UNIVERSITY OF ILINA
FACULTY OF MANAGEMENT SCIENCE AND INFORMATICS

OBJECT TRACKING IN VIDEO

Dissertation Thesis

Study Program:	Applied Informatics
Field of Study:	Informatics
Workplace:	Department of Mathematical Methods and Operations Research
	Faculty of Management Science and Informatics, University of Žilina
Supervisor (title, name):	doc. Mgr. Phd., Ondrej Šuch; Ing. PhD., Peter Tarábek

Žilina, 2022 **Ing., Milan Ondrašovič**

Declaration

I sincerely declare that the thesis *Object Tracking In Video* 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

I would like to acknowledge the significant contribution of both my supervisors for their scrupulous cooperation and guidance throughout my studies and especially in my research. For such a dedication I am deeply grateful.

Abstract

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

Abstract

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 and has consistently been a popular research area over the last two decades. Briefly speaking, the objective is to locate a certain object in all frames of a video, given only its location in the first frame. From a technical perspective, an object is firstly detected in the image (frame), a unique identifier is assigned to it, and subsequently, and then the same identifier has to be correctly assigned if the object is present in future images from the scene. Tracking can also be considered a task of estimating an objects trajectory throughout a sequence of images, such as video frames.

VOT is a task at which people perform reasonably well, but when it comes to computers, it is considered practically unsolved. With this in mind, object tracking is the task of following one or more objects in a scene, from their first appearance to their exit [1]. 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 after which the object re-emerges in the scene again [2].

There is extensive literature covering different approaches. Since 2013, there has even been a standard benchmark in the field, called the VOT Challenge [3], which maintains datasets that researchers and individuals can use to benchmark their algorithms. Another competitive type of benchmark exists under the name Object Tracking Benchmark (OTB) [4]. Additionally, a similar type of evaluation and comparison of numerous approaches for the Multi-Object Tracking (MOT) exists since 2014 under the name MOT Challenge [5], too. Such an extensive endeavor to push the boundaries of object tracking performance supports the relevancy of this topic.

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, an ideal tracking algorithm, should as a result have the properties below [2]:

- 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* [2]. The top-down methodology employs *forward tracking*, which means that an attempt is made at every frame to locate the object given some initial hypothesis [6]. On the other hand, the bottom-up approach utilizes the object detection first and then the adequate correspondence is established with previously detected objects [7]. We plan to analyze and exploit the capabilities of the SOTA object detectors to localize objects of interest (primarily vehicles in our use case) and then build the object tracking pipeline.

The field of computer vision has provided diverse approaches to visual tracking in the past decades, *e.g.* using geometry and manual feature-engineering in combination with Kalman filtering [8] for predicting future states. The Hungarian algorithm [9] is the most commonly used to solve the minimum cost assignment problem, as utilized by [10].

However, the most influential approaches were the ones involving the modern tools of machine learning, specifically deep machine learning. Deep learning can perform global generalization, which makes it a valuable tool as a function approximator, so the utilization is vast. This revolutionary idea which has in the past decade reaped an upsurge in its utility will be the key element of this thesis and we will devote the largest portion of our work to it. Krizhevsky *et al.* [11] showed that an outstanding tool, when it comes to the application of deep learning in computer vision, are Convolutional Neural Networks (CNNs) (Section 3.1.2). It is a predominant approach to extract valuable visual features from the pixel space of images.

Visual tracking of single or multiple objects is often just an intermediate step for various ends, such as traffic analysis [12], whether from a static camera or as part of self-driving cars, pedestrian detection and tracking [13], activity understanding [14], and imitation based on a video [15]. While this is an important problem with annually occurring challenges for the best achievement, the current SOTA solutions still lack high accuracy in unconstrained scenarios with potential object occlusion [16]. 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 [17]:

- *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).

In traffic scenarios, the last two types are prevalent.

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 identification of some object, or re-identification (ReID), is indispensable. Various advances in the creation of latent spaces and so-called *embeddings* using deep neural networks (Section 3.1.1) have shown promising results [18, 19]. 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 [20]. Later on, the Euclidean distance or cosine similarity between any two vectors results in a high degree of similarity if the two visual samples belong to the same object.

A broad range of real-life applications requires tracking multiple objects, which only adds complexity to an already tough problem itself. But [21] shows that approaching the problem of vehicle ReID using embeddings (Section 3.3) trained with contrastive or triplet loss (Section 3.3.3) 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 using tools of deep learning, primarily CNNs. Given the potential for improvement, considering the performance of SOTA approaches as well as the practical demand for an accurate tracking outcome, be it traffic or other scenarios [12], we think that an emphasis should be put on occlusion (partial or full) handling, as it causes major difficulties for existing methods [16].

The secondary objective is to extend the current knowledge in the field of computer vision and deep machine 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` [22] provide only tracking algorithms for single objects, as opposed to MOT, which is demanded in practice, yet concrete solutions exist but are not ubiquitous and easy to use. Visual tracking in the presence of occlusion is, according to our belief, coupled with ReID. As recent work of [21] suggests, for instance, identification of faces has been researched a lot deeper than vehicles. The two tasks are similar, yet the common conviction is that despite the intra-class variations, the identity of a vehicle is less fine-grained compared to other objects, *e.g.* people [21]. Multiple papers, such as [18, 23], show promising results in the training of machine learning models when using embeddings, specifically when applying the triplet loss function referred to earlier.

As the key element of our approach will undoubtedly be deep learning, an adequate dataset will be imperative, whether newly created or re-used, and modified accordingly. Because of this, we could contribute to the current base of datasets available online with our work, too.

The rest of the document is organized as follows. The main goals of this dissertation thesis are described in Chapter 2, discussed next. To equip the reader with the necessary foundational knowledge we provide Chapter 3. Right after this chapter, we composed a short treatise on available and used datasets in Chapter 4. At this point, we start with our first relevant experiment related to homography transformations, to which an

entire Chapter 5 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 in a journal. Then, we have Chapter 6, focused on our developed approaches to visual object tracking. It is by far the most important chapter and presents the greatest part of our work. Our achieved results are discussed subsequently in Chapter 7, which closes the entire document with a conclusion.

