, as well as the detector predictions for the current frame, undergo to derive the final tracker output. At first, all the BBOXes are merged upon which a NMS algorithm is performed. The NMS algorithm preserves the priority that active tracks are processed primarily, then followed by the dormant tracks. The remaining detections are filtered as last. The NMS algorithm does not have to be modified at all, since the priority can be induced by altering the associated score values. After the NMS stage, the dormant tracks with sufficient confidence are reactivated. Conversely, active tracks with low confidence are suspended. Then, if there are detections with high confidence, new tracks are established. The process finishes off by permanently killing dormant tracks that have lasted for more than some specific number of frames without instantiation. Note that even though the solver outputs instances of all three categories, only the active tracks are part of the inference.

A short-term occlusion can be defined as having the visibility confidence for the currently tracked object below the threshold α . In SiamMOT, instead of ceasing the track’s existence, the relevant information is kept in memory and thus the search for the exemplar continues until $\tau > 1$ frames have been processed in hope of re-instantiating the object. The most recently predicted location and its corresponding feature frames extracted from the backbone are utilized as the searching template. As a side note, a very similar solver approach has also been adopted in numerous works, such as [?, ?, ?, ?]. In essence, it is simple yet very effective. As our experiments will later demonstrate, it is exceedingly difficult to surpass its performance in general by a significant margin.

6.1.5 Implementation Details

Pre-training

Overall, the model is not difficult to train. By difficult training, we mainly mean instability and sensitivity to hyperparameters. However, the training itself requires a great deal of Graphics Processing Unit (GPU) Video Random Access Memory (VRAM) available to use sufficiently large minibatches. As far as our development experience with the SiamMOT model is concerned, we used the NVidia [?] RTX 2080Ti graphics cards which provides 4352 Compute Unified Device Architecture (CUDA) cores together with 11GB of GDDR6 memory. For training, we often exploited an existing backbone model pre-trained on ImageNet [?] dataset provided by the original authors. Even though there is the entire model available pre-trained on MS-COCO [?] dataset, we decided to avoid it due to the conflicting nature of object classes. We observed better performance when training vehicle detection from scratch rather than trying to “re-wire” the model to dismiss detecting objects we wanted to avoid in the first place, *e.g.*, pedestrians.

Gradient Accumulation

As the majority of our experiments involved expanding the model by adding additional heads, we often struggled with the amount of available GPU VRAM to preserve a reasonable size of minibatches. To this end, we experimented with gradient accumulation, which allows the user to postpone the model weight updating after more minibatches have been processed (Fig. ??). Therefore, the programmer may simulate using larger minibatches

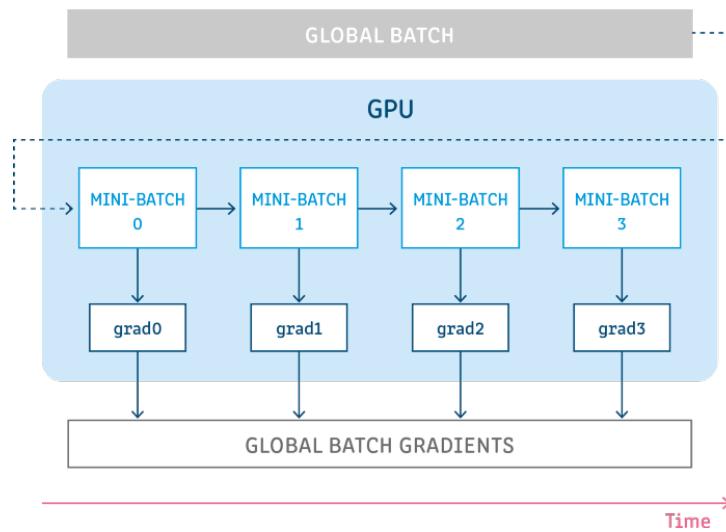


Fig. 6.6: Gradient accumulation is a mechanism that allows splitting the minibatch of samples — used for training a neural network — into several even smaller mini-batches of samples that will be processed sequentially. (source: [?])

than are actually utilized. We emphasize the importance of not updating the model variables must during the accumulation phase so as to ensure that all the mini-batches are processed by the same model to calculate their gradients. Only after accumulating the gradients of the values of the model weights can be adjusted accordingly.

For a brief illustration, let w be a single weight we want to update with respect to the computed gradient using the loss function f . Our goal is to adjust the weight at time t and thus produce the weight at time $t + 1$. The learning rate is denoted by α . So, the gradient update is usually performed as

$$w^{t+1} = w^t - \alpha \nabla f(w^t). \quad (6.11)$$

When using gradient accumulation, the update step is modified as

$$w^{t+1} = w^t - \alpha \sum_{i=1}^n \nabla f(w^t)_i, \quad (6.12)$$

where n is the number of accumulated minibatches.

Considering the advantages outlined above, it is necessary to acknowledge the potential drawbacks. For example, if batch normalization [?] layers are used within the model, then the use of gradient accumulation may have a detrimental effect upon their performance. The reason is that batch normalization layers compute the statistics with respect to a single minibatch. These layers, in their standard formulation and commonly adopted

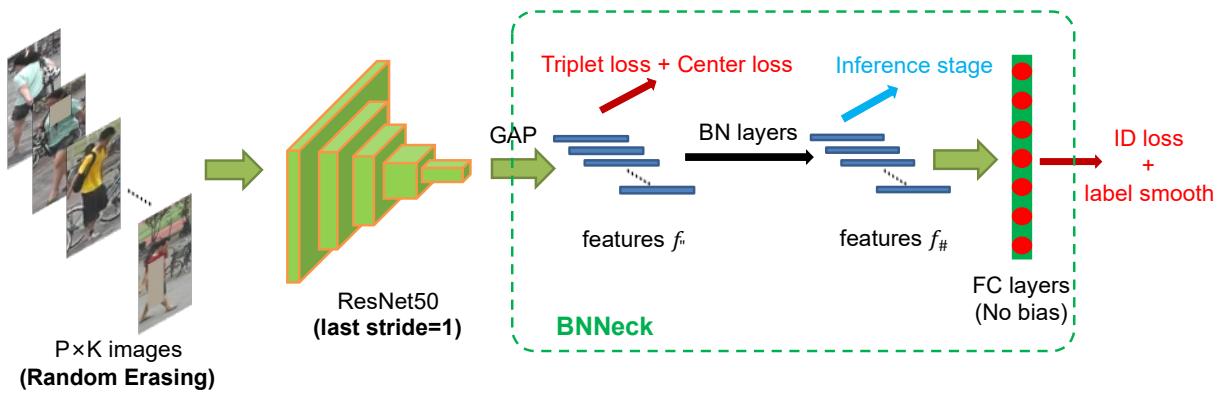


Fig. 6.7: A object ReID baseline which we used for our experiments. (source: [?])

implementation, are incapable of accomodating their inner workings to adequately administer multiple sequential minibatches. However, there is ongoing research in the direction of group normalization as well, namely the works [?] and [?].

6.2 Siamese Multi-Object Tracking and ReID

6.2.1 Motivation

The use of ReID has been emphasized numerous times during our preliminary research for this thesis. We conjectured that once full occlusion ensues, the ReID mechanism could be adopted to recover the lost track since the object appears at the scene as a new one. The presence of occlusion percolates traffic scenes, especially if the camera does not view the road from a higher position. At the beginning of our Ph.D. research, we worked on the *Interreg SK-CZ* project where we tackled vehicle tracking when the camera was positioned next to the road (Fig. ??). Under these circumstances, occlusion was inevitable. Such a setup inexorably led to situations in which a vehicle appeared for a second on the left side of the frame, then ended up fully covered by a truck, and reappeared for a minuscule amount of time at the other side of the frame. The task of maintaining the same object identifier was burdensome and often impossible with a high degree of precision.

6.2.2 Proposed ReID-Enhanced Architecture

We adopted the object ReID approach published in [?], a simple yet very robust framework for person ReID. We employed this architecture (Fig. ??) for vehicle ReID due to its simplicity accompanied with SOTA performance at the time of publishing.

6.2.3 Training Phase

Since this experiment was the first one we embarked upon, we tried to avoid modifications to the underlying tracker framework. Only upon obtaining prospective tracking improvements, we would consider incorporating the object ReID into the whole end-to-end pipeline. As a result, the SiamMOT training phase was not modified. We employed an external object ReID network to perform embedding computation during the inference phase. As far as the training of the ReID network is concerned, we adopted a standard approach of using the triplet loss (Equation ?? on page ??) in conjunction with the categorical cross-entropy loss for additional learning signal. We trained the model using the VeRI-776 dataset (Section ?? on page ??) on which the evaluation was performed, too. The model was trained to produce l_2 -normalized embedding vectors that were then used to measure the degree of similarity between vehicles.

6.2.4 Inference Phase

The online solver algorithm was, from our point of view, the primary suspect for potential improvements due to its inherent simplicity that we considered inferior compared to the rest of the architecture. The aim was to adopt the already trained ReID model to help re-instantiate lost tracks due to occlusion. This external module would be invoked on-demand, in a deterministic way as part of the online solver phase. At this stage of research, we did strive for simplicity rather than efficiency. Consequently, the model inference was substantially impaired by a four-fold speed reduction as measured by FPS, because there was another independent model to process the frames.