Chapter 2

Dissertation Thesis Goals

VOT is increasingly studied in recent years due to its real-world applications. Given the elaboration provided in Chapter 3 that follows, we believe that investing research and development time into the topic of object tracking is worthwhile. The results presented so far are by no means everything that could be said about this topic. Quite to the contrary, 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 [24], segmentation masks [25], or even object contours [26]. We plan to primarily focus on solutions producing axis-aligned BBOXes since they are ubiquitous. Another variation is the speed at which the tracker processes the input. We have mentioned several times 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 ilina. 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 [27, 28, 25] 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 ReID (Section 3.3.1). 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 [29]. Our research of modern publications brought forward in the previous chapters points to the direction of embeddings and metric learning (Section 3.3), specifically when applied to ReID. Besides the possibility to employ motion models for better prediction of locations in the absence of visual input, some research papers indicate 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 [30]. 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. Variation in lightning should not dramatically influence the outcome. 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 [29]. In those situations, it is crucial to exploit temporal information. 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 ReID and similarity learning to support the solution to problems caused by the object occlusion of varying intensity, change in the position, and the viewpoint of the tracked object, and fluctuations in the scene illumination**.

In order to accomplish the aforementioned goals, this dissertation thesis should follow the steps given below:

- Identification of major problems that occur in the task of visual object tracking.
- Outline of requirements for datasets for training and evaluation of models along with a possibility of comparison with other approaches.
- Adequate data preprocessing, either as a modification of existing datasets or creating new ones for object tracking, particularly in traffic-related scenarios.
- Analysis of the current SOTA approaches dealing with visual tracking of objects.
- Review of the modern methods of ReID based on similarity learning. Assessment of the contrastive-based and triplet-based training paradigms.
- The exploitation of recent advances in object ReID to handle object occlusion and dynamic conditions in terms of illuminance and object position.
- Design and implementation of a method for visual object tracking adopting advances in deep machine learning.

- Evaluation of the implemented method with respect to official benchmarks.
- Summarization and dissemination of the achieved results followed by recommendations for their practical applications.

Chapter 3

Theoretical Foundations

3.1 Neural Networks

In this short section, we briefly describe artificial neural networks and their various types relevant to this thesis. However, our assumption is that the reader is sufficiently equipped with the knowledge regarding deep machine learning, and thus, for the sake of keeping this document as short as possible, we will omit detailed elaboration.

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. Within recent decades, under the influence of constantly growing computational power and our data collecting and data processing capabilities, this technology has acquired a status of paramount importance when solving a problem using machine learning. They are a quintessential deep learning model. Artificial neural networks are also known in the literature as the multilayer perceptron [31], feedforward neural networks or deep feedforward neural networks [32].

In mathematical terms, the goal of a neural network is to approximate some unknown function f. For instance, when considering a classier, 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 [32].

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 nally to the output $\mathbf{y}$. Moreover, the feedforward neural networks are called networks since they are typically represented as a composition of multiple different functions. 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, $f(\mathbf{x}) = f^{(3)}\left(f^{(2)}\left(f^{(1)}(\mathbf{x})\right)\right)$. These chain structures are the most commonly used structures in neural network models [32] (see Fig. 3.1). Deep machine learning consists of multiple such layers of neurons stacked onto each other that are trained using the

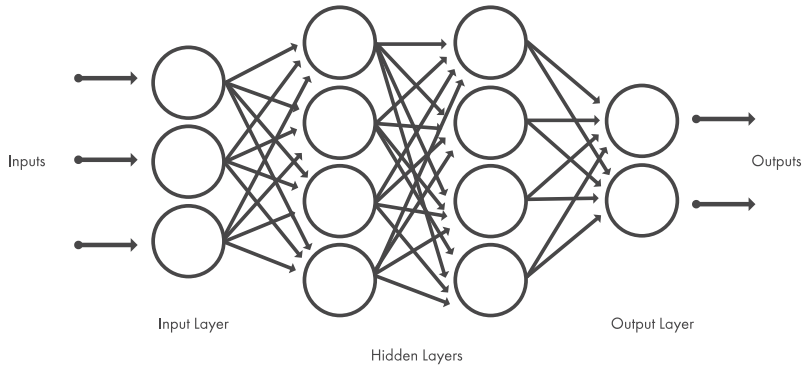


Fig. 3.1: An illustration of a typical fully connected neural network. (*source*: [34])

backpropagation algorithm [33]. Our work will rely heavily upon the aforementioned deep neural networks, especially convolutional neural networks, discussed next.

3.1.2 Convolutional Neural Networks

This type of neural networks has gained popularity in the computer vision community thanks to a never-seen-before performance on image classification task [11]. This machine learning algorithm processes images or other high dimensional, grid-like input and then learns the importance (weights and biases) of various aspects of the input data. Even though CNNs can handle $1D$ time series data and be a counterpart to Recurrent Neural Networks (RNNs) [35], our primary focus is to process $3D$ tensors. Simply put, convolutional networks are artificial neural networks that use in at least one of their layers the convolution operation instead of a general matrix multiplication [32] (see Fig. 3.2).

The successful ability of these networks to capture spatial and temporal 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) [32]:

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

In this setting, f is be 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 (similar to general artificial neural networks) 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 convolution can be written in its discrete form as

$$(f \star g)(i, j) = \sum_{m} \sum_{n} I(m, n)\, K(i-m, j-n). \tag{3.2}$$

Nonetheless, 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 commutative property is not preserved, which is not demanded in neural networks. In that case, the result of the cross-correlation

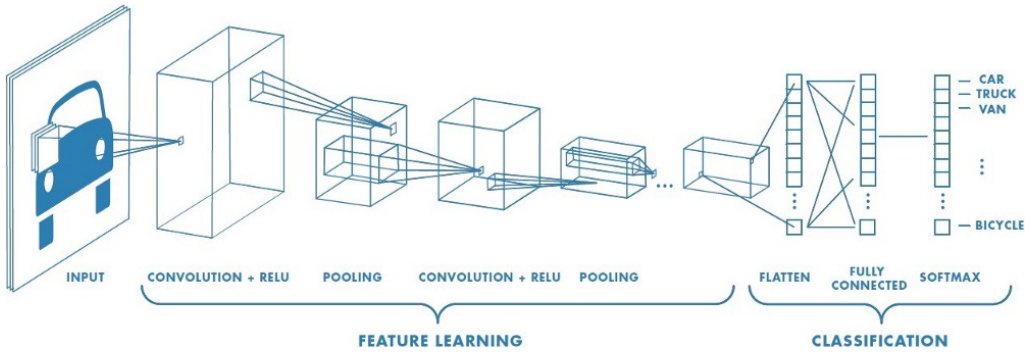


Fig. 3.2: An illustration of a typical convolutional neural network architecture for a quintessential classification task. An image of a car is fed into the model that extracts visual features that are subsequently processed by the fully connected layers resulting in the probability distribution over target classes. (*source*: [37])

is given by [32]

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

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 as well as properly generalize. The exploitation of the mentioned qualities of convolutional layers will be demonstrated in Section 3.6.