We isolated the modifications to the inference phase only to the online solver itself. Thus, we developed a whole new algorithm that handled the newly trained external model as well. The original processing steps of the online solver are outlined in Fig. ???. Our custom inference algorithm is described Algorithm ??.

The algorithm works as follows. The input to the online solver is the set of active tracks, dormant tracks, and newly detected (still unassigned) objects. As in the original implementation, the greatest emphasis during the NMS processing is put on the active tracks, then dormant tracks, and the last ones are the new detections. This ranking establishes the priority with which the NMS algorithm processes the inputs. Our approach

was based on altering the first NMS phase by incorporating only the active tracks and unassigned detections (hence the name “non-dormant” in the pseudocode). Thus, if the overlap between an active track BBOX and detection was too high (above the value of t_{nms}), we considered them to be the same object, and thus the active track was preserved.

Then, after suspending the removed active tracks, we adopted Munkres’ algorithm [?] to solve the linear sum assignment (minimum cost assignment optimization) problem. It is reasonable to assume that a new detection can be assigned to only one dormant track and vice versa, if necessary. The cost matrix was computed based on the cosine similarity (Equation ?? on page ??) between the embedding vectors belonging to the regions of the unassigned detections and the dormant tracks. However, only the visual features where the dormant object was last visible (its visibility confidence was above the threshold t_{track}) were used, as we assumed the object might have been fully occluded. The task was to find an optimal assignment between the new detections with sufficiently high confidence (above the threshold t_{resume}) and the dormant tracks. Then, after obtaining the assignment matrix, we processed all possible pairs and if the similarity between the embedding vectors surpassed the value of t_{sim} , we then resumed the dormant track and re-used the new detection for the object BBOX.

Subsequently, we collected all the remaining unassigned detections the confidence of which was above t_{start} , and initiated new tracks based on them. At last, we searched for active tracks with low visibility confidence (below the threshold t_{track}) to suspend them, thus, making them dormant.

6.2.5 Experimental Evaluation and Discussion

To make a full disclosure right at the beginning, we expected that the adoption of ReID enhancements within the SiamMOT framework would bring improvements in certain areas and for the most part maintain the performance at others. Retrospectively, our experiments showed detrimental effects upon the tracker accuracy. Despite our negative outcome, we still consider the performed research to be a contribution to the MOT area. We will elaborate further on why it is so in order to learn from such an experience as much as possible. Our surprise stems primarily from the fact that the goals of this thesis revolved around the application of ReID in object tracking. Therefore, finding that such an approach does not yield the desired outcome raises multiple questions, specifically the

Algorithm 4 ReID-enhanced online solver inference. This algorithm receives active (\mathcal{A}) and dormant (\mathcal{D}) tracks together with unassigned (\mathcal{U}) detections on the input. Besides, there are four threshold parameters, the NMS overlap (t_{nms}), and three confidence-related inputs, namely the minimum value for the track to stay active (t_{track}), the minimum value for the track to become resumed (t_{resume}), and the minimum value for the track to start from an unassigned detection (t_{start}). Usually, $t_{track} < t_{resume} < t_{start}$. In practice, these values may be 0.3, 0.4, and 0.5, respectively.

```

1: function ONLINEsolverREID( $\mathcal{A}, \mathcal{D}, \mathcal{U}, t_{nms}, t_{track}, t_{resume}, t_{start}, t_{sim}$ )
2:    $\bar{\mathcal{D}} \leftarrow \text{MERGE}(\mathcal{A}, \mathcal{U})$             $\triangleright$  non-dormant items (active tracks, unassigned dets.)
3:    $\bar{\mathcal{D}}_{nms} \leftarrow \text{NMS}(\bar{\mathcal{D}})$          $\triangleright$  non-maximum suppression (Algorithm ?? on page ??)
4:    $\mathcal{A}_{nms}, \mathcal{U}_{nms} \leftarrow \text{SPLIT}(\bar{\mathcal{D}}_{nms})$      $\triangleright$  “after-NMS” active tracks and unassigned dets.
5:    $\mathcal{R} \leftarrow \text{GET\_REMOVED}(\mathcal{A}, \mathcal{A}_{nms})$        $\triangleright$  obtain NMS-removed active tracks
6:    $\text{SUSPEND\_TRACKS}(\mathcal{R})$                        $\triangleright$  suspend NMS-removed active tracks
7:    $\mathcal{U}_{resume} \leftarrow \text{GET\_RESUME\_DETS}(\mathcal{U}_{nms}, t_{resume})$      $\triangleright$  dets. with conf.  $\geq t_{resume}$ 
8:    $\mathcal{E}_u \leftarrow \text{CALC\_EMBEDDINGS}(\mathcal{U}_{resume})$         $\triangleright$  embeddings for the unassigned dets.
9:    $\mathcal{E}_d \leftarrow \text{CALC\_EMBEDDINGS}(\mathcal{D})$               $\triangleright$  embeddings for the dormant tracks
10:   $\mathbf{M} \leftarrow \text{BUILD\_COST\_MATRIX}(\mathcal{E}_u, \mathcal{E}_d)$       $\triangleright$  initialize a cost assignment matrix
11:   $\mathcal{P} \leftarrow \text{MIN\_COST\_ASSIGNMENT}(\mathcal{U}_{nms}, \mathcal{D})$      $\triangleright$  Munkres' algorithm [?]
12:   $\mathcal{U}_{reid\_used} \leftarrow \text{INIT\_EMPTY\_SET}$            $\triangleright$  unassigned detections used by ReID
13:  for all  $u_{idx}, d_{idx} : \mathcal{P}$  do     $\triangleright$  for each pair of unassigned det. and dormant track
14:     $e_u \leftarrow \mathcal{E}_u[u_{idx}]$                    $\triangleright$  embedding vector of the unassigned region
15:     $e_d \leftarrow \mathcal{E}_d[d_{idx}]$                  $\triangleright$  embedding vector of the dormant region
16:    if  $\text{COS\_SIM}(e_u, e_d) \geq t_{sim}$  then  $\triangleright$  similarity check (Equation ?? on page ??)
17:       $\text{RESUME\_TRACK}(d_{idx})$            $\triangleright$  resume dormant track using unassigned det.
18:       $\text{ADD}(\mathcal{U}_{reid\_used}, u_{idx})$      $\triangleright$  unassigned det. was assigned to a dormant track
19:    end if
20:  end for
21:   $\mathcal{U}_{free} \leftarrow \text{GET\_FREE\_UNASSIGNED}(\mathcal{U}_{nms}, \mathcal{U}_{reid\_used})$   $\triangleright$  still free unassigned dets.
22:   $\text{START\_NEW\_TRACKS}(\mathcal{U}_{free}, t_{start})$      $\triangleright$  unassigned dets. with confidence  $\geq t_{start}$ 
23:   $\text{SUSPEND\_TRACKS}(\mathcal{A}_{nms}, t_{track})$          $\triangleright$  active tracks with confidence  $< t_{track}$ 
24: end function

```

following ones:

1. Does the inclusion of ReID into MOT frameworks have a potential to resolve cases of full occlusion without exacerbating other areas?
2. Is the ReID extension suitable to the chosen MOT model, namely, the SiamMOT?
3. Is the target use case in terms of datasets adequate to showcase the potential of ReID applied in the SiamMOT tracker?

Due to an already excessive length of this document and inferior results this experiment brought in terms of tracking accuracy and inference speed, we decided to omit extensive documentation, especially tables with quantitative comparison. There are more important



Fig. 6.8: Partial occlusion in the UA-DETRAC dataset where the object BBOX covers a great part of the region belonging to another object.

experiments that deserve a deeper elaboration.

To say the least, we encountered several situations that indicated improvements. Occasionally, the model was capable of properly identifying the lost object based on the embedding vector similarity. However, such situations were rare. The detrimental effect of the ReID module on the modified NMS phase decreased the tracker’s accuracy. The original online solver uses the NMS algorithm to assign detections to either active or dormant tracks. This phase is very effective and covers a lot of common situations.

Besides the expected slow performance, we noticed approximately 3% reduction in MOTA. The number of misses (FN) also significantly increased as the model failed to properly assign the detection to the dormant track. Overall, there was no metric that would be improved. We either managed to match the original baseline online solver performance, or we performed substantially worse.

Our observations brought the following question. Is it more likely for the occluded object to appear at a completely different position within the frame or somewhere near the position of disappearance? If we constraint ourselves to the *Interreg SK-CZ* project with the road viewed from the side (Fig. ??), then it might hold true most of the time. However, in general traffic analysis, especially in the scenes from the UA-DETRAC dataset (Fig. ?? on page ??), it is scarcely true. Vehicles often re-appear near the position where they last disappeared. As a result, the original online solver dealt with such situations competently. In fact, it outperformed the entire ReID model due to the inherent ability of the Siamese tracker head to implicitly assess the object similarity when producing the response map. Remember that the search region encompasses a square area with sides twice as long as the exemplar region.