Another prominent use case of CNNs is *transfer learning*, where a pretrained model is adopted for a new task, utilizing the already learned features. These pretrained models may come in various flavors, but typical ones are pretrained for an image classification task using the `ImageNet` dataset [36]. 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 also additional *fine-tuning* of the pretrained model of choice.

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. Here we give an overview of various types of image inference to highlight the importance of object detection. The general task of image inference comes in three types, where each builds upon the previous one. The dependency chain is the following: *object classification* → *object detection* → *image segmentation*.

Object detection comes in various flavors discussed in the following sections. From a

holistic point of view, there are *fast* object detectors, such as the most prominent one, You Look Only Once (YOLO) [38] (Section 3.2.2), or Single Shot Detector (SSD) [29]. 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 proposals, such as various versions of `Faster R-CNN` (Section 3.2.3).

Historically speaking, object detection in its early stages of development depended on feature engineering followed by some shallow machine learning model (*e.g.*, Support Vector Machine (SVM) [39]) to classify the proposed object. Feature extraction was achieved using approaches such as Scale-Invariant Feature Transform (SIFT) [**?**], an algorithm capable of producing description of local features in images. Another eminent approach was Histogram of Oriented Gradients (HOG) [40], equally used for feature extraction by taking advantage of a change in orientation of gradients in localized portions of the image and subsequently counting occurrences of various orientations. But, as we stated, robust object detection is what our task demands, and the goal of getting the top performance is unattainable without deep learning-based object detection, the principal topic of this 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 region proposals 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 and the latest results are promising [41]. We will also review fully convolutional trackers in great detail in Section 3.6.2.

3.2.1 Non-Maximum Suppression

Before we dive into the description of the detection itself, we first introduce a general post-processing approach for test phase inference of an object detector. Object detectors have hugely profited from the so-called end-to-end learning paradigm in which object proposals, features, and the classifier become part of one neural network model [42]. A proposal, in this setting, 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 of interest. Moreover, these proposals may have a large overlapping region as measured by intersection over union (IoU) (Section 3.14), rendering most of them useless in terms of conveying new information, as the majority could cover pretty much the same region. To filter such proposals and extract only the relevant ones, the Non-Maximum Suppression (NMS) algorithm introduced below is used (see Fig. 3.3). The central idea is to iteratively select only region proposals the IoU of which is below a specific threshold, *i.e.* the area they share does not exceed a particular value.

Fig. 3.3: 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.

For a more formal description, let $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\}$ be a set of n region proposals described by n BBOXes. Scores for each detection are contained in a set $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$, where s_i denotes a detection score for the i-th box, $\mathbf{b_i}$. Let λ, such that $0 \leq \lambda < 1$, denote a 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 the NMS algorithm below (inspired by [43]):

function NMS($\mathcal{B}$, $\mathcal{S}$, λ)
  $\mathcal{B}_{nms} \leftarrow \emptyset$ ▷ initialize the output (filtered) set of region proposals
  while $\mathcal{B} \neq \emptyset$ **do** ▷ loop until all the proposals are processed
    $m \leftarrow \underset{i \in \{1,2,\ldots,|\mathcal{S}|\}}{arg\,max}\ \mathcal{S}$ ▷ find an index of a proposal with the highest score
    $\mathcal{B} \leftarrow \mathcal{B} - \mathbf{b}_m$, $\mathcal{S} \leftarrow \mathcal{S} - s_m$ ▷ remove the proposal
    $\mathcal{B}_{nms} \leftarrow \mathcal{B}_{nms} \cup \mathbf{b}_m$ ▷ save the proposal with the highest score
    for $i \leftarrow 1$ to $|\mathcal{B}|$ **do** ▷ iterate through remaining proposals
      if IOU($\mathbf{b}_m$, $\mathbf{b}_i$) $\geq \lambda$ **then** ▷ IoU (Equation 3.14) exceeds the threshold
        $\mathcal{B} \leftarrow \mathcal{B} - \mathbf{b}_i$, $\mathcal{S} \leftarrow \mathcal{S} - s_i$ ▷ remove the proposal
      end if
    end for
  end while
  return $\mathcal{B}_{nms}$
end function

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 very accurate. Its speed is primarily a consequence that *it looks only once* at a given input image. Compared to region proposal-based approaches, where each object proposal has to be classified separately, resulting in multiple inferences by a potentially huge neural network model, the authors of YOLO devised a CNN model capable of performing extraction of region proposals as well as classification in a single run [38]. Besides, the backbone CNN model 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 propagation through the neural network is executed, 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 [38],

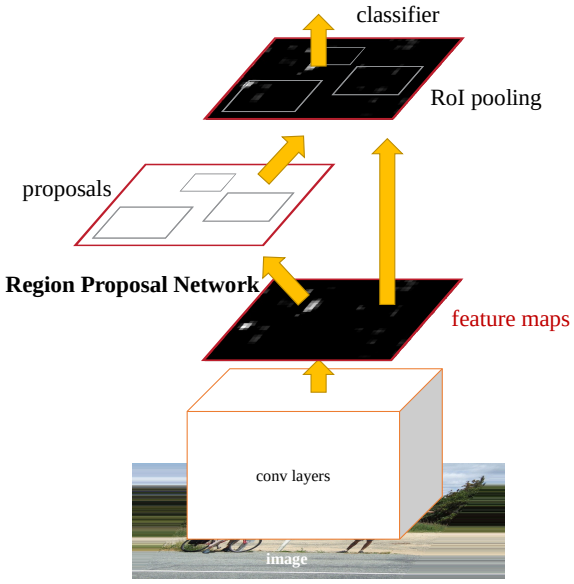


Fig. 3.4: The `Faster R-CNN` is built as a unified network for object detection, with its RPN module serving the purpose of an *attention*. (*source*: [48])

multiple updates have been brought forward, either from the main author himself [44, 45], or from other researchers [46, 47]. Such an interest supports the assertion that YOLO has great potential for further research as well as real-world applications. Despite its populality, for our experiments, we employed the paragon of two-stage object detection, discussed next.

3.2.3 Faster R-CNN

The `Faster R-CNN` [48] is the most prominent *two-stage* object detector. The first stage consists of generating region proposals using the Region Proposal Network (RPN) (see Fig. 3.4), where the input image is processed by a feature extractor [49]. 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) (see Fig. 3.5 and Fig. 3.6).

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 map which are then fed to the remaining feature extractor to predict the class. Based on this class prediction, the proposed box is further refined. The name for this stage is *box classifier*.

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.