Another problem appears during partial occlusion. A vehicle is often severely occluded by another vehicle, so the two BBOXes enclose both objects (Fig. ??). Thus, the embed-

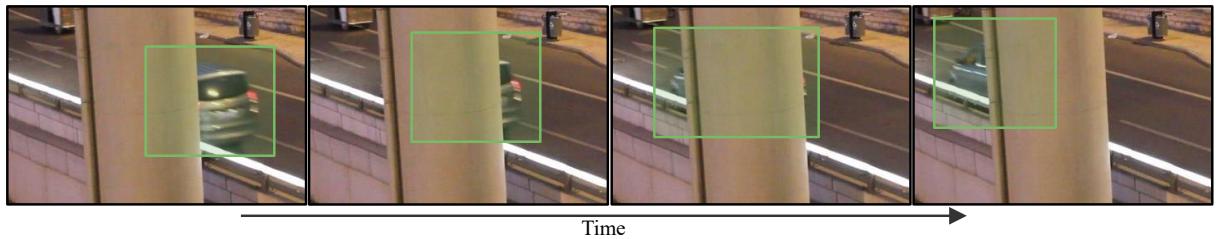


Fig. 6.9: A demonstration of almost complete occlusion in the UA-DETRAC dataset where the object passes behind a pillar. As we can see, the tracker correctly infers the object position based on minor visible cues on the sides of the pillar. However, even if the full occlusion happened, thanks to the search region being four times as large as the exemplar region in terms of the area, this object’s identity would be flawlessly recovered, too.

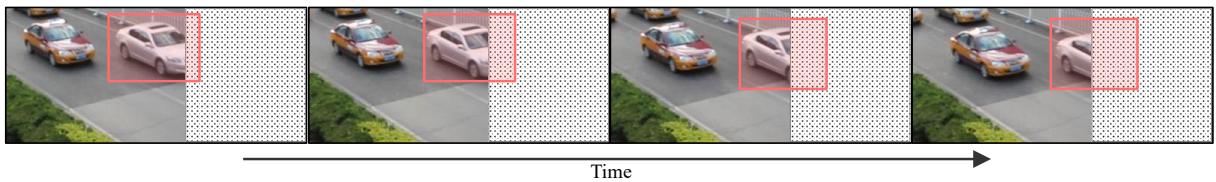


Fig. 6.10: Progressing partial occlusion in the UA-DETRAC dataset where a large part of the object BBOX is covered by a background. In this case, the occlusion comes in a form of a synthetically generated region.

ding distances for the two objects are very close. In terms of the ReID mechanism, the two delineated regions given by the two BBOXes of closely positioned objects with severe occlusion are sometimes considered to be the same object. Consequently, the two tracks are merged into one.

Besides all this, we noticed that the tracker works well for handling partial occlusion. When the tracked object is only partially occluded, the Siamese tracker properly predicts the possible span of the object (Fig. ??). This, despite the obvious benefits, posed another challenge to our ReID extension. We had to store the last visible frame crop for the target object in memory along with its corresponding features for subsequent ReID in the future. However, the situation depicted in Fig. ?? shows one of the situations in which the “last visible exemplar” was polluted by the occlusion. We would need another mechanism to check for partial occlusion and store only the “last visible non-occluded exemplar”, or a sequence of multiple exemplars before the visibility confidence dropped below a specific threshold value. The tracker itself maintains a relatively high confidence score anyway, making the whole approach much more complicated.

Thus, to employ ReID that would utilize the object’s most-recent history to re-establish its identity later, there is a need to systematically detect the level of occlusion. With this in mind, the related methods of template updating are worth exploring, too. For

instance, Zhang *et al.* [?] proposed a replacement for deterministic template updating strategies. They developed a neural network model to learn the optimal template for the next frame while exploiting the current and historical observations of the target object. The template updating based on Gaussian Mixture Model (GMM) utilized in the Siamese Network with Re-detection Mechanism (SiamRM) tracker together with conditional re-detection could handle long-term object tracking in presence of severe occlusion [?]. The GMM served for dynamically updating the template instead of using a fixed template from the initial frame. But, the template was not updated in every frame or a fixed interval. Instead, only templates with sufficiently high confidence were considered. Even though template updating provides benefits to model adaptation, online tracking may become very inefficient. In addition, it still does not solve the tracking drift problem caused by similar interference completely [?]. Moreover, in case of severe occlusion, there is a high risk of the template becoming polluted. We dedicated an entire section to various template updating strategies in our survey [?], and we venture to claim that no consensus has been reached so far.

We conjecture that the ReID is useful for multi-camera scenarios where the task is to re-identify the object from a totally different angle, often without partial occlusion. In common crossroad situations that are abundant in the UA-DETRAC dataset, the first formulation of the online solver approach handles partial occlusion with a lot higher precision than our ReID extension.

This experiment showed that the inclusion of ReID mechanism into the inference algorithm brings a lot more disadvantages than benefits and we did not continue with this path. It was not possible to justify the additional complexity we introduced to the model by being able to improve very few cases according to our expectations and by simultaneously exacerbating the performance in other, more frequent situations. In addition, incorporating this proposal into the tracking pipeline was cumbersome as the underlying design principles had conflicting requirements. For instance, the Siamese tracker head generates predictions even for the dormant tracks. Conversely, we demanded the dormant tracks to become frozen to the last frame where the visibility confidence was satisfactory. This modification was more difficult than we initially expected. There were many other tweaks to the model to make it work.

We think that the architecture in its original formulation is not suited for this type

of extension. In fact, we have only observed the integration of ReID mechanism into the multi-object tracker when the pure tracking-by-detection approach is adopted. After our failure, we are starting to see why. The predictions from the Siamese tracker head have conflicting interests with the ReID extension. Nonetheless, to address the question of whether a joint training of the tracker and embeddings would bring an improvement, we provide the upcoming experiment, discussed next.

6.3 Siamese Multi-Object Tracking and Embedding

6.3.1 Motivation

One of our experiments involved an end-to-end training of the SiamMOT model together with a custom head aimed at embeddings based on ROI-pooled backbone-extracted features for the object BBOX. The goal was to force the training process into extracting features that are not only satisfactory for detection and tracking but also contain the necessary information to create embeddings for ReID purposes during the inference.

We strived for simplicity by extending the processing pipeline without altering the existing infrastructure. From the standpoint of implementation, the SiamMOT project itself is organized in a proper object-oriented and modular fashion, which made this particular development task easy in this aspect. However, as we will discuss further, we encountered setbacks in terms of training stability and we had to take appropriate measures. Furthermore, extending a huge model that already requires a significant amount of GPU VRAM made it even more demanding.

During the research related to our Siamese tracking survey [?], we noticed one work where the exemplar features were projected using Global Average Pooling (GAP) operation into an embedding space consisting of fewer dimensions [?]. The embedding vector was produced using the feature tensor that represented the kernel for the cross-correlation operation utilized by the majority of the Siamese trackers discussed so far.

More concretely, suppose the extracted features were represented by a tensor of shape $8 \times 8 \times 256$. Then, the GAP operation along the channel dimension would produce a tensor of shape $1 \times 1 \times 256$, which could then be further flattened into a single 256-dimensional vector. In the end, the obtained vector was l_2 -normalized and thus projected onto a unit hypersphere. In the work of Li *et al.* [?], these embedding vectors were exploited for

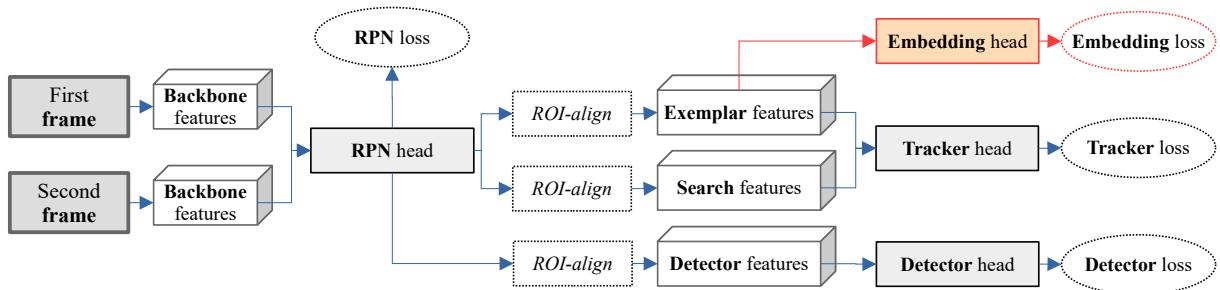


Fig. 6.11: Our extension (shown in red) to the underlying SiamMOT architecture that incorporates vector embeddings to the end-to-end training. This diagram shows the pipeline that is used during the training, not inference.

template updating and for combining multiple templates within a pool of size n in an exponential fashion.

This observation led us to the following hypothesis. Given the fact the Siamese exemplar features do contain some, although probably not sufficient information for pure object ReID, would it be possible to map them further using a non-linear function to produce embedding vectors that could serve for ReID? Such features are just a learned template, therefore, some notion of similarity needs to be already built into it.

6.3.2 Feature Embedding Head Architecture

As far as the vector embedding computation was concerned, we attached the embedding head (Table ??) into the backbone features but after the ROI-pooling operation (Fig. ??). This ensured fixed tensor shapes and allowed us to process the very same features that the object detector and Siamese tracker utilized, too. Simply put, for every proposal made for a particular frame, we looked at the delineated BBOX through the lens of ROI-pooling to extract backbone features. In fact, we simply reused the extracted exemplar features. Later on, we processed these features using our newly devised embedding head to produce feature embeddings. The resulting embeddings were subjected to the triplet loss computation (discussed next) with all the necessary operations such as various types of hard negative mining.