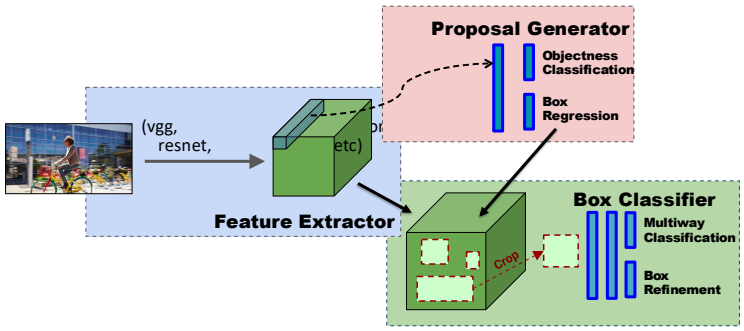


Fig. 3.5: A meta-architecture of the `Faster R-CNN` model. (*source*: [49])

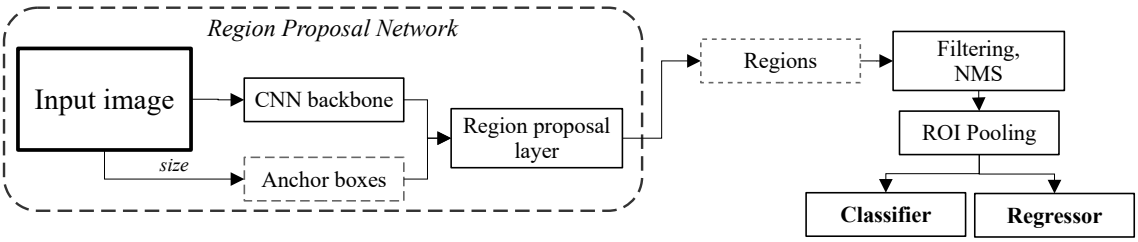


Fig. 3.6: A conceptual diagram of the `Faster R-CNN` model demonstrating the information pipeline when processing an input image.

3.3 Latent Spaces and Embeddings

3.3.1 Learning Metric Embedding

As [23] describes 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 $\mathbb{R}^D$. The function $f_\theta(\cdot)$ ideally maps *similar* (a measure of similarity has to be defined, as in Section 3.3.2) points in the input space to nearby points on the manifold [20].

Suppose the use of this transformation for vehicle ReID. This embedding would be achieved 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 [50], clustering [18], face recognition [51] and last, but not least, ReID [21].

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.

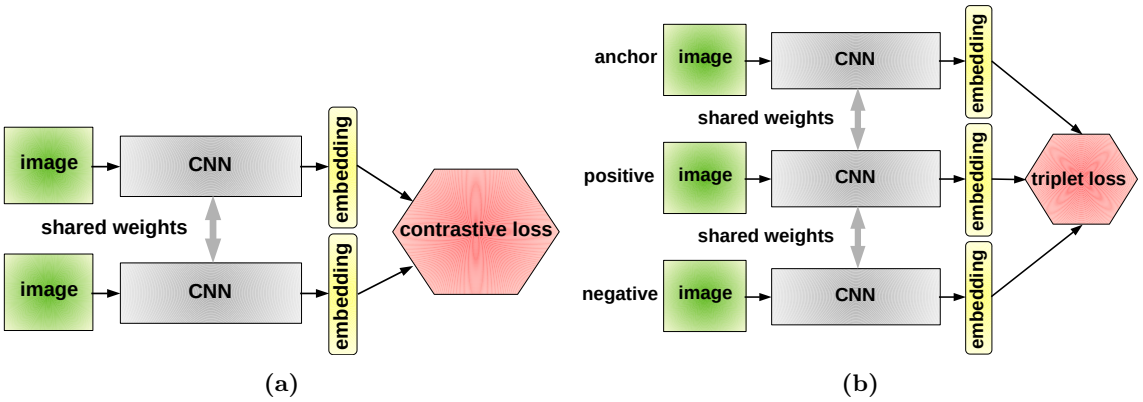


Fig. 3.7: Comparison of the Siamese **(a)** and the triplet **(b)** network architectures. The concept of weight sharing implies that the same network is used for inference and only one set of weights is trained. Architectures are depicted as containing multiple models only for conceptual understanding.

The Euclidean distance between the vectord $\mathbf{u}$ and $\mathbf{v}$ is defined as

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

and the cosine similarity is defined as

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

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

The major distinction between these measures is that cosine similarity only measures the angular difference between the vectors regardless of their magnitude, whereas the Euclidean distance takes the vector length into account as it quantifies the distance between the "tips" of the vectors. However, one important property is that as long as the two vectors under consideration are normalized, *i.e.*, they are of unit length, then the squared Euclidean distance is proportional to the cosine similarity as given by

$$\left\|\frac{\mathbf{u}}{\|\mathbf{u}\|_2}-\frac{\mathbf{v}}{\|\mathbf{v}\|_2}\right\|_2^2 = 2-2\frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|_2\|\mathbf{v}\|_2} = 2-2\cos\angle(\mathbf{u},\mathbf{v}) \in \langle 0,4\rangle\,. \tag{3.6}$$

3.3.3 Siamese and Triplet Networks

For the upcoming discussion, let $D(x,y):\mathbb{R}^D\times\mathbb{R}^D\to\mathbb{R}$ 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

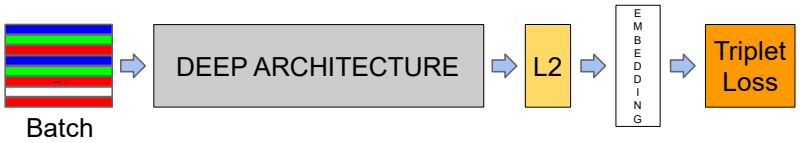


Fig. 3.8: A model structure developed by [18]. This network consists of a batch input layer and a deep CNN followed by an L_2 normalization, which produces the final embedding. A triplet loss is used during the training. (*source*: [18])

space to separate positive from negative samples. Then the contrastive function for any sample is defined as [20]

$$\mathcal{L}_{contr}(\theta) = \frac{1}{2} y D\left(f_\theta(x_0), f_\theta(x_1)\right)^2 + \frac{1}{2}(1-y)\left([\alpha - D\left(f_\theta(x_0), f_\theta(x_1)\right)]_+\right)^2. \tag{3.7}$$

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. 3.7 **(a)**). Positive samples should have a small distance between each other as measured by the $D(\cdot)$ to decrease the loss towards 0. On the other hand, negative samples should have a distance beyond the threshold α to incur a loss of 0.

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. 3.7 **(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 (see Fig. 3.8).