6.3.3 Training Phase

The training phase was altered by adding another loss function to the sum of already existing three losses from the original model. In particular, the general SiamMOT loss

layer	tensor shape	parameters no.
input	[B , 128, 15, 15]	0
conv 3×3	[B , 128, 13, 13]	147 456
ReLU	[B , 128, 13, 13]	0
conv 3×3	[B , 256, 11, 11]	294 912
ReLU	[B , 256, 11, 11]	0
flatten	[B , 30976]	0
linear	[B , 1024]	31 720 448
l_2 -normalize	[B , 1024]	0
total		32 162 816

Table 6.1: Our custom embedding head that we used to process backbone-extracted features to produce embedding vectors. It is built from two convolutional layers separated by a ReLU nonlinearity followed by a fully connected layer which produces a 1024 dimensional feature embedding. The batch size dimension is given by B in the tensor shape. Since each embedding vector is normalized to unit length, we avoided learning biases throughout the whole network.

function defined in Equation ?? was reformulated as

$$\mathcal{L} = l_{rpn} + l_{detect} + l_{motion} + l_{emb}. \quad (6.13)$$

The l_{emb} loss incorporated triplet loss (Equation ?? on page ??). We also experimented with the contrastive loss (Equation ?? on page ??), but the effect was detrimental in every aspect, so we will not discuss it any further. As we remarked in Section ?? on page ?? aimed at latent spaces and embeddings, it is crucial to adopt appropriate sample mining strategies when using the triplet loss. The rationale is that for the training to keep progressing, the model needs to encounter harder and harder triplets to generate sufficient learning signals. To this end, we went for the semi-hard triplet mining strategy (Equation ?? on page ??). However, we struggled with collapsing embeddings [?]. This phenomenon happens when the embedding training forces the model to project all the features onto a single point in the embedding space, thus incurring the loss equal to the used margin. We claim that the use of semi-hard negative mining produced triplets that were too difficult. Since we used all the RPN proposals to generate triplets, one

may imagine that there would always be proposals covering only some small part of the object, making it problematic for the network to learn the concept of “similarity” and “difference” if it only processes very hard images. Nevertheless, these situations are very common in margin-based losses [?]. The computed loss is so high that it is more suitable for the model to map all the features onto a single embedding vector. To remedy this, we implemented batch-all online mining strategy (Equation ?? on page ??), which stabilized the training. We recommend first utilizing batch-all mining strategy during the training, and then slowly proceeding to batch-hard strategy after a certain point. We still have not implemented this approach, because it would be time-consuming to find the right hyperparameters. There are many open questions, such as to mine the RPN proposals in a better way or to set the margin to a higher value. Loss functions aimed at object ReID are notoriously cumbersome to train. One last remark is that we implemented the entire mining algorithm followed by the loss computation in a GPU-only fashion for fast execution and easy integration into the pipeline.

6.3.4 Inference Phase

Feature-based Non-Maximum Suppression

Salscheider [?] proposed an extended NMS algorithm that incorporates a distance between feature embeddings dubbed as Feature-NMS. Considering our idea introduced above, we had to encompass the vector embeddings into the solver reasoning. In the beginning, we came up with the solution that exactly copied the one the mentioned author proposed. That provided further justification for attempting to implement the algorithm and test it in practice. The advantage is that this approach is restricted to the inference phase, thus experimenting with it did not require model re-training.

We assume the reader is acquainted with the original NMS algorithm (more in Section ??.). Nevertheless, here we repeat the same definitions for clarity. Let $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_n\}$ be a set of n region proposals described by n BBOXes. Scores for each detection are contained in a set $\mathcal{S} = \{s_1, s_2, \dots, s_n\}$, where s_i denotes a detection score for the i -th box, \mathbf{b}_i . This time, we are also going to need the associated feature embedding vectors with each BBOX, represented by a set $\mathcal{E} = \{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$. Let \mathcal{B}_{fnms} be the set of filtered proposal instances from the set \mathcal{B} produced using the Feature-NMS algorithm.

The distinction in terms of parameters is the following. The original algorithm required only one threshold for the maximum allowed portion of the overlap between regions. The Feature-NMS requires three parameters discussed below.

- A minimum threshold τ_{lower} denoting the boundary below which the two objects are deemed as different. This value should be low, for example, 0.2, which means that if the IoU between the two objects is less than 0.2, then the two instances should be treated as different objects.
- A maximum threshold τ_{upper} denoting a boundary above which the two objects are considered to be identical. Conversely to the τ_{lower} , this value should be high, for instance, 0.8, which indicates that if the IoU of the two object instances surpasses this threshold, then it should be the same object, and thus, the BBOX with the lower confidence is discarded.
- A threshold δ is used as a decision boundary between the embedding vectors. This threshold should reflect a measure of similarity. If the adopted measure of similarity (cosine distance, Euclidean distance, ...) falls below δ , then the two objects are different, otherwise, they are considered the same one. This value of δ is used only if the two conditions above do not hold.

Here we provide a pseudocode of the Feature-NMS algorithm:

```

1: function FEATURE-NMS( $\mathcal{B}$ ,  $\mathcal{S}$ ,  $\mathcal{E}$ ,  $\tau_{\text{lower}}$ ,  $\tau_{\text{upper}}$ ,  $\delta$ )
2:    $\mathcal{B}_{fnms} \leftarrow \emptyset$             $\triangleright$  initialize the output (filtered) set of region proposals
3:   while  $\mathcal{B} \neq \emptyset$  do            $\triangleright$  loop until all the proposals are processed
4:      $m \leftarrow \arg \max_{i \in \{1, 2, \dots, |\mathcal{S}|}\} \mathcal{S}_i$             $\triangleright$  find an index of a proposal with the highest score
5:      $\mathcal{B} \leftarrow \mathcal{B} - \mathbf{b}_m$ ,  $\mathcal{S} \leftarrow \mathcal{S} - s_m$ ,  $\mathcal{E} \leftarrow \mathcal{E} - \mathbf{e}_m$             $\triangleright$  remove the proposal
6:      $\mathcal{B}_{fnms} \leftarrow \mathcal{B}_{fnms} \cup \mathbf{b}_m$             $\triangleright$  save the proposal with the highest score
7:     for  $i \leftarrow 1$  to  $|\mathcal{B}|$  do            $\triangleright$  iterate through remaining proposals
8:       if  $\text{IOU}(\mathbf{b}_m, \mathbf{b}_i) \geq \tau_{\text{lower}}$  then            $\triangleright$  above the lower-bound threshold
9:         if  $\text{IOU}(\mathbf{b}_m, \mathbf{b}_i) \geq \tau_{\text{upper}}$  then            $\triangleright$  above the upper-bound threshold
10:         $\mathcal{B} \leftarrow \mathcal{B} - \mathbf{b}_i$ ,  $\mathcal{S} \leftarrow \mathcal{S} - s_i$ ,  $\mathcal{E} \leftarrow \mathcal{E} - \mathbf{e}_i$             $\triangleright$  remove the proposal
11:      else
12:        if  $\text{SIMILARITY}(\mathbf{e}_m, \mathbf{e}_i) \geq \delta$  then            $\triangleright$  similarity above threshold
13:           $\mathcal{B} \leftarrow \mathcal{B} - \mathbf{b}_i$ ,  $\mathcal{S} \leftarrow \mathcal{S} - s_i$ ,  $\mathcal{E} \leftarrow \mathcal{E} - \mathbf{e}_i$             $\triangleright$  remove the proposal

```

```

14:           end if
15:       end if
16:   end if
17: end for
18: end while
19: return  $\mathcal{B}_{fnms}$ 
20: end function

```

6.3.5 Experimental Evaluation and Discussion

The proposed embedding-based enhancement was evaluated against the baseline model without the embedding head. For a fair comparison, we made sure that all the hyperparameters were identical. The only hyperparameters that we had to change compared to the original model were the batch size and learning rate. Since the triplet loss requires the computation of a great number of triplets, especially the batch-all strategy, we had to decrease the batch size to avoid crashes due to not having enough GPU VRAM available.

We acknowledge that it is difficult to compare the baseline model with the proposed extension since they progress differently during the training. Therefore, we saved the model state after K training iterations and then evaluated the performance using the validation dataset. The same process was repeated for the baseline architecture, which we trained from scratch using smaller batch sizes. Furthermore, the embedding evaluation also requires the cosine similarity threshold. We tried multiple different values to see the effect.

Add “baseline vs. embedding” 2D scatter plot - MOTA vs. MOTP.

Add “baseline vs. embedding” 2D scatter plot - precision vs. recall.

Add a table with quantitative comparison.

Add an ablation study “ReID training + original solver” to see the effect.

6.3.6 Discussion

We conjecture that our inability to improve the tracker performance was caused by the introduction of the feature embedding itself, as it was the only change during the training phase. Our ablation study also demonstrated that joint training with the embedding head had a negative impact on the tracker, even if the original solver was adopted during the

evaluation. This raises the question of potential task/feature conflict between the heads. We venture to claim the inherent incompatibility of the detection, tracking, and ReID tasks, despite the existence of a recently published work Lu *et al.* [?], who developed their RetinaTrack tracker. This framework exploited the base visual object detector called a RetinaNet [?] and then added, in principle, a very similar head as we did for the purpose of producing feature embeddings. However, there are obvious differences between the two trackers in terms of how the inference phase is executed, and that is where we see one of the possible causes of our failure.