Let N be the number of all possible valid triplets $\left(x_a^i, x_p^i, x_n^i\right)$ 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 $y\left(x_a^i\right)$, x_p^i be the positive sample of the same object with label $y\left(x_p^i\right)$, such that $x_a^i \neq x_p^i \wedge y\left(x_a^i\right) = y\left(x_p^i\right)$, and let x_n^i with label $y\left(x_n^i\right)$ be a sample of any other object, satisfying $y\left(x_a^i\right) \neq y\left(x_n^i\right)$, $\forall i = 1, \ldots, N$. Let α be the margin value that is enforced between positive and negative pairs. Then, we want the relationship

$$D\left(f_\theta\left(x_a^i\right), f_\theta\left(x_p^i\right)\right) + \alpha < D\left(f_\theta\left(x_a^i\right), f_\theta\left(x_n^i\right)\right), \forall i = 1, \ldots, N, \tag{3.8}$$

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

$$\mathcal{L}_{triplet}(\theta) = \sum_{i=1}^{N} \left[\alpha + D\left(f_\theta\left(x_a^i\right), f_\theta\left(x_p^i\right)\right) - D\left(f_\theta\left(x_a^i\right), f_\theta\left(x_n^i\right)\right)\right]_+. \tag{3.9}$$

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

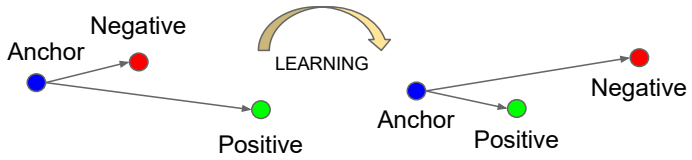


Fig. 3.9: The objective is to learn embeddings such that the anchor is closer to the positive example than it is to the negative example by some specified margin value. The triplet loss, therefore, minimizes the distance between the anchor and the positive sample of the same identity and maximizes the distance between the anchor and the negative sample of a different identity. (*source*: [18])

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

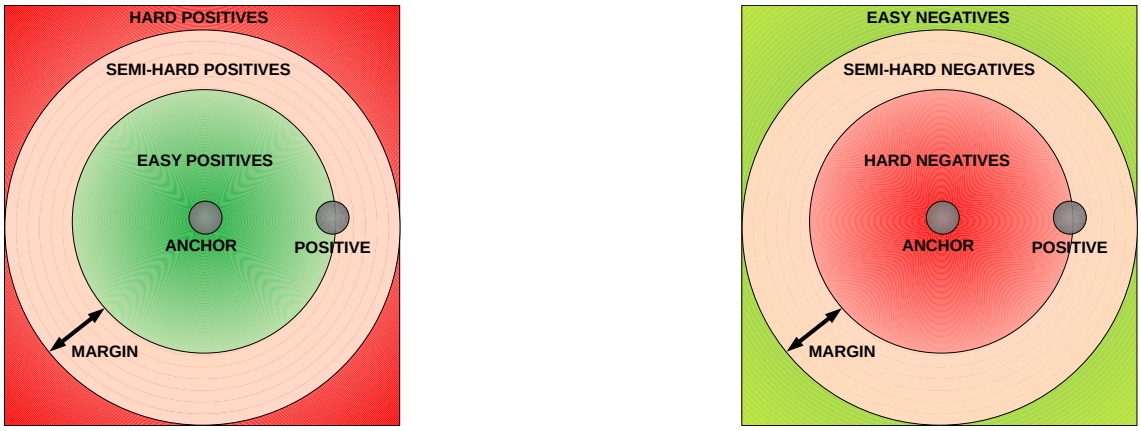
Table 3.1: Definitions of various categories of triplets (regardless whether it is positive or negative) as imposed by their distance relationship.

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 (see Fig. 3.9).

3.3.4 Triplet Mining Strategies

Contrastive (Equation 3.7) and triplet (Equation 3.9) 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.* [23] and Manmatha *et al.* [52] showed, it may significantly influence the training. The sampling strategy matters at least as much as the loss function itself. 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, because generating all possible triplets produces many triplets that fulfill the constraint in Equation 3.8. To paraphrase the analogy from [23], 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 3.1, using a general $\left(x_a^i, x_p^i, x_n^i\right)$ triplet for clarity. We encourage the reader to observe Fig. 3.10, too.

In order to attain an effective convergence during the training, it is necessary to select triplets that violate the triplet constraint in Equation 3.8. This means that given x_a^i, the goal is to select a *hard positive* x_p^i given by $\arg\max_{x_p^i}\left\{D\left(f_\theta\left(x_a^i\right), f_\theta\left(x_p^i\right)\right)\right\}$ and a *hard negative* x_n^i as a result of $\arg\min_{x_n^i}\left\{D\left(f_\theta\left(x_a^i\right), f_\theta\left(x_n^i\right)\right)\right\}$. Admittedly, it is often infeasible to compute the $\arg\min\{\cdot\}$ and $\arg\max\{\cdot\}$ over the entire training set. In this



(a) **(b)**

Fig. 3.10: 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.

regard, there are two possible approaches to tackle this problem, either by selecting these hard triplets online or doing it offline [18].

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 network checkpoint. Each sample in a batch consists of a triplet, so the embeddings have to be computed three times. Subsequently, hard or semi-hard triplets may be selected. Since this strategy has been shown on multiple occasions [18, 23, 21] 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. 3.11). 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.* [18] 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*.

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 harder triplets [23]. This

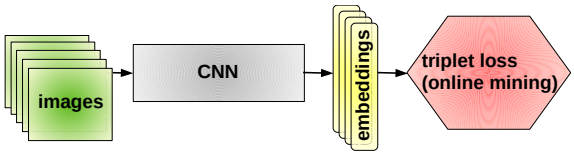


Fig. 3.11: The architecture of a triplet network where an online triplet loss function is used for training. In this architecture, no weight sharing is required as the triplet selection happens *online* solely in the loss function.

approach produces a total of $PK\left(K-1\right)\left(PK-K\right)$ triplets (PK anchors, $K-1$ positives per anchors, $PK-K$ negatives) incorporated in the loss function as

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

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. Additionally, Hermans *et al.* [23] say that the selected triplets can be considered moderate since they are the hardest within a small subset of the data, which is the best for learning with the triplet loss.

$$\mathcal{L}_{batchhard}\left(\theta\right)=\sum_{i=1}^{P}\sum_{a=1}^{K}\Bigg[\alpha+ \max_{p=1,\ldots,K}\left\{D\left(f_\theta\left(x_a^i\right),f_\theta\left(x_p^i\right)\right)\right\}- \min_{\substack{j=1,\ldots,P\\n=1,\ldots,K\\j\neq i}}\left\{D\left(f_\theta\left(x_a^i\right),f_\theta\left(x_n^j\right)\right)\right\}\Bigg]_+ \tag{3.11}$$

3.4 Evaluating Information Retrieval

In this section, we assume the reader is fully equipped with the basic notions of accuracy, precision, recall, and some other derived metrics. We will use the standard terminology to denote whether the classification is correct (true) or not (false), for both positive and negative categories. All these measures can be obtained from the confusion matrix (see Table 3.2), which we provide for completeness.