Unfairness between the detection and re-identification

There is recent publication [?] by Zhang *et al.* that discusses the notion of “fairness” of detection and ReID. According to their empirical evidence, there are multiple levels in which the “unfairness” manifests itself. This term was used to describe a situation in which the importance of either ReID or detection task is lessened. We would like to emphasize the recency of this work since several of their remarks and conclusions coincide with ours. Had we encountered this paper prior to conducting our experiments, we would have taken a different path. At least, we would have striven to find another way to incorporate the embedding creation into an end-to-end pipeline. The aforementioned work, among other very important things, introduces a joint tracker that effectively circumvents the obstacles related to conflicting tasks. This framework avoids the use of anchor proposals and is based on Center Point-based Network (CenterNet) [?] object detector.

They argue that the ReID task may become overlooked if the embedding vectors are produced solely from the ROI-pooled regions, as in our case. Consequently, the detection task becomes central in terms of influence upon the loss function, since generating embedding vectors on incorrect BBOXes is meaningless. Thus, the bias to produce accurate object proposals is inescapable.

Next, a problem arises when one anchor corresponds to multiple identities, which is an obstacle we have already mentioned. As a result, the extracted features are not optimal in terms of their accuracy as well as discriminative representativeness. To remedy this, features should be extracted at a single point, *i.e.*, at the estimated object centers (hence the use of CenterNet), instead of region proposals.

An analogical problem is when multiple anchors correspond to one identity. A high

overlap in terms of IoU directs the model to estimate the same identity for nearby anchors. But, even small perturbation may result in falling below a specific threshold, and the anchors became marked as belonging to different objects. In fact, this situation is very common as proposals are generated on the feature level (a coarse representation due to downsampling), and not the pixel level.

Another inherent problem when merging object detection and ReID is feature conflict. We have highlighted numerous times the significance of multi-layer feature fusion. However, not all tasks benefit from this approach equally. In particular, object detection utilizes deep features to estimate object classes and locations, whereas the ReID is more prone to utilizing low-level features owing to their discriminative power. For that reason, it is important to balance the loss optimization.

And last but not least, the feature dimension poses another source of imbalance. The authors remarked that actually learning lower-dimensional features for ReID is better than higher-dimensional. In our research, we thought the opposite is true. High-dimensional embedding vectors notably harm the object detection accuracy due to the competition between the two tasks. The number of feature dimensions for detection is usually very low (compared to 1024D or 2048D embedding vectors). The rationale behind proposing to use, *e.g.*, 64D embedding vectors, is that the MOT task performs just a few one-to-one matchings between two consecutive frames. However, there is a tacit assumption of having all objects sufficiently visible, as this work did not cover occlusion handling and their benchmark evaluations were not specifically targeted at full object occlusion. This is a path definitely worth exploring. Nonetheless, the proposed framework achieved SOTA performance among MOT approaches, surpassing the competitive frameworks by a large margin even in terms of FPS.

6.4 Siamese Multi-Object Tracking and Attention

6.4.1 Motivation

During several evaluation runs and our manual inspection of the tracker performance, we noticed a ubiquitous pattern. We remind that the scenes on which we trained as well as tested our tracker were captured by a static camera. Consequently, several video sequences contained multiple vehicles standing still, due to a traffic jam or an ongoing



Fig. 6.12: An example of a situation where multiple vehicles are standing still on a cross-road. In this scenario, even though just a slight degree of occlusion is present, the biggest issues are caused by the need to delineate ROIs using axis-aligned BBOXes. This inevitably captures the neighboring vehicles, increasing the likelihood of drifting to semantic background due to presence of similar interference, *i.e.*, distractors.

red light, but viewed under an angle somewhere in the range of 30 – 60 degrees (Fig. ??). Therefore, it resulted in a partial occlusion. However, what we considered even more problematic was the inability of the axis-aligned BBOX to properly define the vehicle as the angle under which the car was visible caused the BBOX to capture a great portion of the neighboring vehicles even without severe occlusion happening.

Situations described above reminded us of the SiamMask [?] single object tracker targeted at predicting segmentation mask along with the usual single-object Siamese tracking routine. Such prediction was subsequently exploited to produce a rotated BBOX instead of an axis-aligned one. Even though the evaluation benchmarks only consider axis-aligned predictions, the rotated region served the purpose of enhancing the discriminative power of the tracker, primarily when dealing with occlusion. In the scenario shown in Fig. ??, a rotated BBOX would inexorably lead to an improved tracking accuracy. This approach was deemed successful for general object tracking, thus it also spawned another follow-up work of SiamMask-E [?] which altered the original formulation of predicting the rotated BBOX by use of ellipse fitting for even better accuracy.

However, there is a lack of datasets providing rotated annotations. The UA-DETRAC dataset is no exception. As a result, we sidestepped this approach and searched for an alternative solution that would enhance the discriminative power of the tracker when faced with partial occlusion. One such approach was the use of attention [?], especially spatial attention, which we found effective during our survey research [?]. Apart from the attention mechanism, we also remembered the more general formulation of the convolution operation, which has been shown to significantly better object detection tasks due to the semi-dense prediction requirements, dubbed as deformable convolution [?]. In what

follows, we will discuss these two methods (Section ?? and Section ??) as a foundation for our subsequent experiments that yielded a positive outcome.

6.4.2 Attention

An attention mechanism was first introduced by Vaswani *et al.* [?]. The use of encoder-decoder architectures to capture a complete sequence of information by a single vector spurred the development of the attention module. This use case poses problems in holding on to information at the beginning of the sequence and encoding long-range dependencies. To address this, the attention module computes attentions, or, in other words, the degree of relevance between “queries” and “keys”, to retrieve “values” in adequate proportions.

The concept of “queries, keys and values” comes from information retrieval systems. Let us provide a demonstrative example based on a YouTube video search. Assume a specific query signaling the demand to retrieve a particular YouTube video. The system then maps this query against a set of keys represented by various features, *e.g.*, video title, description, upload time, etc. These keys are directly associated with the stored candidate videos within the database. The output of this operation is a set of values, *i.e.*, found videos, that best match the given query.

The attention aims to exploit deep learning to learn a transformation of the input (not necessarily the same) into three separate vector spaces, each of them dedicated to a different purpose. The first space is to capture the query, therefore, it should represent features that best describe the query to facilitate information retrieval. The obvious compatriot is the key vector space which is trained to represent the value in the most accurate way to initiate the search accurately. Last but not least, the value vector space extracts features that are most useful for the task at hand. They do not need to capture features pertinent to the search. For that, there are the two other mappings.

For a more concrete demonstration, we will use a scaled dot-product attention. The input consists of queries and keys of dimension d_k , and values of dimension d_v . The query is used to compute a dot product with all the keys. These computations are scaled by $\sqrt{d_k}$ to provide a temperature scaling for the following softmax transformation to obtain the weights that will be used to retrieve values (Fig. ??). For optimal performance, it is reasonable to compute the attention function for the set of queries simultaneously as they can be easily stored in a matrix, denoted by **Q**. Analogically, keys and values can be also

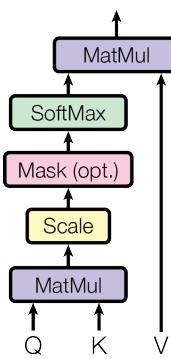


Fig. 6.13: An example of the input transformation by the scaled dot-product attention module. The pair of queries and keys is used to produce the probability distribution over the individual values for the final weighted sum. (source: [?])

packed together into matrices given by \mathbf{K} and \mathbf{V} , respectively. Thus, the attention can be formulated as a function of queries, keys and values, and is defined as

$$\text{attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}. \quad (6.14)$$

The two most prominent variants of attention are the additive attention [?] and the multiplicative (dot-product) attention, with the latter being identical to the one described above except for the temperature scaling. Just for the record, we experimented with both approaches and we observed differences in performance. On balance, both attentions are similar in theory, however, dot-product is much faster and more space-efficient in practice. On the other hand, additive attention outperforms the dot-product attention as long as temperature scaling is not employed for larger values of d_k , since the dot-products tend to push the softmax function to regions of extremely small gradients.

In our work, we also exploited the notion of self-attention. Since attention was first targeted at natural language translation, let us provide an example from this area. Originally, the attention was computed between the input and output sentences. Regarding self-attention, attention is computed with respect to the sentence itself. In terms of computer vision, the spatial self-attention represents a weight map over a 2D feature map indicating how important each feature element for the particular task is. Analogically, the channel self-attention may be used to attribute importance to individual channels, as they often are not equally important. Moreover, it yields more interpretable models as a by-product [?]. These ideas will be explored later.

6.4.3 Deformable Convolutional Neural Networks

Deformable Convolutional Neural Networks (DCNNs) [?] are gaining popularity and are being applied to numerous sophisticated computer vision tasks, *e.g.*, object segmentation (dense predictions) and object detection (semi-dense predictions). Since object tracking revolves around the same requirements in terms of pixel-wise precision, we contemplated using this advancement, too.

Although CNNs (Section ?? on page ??) are an excellent tool for a plethora of deep learning tasks involving image processing, they are still limited in their capabilities to model a broad range geometric transformation. To address this, practitioners apply a broad range of data augmentation techniques (*e.g.*, rotation, translation, scaling, shearing, and cropping) to provide the necessary samples of some particular transformation during the training. However, such an approach is limited to tailor-made transformations that may not cover the entire set of possibilities the model may face at test time.

The first work to learn spatial transformation from the training data in a deep learning fashion is known under the name Spatial Transform Networks (STNs) [?]. It warps the feature map via a global parametric transformation such as affine transformation. In the realm of convolutional operations, there is the atrous convolution operation [?] that enhances the standard convolution by expanding the receptive field while maintaining the same number of parameters by use of greater offsets. However, these offsets are fixed. An obvious successor of this approach is the active convolution [?] that treats convolution offsets as learnable parameters instead of constants. But, in this setting, the learned offsets are shared across different spatial locations. Thus, the most general approach is to determine the offsets at each location independently and then proceed as usual. This is where deformable convolution (Fig. ??) comes into place.

In concrete terms, a 2D convolution consists of sampling using a regular offset grid \mathcal{R} defining the receptive field as well as dilation over the input features \mathbf{x} followed by the summation of the samples values weighted by \mathbf{w} . For example, a standard 3×3 convolution with dilation 1 would employ offsets given by

$$\mathcal{R} = \{(-1, -1), (-1, 0), \dots, (0, 1), (1, 1)\}. \quad (6.15)$$

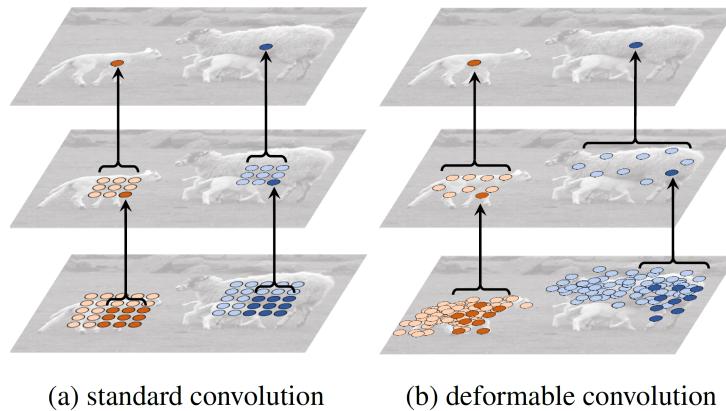


Fig. 6.14: Visualization of the difference between the fixed **(a)** and adaptive **(b)** receptive fields. Stacking multiple deformable convolutions results in profound amplification of deformation, making the transformation capture diverse shapes that would otherwise be very coarsely approximated by a standard convolution. (*source: [?]*)

Then, for each location \mathbf{p}_0 within the output feature map \mathbf{y} is calculated as

$$\mathbf{y}(\mathbf{p}_0) = \sum_{\forall \mathbf{p}_n \in \mathcal{R}} \mathbf{w}(\mathbf{p}_n) \cdot \mathbf{x}(\mathbf{p}_0 + \mathbf{p}_n), \quad (6.16)$$

where the locations in \mathcal{R} are iterated over by \mathbf{p}_n . Conversely, the deformable convolution extends the standard one by augmenting the original sampling grid \mathcal{R} with additional offsets $\{\Delta\mathbf{p}_n \mid n = 1, \dots, |\mathcal{R}|\}$ (Fig. ??). Thus, Equation ?? is reformulated as

$$\mathbf{y}(\mathbf{p}_0) = \sum_{\forall \mathbf{p}_n \in \mathcal{R}} \mathbf{w}(\mathbf{p}_n) \cdot \mathbf{x}(\mathbf{p}_0 + \mathbf{p}_n + \Delta\mathbf{p}_n). \quad (6.17)$$

Nonetheless, the user needs to keep in mind that the sampling offsets now become fractions and thus have to be handled accordingly. One approach is to employ bilinear interpolation, where the position in the input feature map \mathbf{x} is determined by

$$\mathbf{x}(\mathbf{p}) = \sum_{\forall \mathbf{q}} G(\mathbf{q}, \mathbf{p}) \cdot \mathbf{x}(\mathbf{p}), \quad (6.18)$$

in which \mathbf{q} enumerates all integral locations and $G(\cdot)$ represents the interpolation kernel. The interpolation processing can be efficiently implemented owing to the sparsity. The performance overhead is negligible compared to the reaped benefits of adaptive sampling locations capable of covering very complicated transformations (Fig. ??).

The original paper [?], in which DCNNs were introduced, showed that learning dense spatial transformation in using deep learning by use of CNNs or sophisticated vision tasks

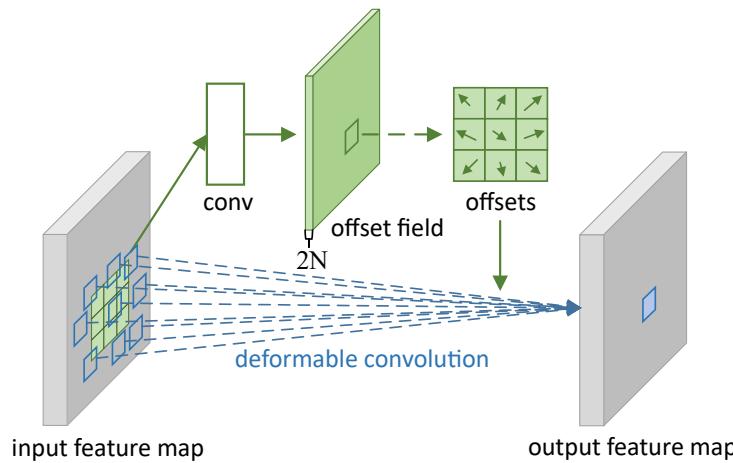


Fig. 6.15: Illustration of a 3×3 deformable convolution operation. Unlike the standard convolution operation used in neural networks, this one employs one additional step of predicting variable offsets instead of using a fixed rectangular grid. (*source: [?]*)

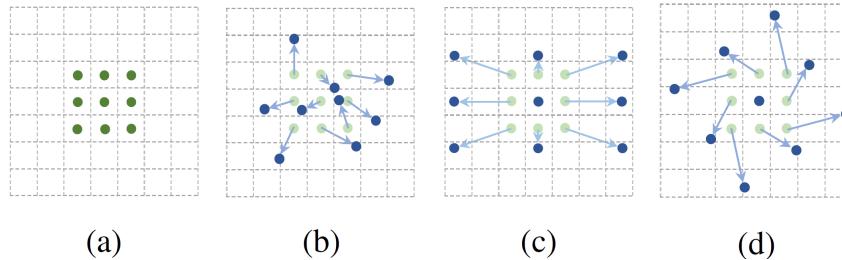


Fig. 6.16: Deformable convolution is effective at learning appropriate sampling locations reflecting the underlying transformation. (a) shows the regular sampling grid of a standard convolution; (b) is an example of irregularly deformed sampling region; (c) and (d) represent an expected pattern corresponding to scaling and rotation operations, respectively. (*source: [?]*)

such as object detection and semantic segmentation is feasible as well as effective. After our experience, we add that object tracking may benefit from this extension, too.

6.4.4 Modulated Deformable Convolutional Neural Networks

The original paper by Zhu *et al.* [?] aptly described their contribution in the headline that contained the words “mode deformable, better results”. Because of this, we will provide a concise description of the Modulated Deformable Convolutional Neural Networks (MDCNNs), an extension to the previously discussed DCNNs.

Since we are simply adding a slight modifications to an already introduced equation, we will try to avoid repetition. Thus, let \mathbf{p}_0 , \mathbf{p}_n and $\Delta\mathbf{p}_n$ have the same meaning as in Equation ???. Then, the modified equation becomes

$$\mathbf{y}(\mathbf{p}_0) = \sum_{\forall \mathbf{p}_n \in \mathcal{R}} \mathbf{w}(\mathbf{p}_n) \cdot \mathbf{x}(\mathbf{p}_0 + \mathbf{p}_n + \Delta\mathbf{p}_n) \cdot \Delta\mathbf{m}_n, \quad (6.19)$$

where $\Delta\mathbf{m}_n$ is the modulation scalar for the current location, such that $\Delta\mathbf{m}_n \in (0, 1)$. Thus, there are two types of learnable parameters. The already described offsets, given by the $\Delta\mathbf{p}_n$ term, and the new modulation (weighting) coefficients, represented by the term $\Delta\mathbf{m}_n$. This trivial extension allows the system to not only learn how to sample features in a non-regular fashion if needed, but it also allows applying distinct weight to each sampling location to further adaptively intensify the deforming effect.

From an implementation standpoint, DCNNs as well as MDCNNs have learnable offsets (and modulation coefficients, if used) set to zero during initialization. This produces no deformable effect, so the convolution behaves as usual in terms of location sampling. However, the modulation aspect is slightly different. Since the sigmoid function is commonly adopted to project the modulation weights into the $(0, 1)$ interval, it multiplies each location by the value of $\text{sigmoid}(0) = 1/2$ at the beginning.