	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. (*source*: [53])

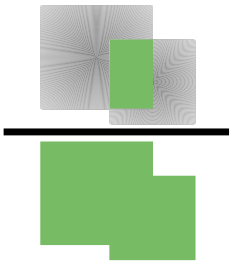


Fig. 3.12: Computation of the IoU metric between two BBOXes using of ratio of the area of overlap and the area of the union.

3.4.1 Evaluating Bounding Box Prediction

Intersection Over Union

Intersection over union (IoU) measures the overlap between two boundaries. It is a quantitative measure of overlap between two BBOXes, where one is the ground truth BBOX and the other is the predicted one (see Fig. 3.12).

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 integer 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

$$\begin{aligned}\mathbf{b}_1 \cap \mathbf{b}_2 = &\max\left\{0, \min\left\{x_1 + w_1, x_2 + w_2\right\} - \max\left\{x_1, x_2\right\} + 1\right\} \times \\ &\max\left\{0, \min\left\{y_1 + h_1, y_2 + h_2\right\} - \max\left\{y_1, y_2\right\} + 1\right\},\end{aligned} \tag{3.12}$$

and the area of their union is given by

$$\mathbf{b}_1 \cup \mathbf{b}_2 = w_1 h_1 + w_2 h_2 - \mathbf{b}_1 \cap \mathbf{b}_2. \tag{3.13}$$

Then, the final IoU metric between $\mathbf{b_1}$ and $\mathbf{b_2}$ is computed as

$$\mathrm{IoU}\left(\mathbf{b}_1, \mathbf{b}_2\right) = \frac{\mathbf{b}_1 \cap \mathbf{b}_2}{\mathbf{b}_1 \cup \mathbf{b}_2}, \tag{3.14}$$

where $0 \leq \mathrm{IoU}\left(\mathbf{b}_1, \mathbf{b}_2\right) \leq 1$, 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, 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) belongs to commonly used approaches 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. Such evaluation measures are used to determine how well the search results satisfied the user's query intent.

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

$$\mathrm{AP}\left(q\right) = \frac{1}{N_{gt}\left(q\right)} \sum_{k} P\left(k\right) \times \delta_k, \tag{3.15}$$

where $P\left(k\right)$ represents precision at rank k, $N_{gt}\left(q\right)$ 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

$$\mathrm{mAP} = \frac{1}{Q} \sum_{q} \mathrm{AP}\left(q\right), \tag{3.16}$$

where Q is the total number of query images, as described in [21]. Equation 3.16 essentially tells us is that, for a given query q, we calculate its corresponding AP (defined in Equation 3.15), and then take the mean of the all these AP scores, quantifying how well our model responds to queries on average.

Object Detection

Object detection models seek to identify the presence of objects in images and then classify them. 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. Even though we defined mAP for object ReID (Section 3.4.2), the mAP is also used for object detection [54].

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 [54], the AP is computed for 11 equally

spaced discrete recall levels, specifically $[0.0, 0.1, 0.2, \ldots, 1.0]$, using

$$\mathrm{AP} = \frac{1}{11} \sum_{r \in \{0.0, 0.1, \ldots, 1.0\}} p_{interp}\left(r\right), \tag{3.17}$$

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

$$p_{interp}\left(r\right) = \underset{\tilde{r}:\tilde{r} \geq r}{max} \quad p\left(\tilde{r}\right), \tag{3.18}$$

with $p\left(\tilde{r}\right)$ representing the measures precision at a specific recall level $\tilde{r}$ [55].

3.5 Evaluating Visual Multiple Object Tracking

In our work, we tackle problems related to VOT, concretely MOT, for which there are no established solutions neither for prediction itself nor for its evaluation. When it comes to evaluating MOT performance, even after many years, 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) MOT metric [56] (further referred to as just CLEAR), that we will employ to evaluate and quantitatively assess the performance of a MOT system. The reasons are the following:

- This metric is still considered a reasonably effective and intuitive metric to use, despite multiple proposals for improvements [57].
- Numerous works in object tracking, especially tracking of people, report statistics from the MOT challenges that historically have utilized this metric.
- From the engineering perspective, there are standard frameworks (*e.g.*, [58]) that provide an evaluation of a custom MOT tracker inference with a plethora of configurations and additional metrics as a bonus making peeking into the performance of the tracker a lot easier.

Bernardin *et al.* [56], the authors of CLEAR discussed above, designed two crucial criteria that performance metrics should meet. Here we present their list in which the first two items are considered primary, whereas the remaining are expected properties of useful metrics. 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,

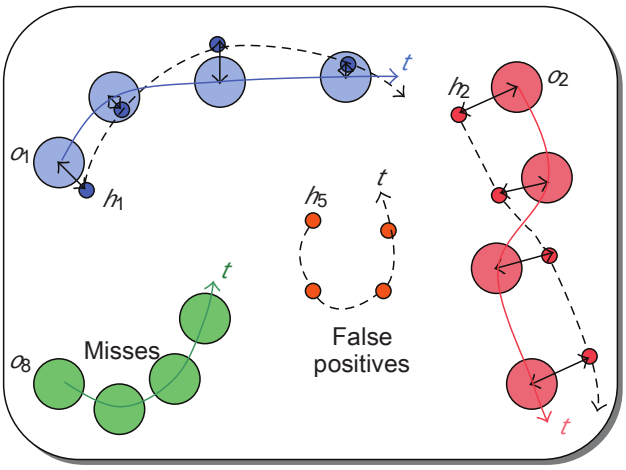


Fig. 3.13: With a demonstration of a correct tracker inference at the top, the CLEAR metric distinguishes between three fundamental types of errors, misses (false negatives), false positives and ID switches, shown in this order respectively. (*source*: [56])

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, *e.g.*, 2D or 3D, 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 $\{h_1, h_2, \ldots, h_m\}$ for a set of visible objects $\{o_1, o_2, \ldots, o_n\}$. The evaluation procedure can be briefly described in the following pseudocode.

For each time frame t:

1. Establish the best possible correspondence between hypotheses h_i and objects o_j, where $i = 1, 2, \ldots m$ and $j = 1, 2, \ldots, n$.

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 (see Fig. 3.13) in the found correspondences:

   (a) count all false negatives (misses), *i.e.*, objects for which there was no hypothesis,

   (b) count all 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 rather we could 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. We acknowledge that there may be numerous approaches to this problem. Bernardin *et al.* [56] 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. 3.14. 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. We agree with such objection, since local discrepancy in object-hypothesis correspondence may indicate just a temporary drift the tracker, whereas its ability to preserve the object's identity does not necessarily have to be as poor as the original metric would imply.

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 contradition 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 that the new one, even if the distance metric alone would indicate otherwise (see Fig. 3.15 for illustration).

3.5.3 Mapping Procedure

In what follows, we will describe a recipe for the mapping procedure.

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

1. Verify if every mapping in (h_i, o_j) in M_{t-1} is still valid. Such 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.

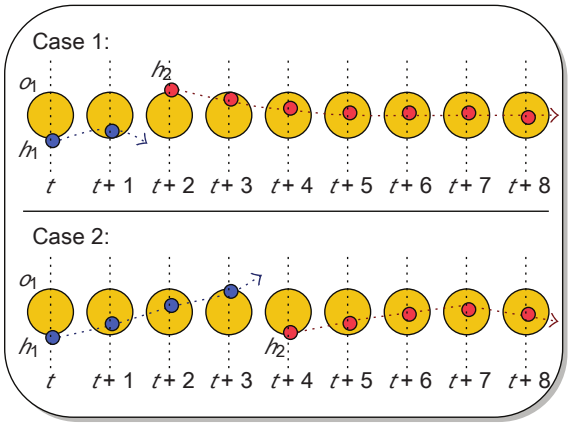


Fig. 3.14: Illustration of the inherent "unfairness" when relying on sequence-level "best" object-hypothesis mapping induced by the most frequent correspondence. As shown in the case 1, the correct hypothesis is the h_2, and thus only 2 errors are incurred for the first mismatch. The case 2 is practically identical, the h_2 also represents the most common assignment. However, 4 errors are accumulated for the alleged mismatch for h_1. (*source*: [56])

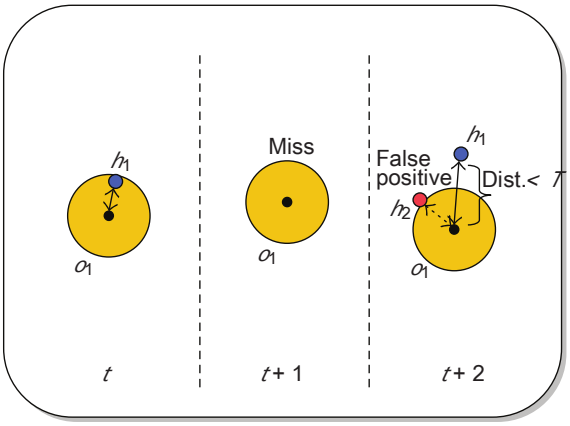


Fig. 3.15: Demonstration of a track reinitialization. At time t, the identity of the object o_1 is accounted for by the hypothesis h_1. At time $t+1$, the object disappears and the track is temporarily lost. At time $t+2$, the tracker is responsible for reinstantiating the object identity. During evaluation, 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*: [56])

2. If there are objects for which no correspondence has been made so far, then a suitable 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 minimum cost assignment problem and Munkres's algorithm [59] can be adopted. 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, *i.e.*, false negatives. Thus, let fp_t and m_t be the number of false positives and misses, respectively. For the sake of computation, let us define g_t as the number of ground-truth objects visible at time t.

Please note that no mismatch errors can occur at the initial frame as the set of mappings M_0 is empty and, therefore, all correspondences are treated as initializations.

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, namely the "tracking precision" and "tracking accuracy".

In general terms, the role of tracking precision is to measure the alignment between the predicted object position (*e.g.*, its BBOX) 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 properly. It is designed to evaluate the suitability of the delineation of the object BBOX the detection of which was correct in the first place. The Multiple Object Tracking Precision (MOTP) metric can thus be defined as

$$\text{MOTP} = \frac{\sum_{\forall t} \sum_{\forall i} d_{ti}}{\sum_{\forall t} c_t}, \tag{3.19}$$

Equation 3.19 represents 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. As stated above, it gauges the ability of the tracker to estimate precise object positions regardless of its capability of recognizing them or keeping their trajectories consistent.

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

$$\text{MOTA} = 1 - \frac{\sum_{\forall t} (m_t + fp_t + mme_t)}{\sum_{\forall t} g_t}, \tag{3.20}$$

the possible values of which lie within the interval $[-\infty, 1]$. For example, if a tracker produces a lot of false positives, the value in the sum may actually exceed one, and therefore, the final metric would be negative.

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 (in both equations 3.19 and 3.20) would lead to non-intuitive outcome (see Fig. 3.16 for details).

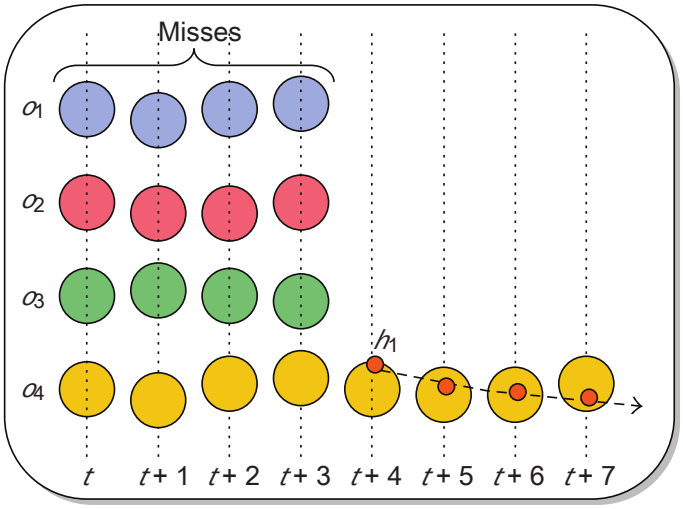


Fig. 3.16: Computing of error ratios needs to be performed on a global level, rather than on a local, frame level. Assume a sequence consisting of 8 frames. Moreover, assume that objects $o_1, \ldots, o_4$ are visible on the frames from t_1 to t_4, but none of them is being tracked. The situation changes at frame t_4 where only the object o_4 is being tracked properly by the hypothesis h_1. As a result, in frames $t_1, \ldots, t_4$, the resulting miss rate is 100%, whereas in frames $t_5, \ldots, t_8$ 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*: [56])

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 3.3.

3.6 Single Object Tracking

In this section, we will focus on the main topic of this thesis: tracking objects visually. Many research activities have been devoted to dealing with tracking of objects where visual input is the only information a model is provided with. Since the field of computer vision as a whole has transitioned from manual feature engineering in combination with shallow machine learning algorithms to using deep neural networks in end-to-end style, VOT as a subfield is no exception. We are going to discuss primarily SOTA approaches, and if we do mention older, maybe even obsolete methods, it will be only for the sake of reference and historical motivation. Furthermore, unless stated otherwise, we will assume a general object tracking model.