6.4.5 Deformable Siamese Attention

The two independent ideas above led us to experiment with a self-attention mechanism aimed at enhancing feature selection in both spatial and channel domain. Such experiments resulted in slight improvements for the reasons outlined in the motivation section. To support that our ideas were based on properly identified causes, there is a recently published work demonstrating the effectiveness of the very same approach.

Yu *et al.* [?] formulated their Deformable Siamese Attention (DSA), which covered both of our suggestions above and additionally introduced the notion of cross-attention as an enhancement to the self-attention itself. What primarily motivated the introduction of the cross-attention was that the exemplar and search region features in Siamese trackers are computed separately, yet they may frequently compensate each other. It is reasonable to assume that multiple objects appear at the same time even in SOT, let alone MOT. Consequently, it is of paramount importance for the search branch to have as much information as possible about the exemplar during the computation of the response map for better discrimination. By the same token, the exemplar features may be enhanced by information from the search features. To this end, the cross-attention, at the acceptable computational cost, serves properly in a predictable fashion.

Considering their contribution and promising outcomes for the SOT demonstrated on the SiamRPN framework (Section ?? on page ??), we decided to implement their proposed

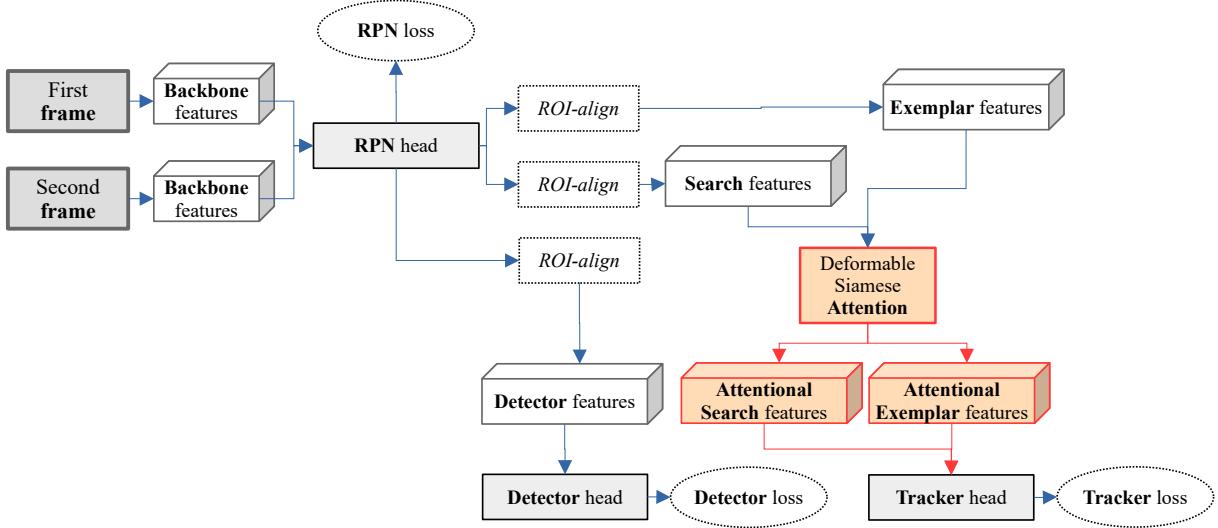


Fig. 6.17: Our proposal to incorporate attention into the SiamMOT pipeline. This diagram shows the relationships between individual parts of the framework during the training phase.

module into the SiamMOT tracker as described in their paper (Fig. ??). However, with the prospect of greater improvement, we adopted DCNNs, instead of pure DCNNs, because the MDCNNs have all the advantages of the standard deformable convolution, but they additionally learn a modulation (weighting) for individual elements of the feature map while taking the underlying features into account. For our purposes, this seemed to intensify the spatial attention effect, since not only the deformable part was responsible for choosing features using irregular sampling patterns, the network was even allowed to weigh them differently. We conjectured that such an extension may either have no dramatic effect or influence it only positively.

Self-Attention

Self-attention is computed separately on the exemplar and search branch independently. This operation can be easily achieved since exemplar and search tensors only differ in width and height. The following description of the self-attention computation conforms to the established attention principles regarding “queries, keys and values” introduced in Section ???. For better understanding of the computation, see the diagram in Fig. ??.

Let $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$ be the input features. To produce query features \mathbf{Q} and key features \mathbf{K} , such that $\mathbf{Q}, \mathbf{K} \in \mathbb{R}^{C' \times H \times W}$ and $C' = \frac{1}{4}C$, where C' is the reduced number of channels, two separate 1×1 convolution layers are applied. The two obtained features then need to be reshaped into $\bar{\mathbf{Q}}, \bar{\mathbf{K}} \in \mathbb{R}^{C' \times N}$, where $N = H \times W$. The spatial self-attention $\mathbf{A}_S^S \in$

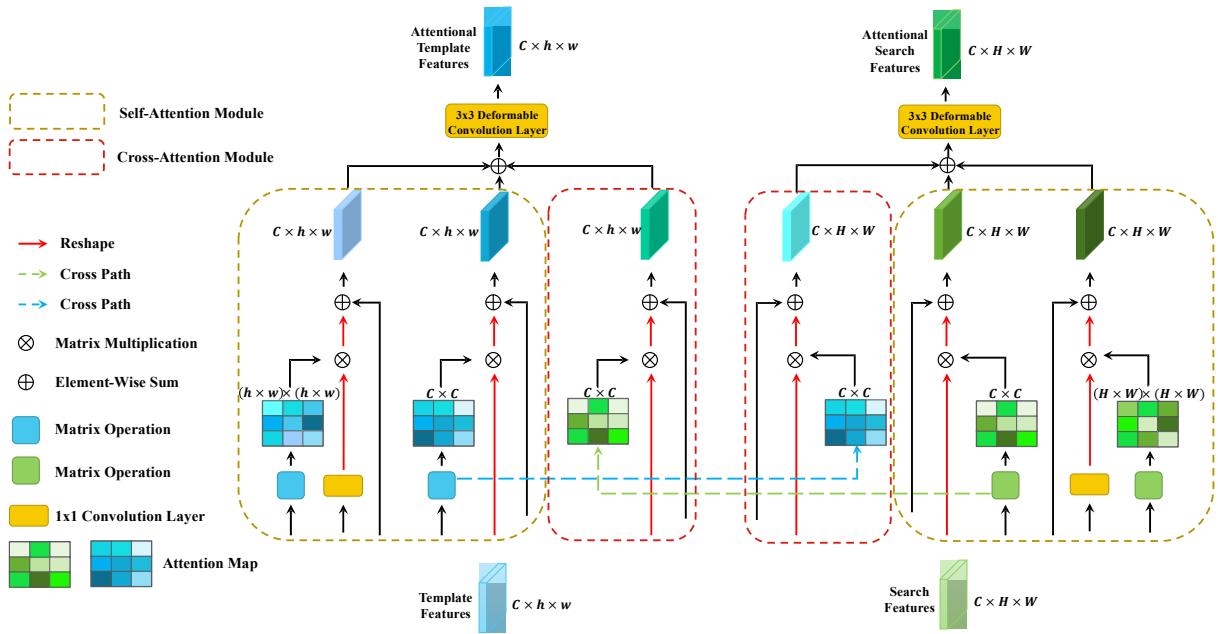


Fig. 6.18: The DSA extension introduces two sub-modules for both exemplar and search branch. The self-attention is further divided into two operations, namely the spatial and channel attention. The very same attention network is used to process both features independently. Notice how the channel attention is computed only once as part of the self-attention process and then directly fused with the channel self-attention of the other branch, creating the cross-attention effect, which is the strongest one from all three, according to authors. (source: [?])

$\mathbb{R}^{N \times N}$ is then produced via matrix multiplication and a column-wise softmax operation as

$$\mathbf{A}_S^S = \text{softmax}_{\text{col}}(\bar{\mathbf{Q}}^T \bar{\mathbf{K}}). \quad (6.20)$$

Authors used $C' = \frac{1}{8}C$, but in MOT, the number of objects to track is often a lot greater, thus the computation graph grows dramatically with the higher number of channels. Furthermore, in our case $C = 128$, and we considered using 32 channels for the attention to be the bare minimum.

Meanwhile, an analogous sequence of operations is adopted to produce the value features. A 1×1 convolution layer without the subsequent reshape operation transforms the input features \mathbf{X} into the value features $\bar{\mathbf{V}} \in \mathbb{R}^{C \times N}$. At this point, we have matched the queries with keys and computed the values. We may proceed further to the weighted selection from the values and to incorporate the attention into the features as follows

$$\bar{\mathbf{X}}_S^S = \alpha \bar{\mathbf{V}} \mathbf{A}_S^S + \bar{\mathbf{X}}, \quad (6.21)$$

where α is a learnable scalar parameter, and $\bar{\mathbf{X}}_S^S \in \mathbb{R}^{C \times N}$. The outputs $\bar{\mathbf{X}}_S^S$ are then

reshaped back to the original size, specifically $\mathbf{X}_S^S \in \mathbb{R}^{C \times H \times W}$. From our experience, the parameter α is very useful for training stabilization.

The corresponding channel self-attention \mathbf{A}_C^S and the channel-wise attentional features \mathbf{X}_C^S are obtained similarly. Due to space limitations and the fact that the upcoming formulation of the cross-attention exploits the channel self-attention, we will omit a detailed description. We will just point out that the “queries, keys and values” for the channel self-attention are produced directly from the features on the input, with no 1×1 convolutions whatsoever. The final self-attentional features are generated by an element-wise sum using the partial spatial and channel self-attentions, \mathbf{X}_S^S and \mathbf{X}_C^S , respectively.

Cross-Attention

Let $\mathbf{Z} \in \mathbb{R}^{C \times h \times w}$, $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$ denote the exemplar and search region features, respectively. The following description introduces the computation of the cross-attention from the perspective of the search branch. First, the exemplar features \mathbf{Z} are reshaped into $\bar{\mathbf{Z}} \in \mathbb{R}^{C \times n}$, where $n = h \times w$. Then, the cross-attention from the exemplar branch is computed. We emphasize that the channel attention is reused, therefore, the computation below serves as a recipe for how to compute the channel self-attention. So, we compute the channel cross-attention $\mathbf{A}^C \in \mathbb{R}^{C \times C}$ as

$$\mathbf{A}^C = \text{softmax}_{\text{row}} (\bar{\mathbf{Z}} \bar{\mathbf{Z}}^T). \quad (6.22)$$

The real benefit comes from the merging stage, where the above-computed attention is merged with the other, in this case, the search branch as

$$\bar{\mathbf{X}}^C = \gamma \mathbf{A}^C \bar{\mathbf{X}} + \bar{\mathbf{X}}, \quad (6.23)$$

where γ is a learnable scalar parameter. The merged features $\bar{\mathbf{X}}^C$, once again, have to be reshaped, so the features $\mathbf{X}^C \in \mathbb{R}^{C \times H \times W}$ are the final output.

At last, the self-attentional features are combined with the cross-attentional features using element-wise sum. The cross-attention from the perspective of the exemplar branch can be obtained using a similar sequence of operations. In total, there are six steps that involve addition for the purpose of feature merging (Fig. ??).

Once the attention is applied to tracking, the corresponding response map is altered as

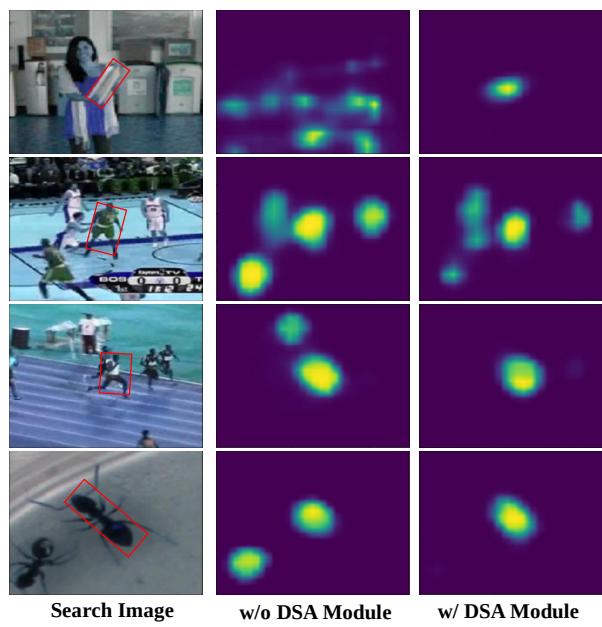


Fig. 6.19: Visualization of response (confidence) maps. The first column represents the search image, the second column represents the activation levels without the DSA module, whereas the third column clearly demonstrates the improved target-background discriminability in the computed attentional features. (*source: [?]*)

expected. Discriminative power of the tracker is enhanced by appropriate suppression of the semantic background. As Fig. ?? shows, activations in the search regions as viewed through the response map vary if the DSA module is included in the computation, making the tracker less prone to drifting to scene or semantic background objects.

Deformable Convolution Phase

The attention is finalized by processing the obtained feature tensors by another layer of modulated deformable convolution. We modified the deformable convolution operation to include the modulated version, which is our modification compared to the original formulation (Fig. ??, top yellow boxes). The resulting features with shape identical to the input shape are used to compute the response map. As a result, this extension can be easily integrated into an existing pipeline thanks to its ability to preserve tensor shapes.

6.4.6 Experimental Evaluation

Add experimental evaluation. Don't forget about gradient accumulation.

6.4.7 Discussion

Add discussion.

Chapter 7

Discussion and Conclusion

7.1 Discussion

Add general discussion regarding the achieved results.

7.2 Conclusion

Finish conclusion.

The main objective of this thesis was to contribute to the field of visual object tracking. Specifically, we aimed at improving a SOTA multi-object tracker of our choice, the SiamMOT framework, for traffic analysis. The core of our contribution stands at enhancing the discriminative ability of the tracker to handle scenes where objects of interest become occluded. In terms of methodology, we heavily relied on deep machine learning and Siamese neural networks, which the SiamMOT tracker is built upon.

At the beginning of this write-up, we introduced several concepts regarding object tracking, primarily when dealing with visual input in the form of a video. We have identified a plethora of approaches, but the most promising seems to be Siamese-based fully convolutional trackers.

Siamese neural networks form the basis of the leading branch of trackers exploiting the properties of similarity learning. The approach of similarity learning facilitates the creation of metric spaces with specific traits. Without loss of generality, these traits are defined during the training by the data that is fed to the neural network model. The aim

is to create a space where a trivial distance measurement between feature vectors of the embedded objects reflects their similarity. Such a similarity measure should be invariant to various distortions in lightning and object position, as well as occlusion of varying severity. Still, it should capture enough information to accurately indicate whether two objects are the same or not. Among multiple problems hindering the performance of object tracking, we have identified the occlusion as the one we would focus on.

For the most part, we surveyed general object trackers, yet we intend to apply our work to tracking vehicles. Even though we also researched vehicle tracking, the relative lack of detailed elaboration is evidence of the shortage of published research in that area. More than once when we mentioned that even face ReID had been explored a lot more than vehicle ReID. Taking this into consideration, we would venture to claim that vehicle tracking is still a largely unexplored area, too. We remark that it is an important area with a vast practical impact.

Investigation of general principles of object tracking has provided a foundation for what we can expect from the currently best object trackers. Examination of the concept of learning metric spaces, in particular, object ReID, offered insight into the possibilities of task-specific embeddings. At the same time, occlusion is tightly coupled with similar interference. To prevent the tracker from drifting to the background (whether semantic or not), mechanisms based on attention have shown promising results. Not only does our survey [?] cover such approaches explicitly in one section, our most relevant contribution to object tracking exploits the attention as well.

Appendix

List of Author's Publications

Home Conferences

AFD001 Foundations for homography estimation in presence of redundant point correspondencies / Milan Ondrašovič, Peter Tarábek.

In: **Mathematics in science and technologies: proceedings of the MIST conference 2020: proceedings of the MIST conference 2020 / Katarína Bachratá, Katarína Jasenčáková and Monika Smiešková.** - 1. vyd. - [S.l.] : [s.n.], 2020. - 70 s. [print]. - ISBN 9798648566026. - s. 52-57 [print].

[Ondrašovič Milan (50%) - Tarábek Peter (50%)]

AFD002 Object position estimation from a single moving camera / Milan Ondrašovič, Peter Tarábek, Ondrej Šuch.

In: **Information and digital technologies 2021: proceedings of the international conference: proceedings of the international conference / [bez zastavovateľa].** - 1. vyd. - Danvers : Institute of Electrical and Electronics Engineers, 2021. - 370 s. - ISBN 978-1-6654-3692-2. - s. 31-37. Zaradené v: SCOPUS

[Ondrašovič Milan (50%) - Tarábek Peter (40%) - Šuch Ondrej (10%)]

International Impacted Journals

ADC001 Homography ranking based on multiple groups of point correspondences / Milan Ondrašovič and Peter Tarábek.

In: **Sensors.** - Bazilej: Multidisciplinary Digital Publishing Institute. - [online, print]. - ISSN 1424-3210. - Roč. 21, č. 17 (2021), s. [1-17] [online, print]. Zaradené v: Current Content Connect ; SCOPUS ; Web of Science Core Collection

[Ondrašovič Milan (50%) - Tarábek Peter (50%)]

ADC002 Siamese visual object tracking: a survey / Milan Ondrašovič and Peter Tarábek.

In: **IEEE Access : practical innovations, open solutions: practical innovations, open solutions.** - Piscataway: Institute of Electrical and Electronics Engineers. - [online]. -

ISSN 2169-3536 (online). - Roč. 9 (2021), s. 110149-110172 [online]. Zaradené v: Current Content Connect ; SCOPUS ; Web of Science Core Collection
[Ondrašovič Milan (50%) - Tarábek Peter (50%)]

There are already **three existing international citations** for this paper:

1. *Zhou, Dong, Gunaghui Sun, and Xiaopeng Hong.* **3D Visual Tracking Framework with Deep Learning for Asteroid Exploration.** arXiv (2021).
2. *Sun, Xinglong, Guangliang Han, and Lihong Guo.* **Siamese Visual Tracking with Residual Fusion Learning.** IEEE Access (2021).
3. *Choi, Janghoon, et al.* **Visual Tracking by Adaptive Continual Meta-Learning.** IEEE Access (2022).