3.6.1 Initial Deep Learning-Based Solutions

At a time of publishing [60], 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 development of the famous Generic Object Tracking Using Regression Networks (GOTURN) [60]. This approach used to be

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.

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. 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. The previous frame is cropped and scaled so that it is centered on the target object. A small padding is made to allow for contextual information (see Fig. 3.17). A key concept to highlight is that GOTURN addresses the tracking as a box regression problem. In contrast, as will be presented later, similarity learning performs even better (Section 3.6.3).

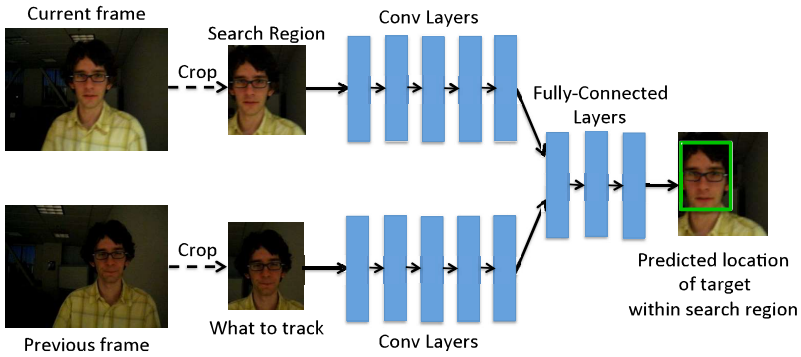


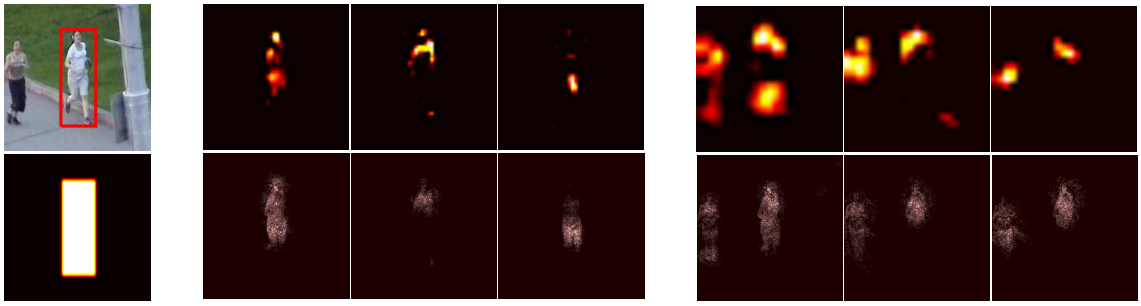
Fig. 3.17: The architecture of GOTURN showing the input as a search region from the current frame where the target from the previous frame is to be searched for. The network essentially learns to localize the target within the cropped region. (*source*: [60])

3.6.2 Fully Convolutional Tracking

We will start with a discussion about approaches that are so-called *fully convolutional*. 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 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 input of diverse shape. 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 [61]. 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 [62].

The CNNs provide valuable spatial clues about the image content. As we touched upon in Section 3.1.2, convolutional layers placed at the bottom (beginning) of the model tend to capture elementary details useful for discriminating between targets of similar appearance, whilst top layers (at the end) seize more abstract information regarding separate object categories. Thus, interclass variations are thoroughly captured in the top layers, and intraclass variations conversely in the bottom layers (see Fig. 3.18). This concept led the authors of [63] to propose a fully convolutional visual object tracker that exploits different layers of the pre-trained VGG network [64] (see Fig. 3.19). Thus, 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 [36]. It was found out, as suggested above, that characterization from different perspectives is provided by convolutional layers at different levels.

The motivation for the utilization of different convolutional layers was that existing appearance-based tracking methods adopted either generative or discriminative models for separating the foreground from the background [63]. To denote the distinction between



(a) **(b)** **(c)**

Fig. 3.18: A CNN model trained on image classification task carries spatial information. **(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*: [63])

the two, assume a classification task with the goal of estimating a function $f : X \to Y$, or $P(Y \mid X)$, where X represents the data and Y the associated labels. A generative classifier would estimate parameters $P(X \mid Y)$ and $P(Y)$ (in essence, the joint probability $P(X, Y)$) from the training data and then use the Bayes' rule to compute $P(Y \mid X)$. On the other hand, a discriminative classifier would estimate parameters of $P(Y \mid X)$ from the training data directly [65].

Authors of [63] 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, localized, and correlated to the regions of semantic objects.
- 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 interclass variations discussed at the beginning).

The proposed architecture in accordance to the aforementioned observation is described in Fig. 3.18 and Fig. 3.19.

3.6.3 Tracking Using Siamese Networks

Even though CNNs condense valuable visual information into low dimensional space upon which a tracker may be built, 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 3.3 dedicated to metric spaces and embeddings, an object representation supporting ReID requires different types of models, one of which is a *Siamese network* (Section 3.3.3). We have already mentioned our intention of utilizing custom

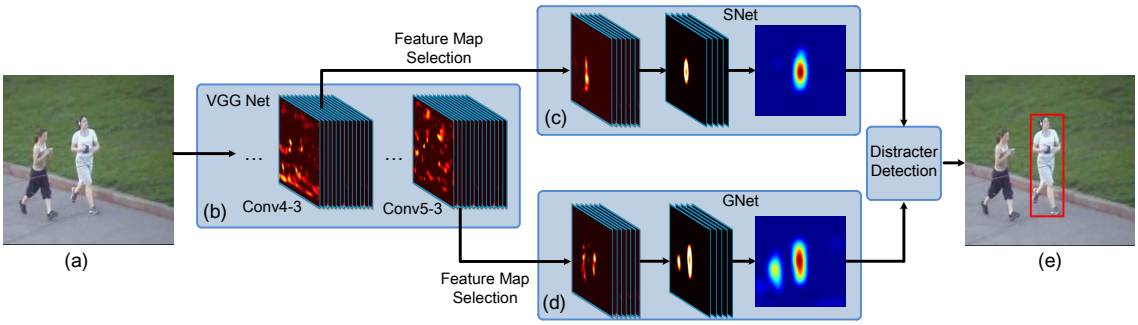


Fig. 3.19: For a given image, a feature map selection is performed on two different layers of the VGG network [64] to select the most relevant ones. This selection is handled as a heat map regression problem. A general network (GNet) captures the category information, whilst the specific network (SNet) captures visual traits based on which the foreground is separated from the background. Two heat maps are produced by these networks forming a basis for subsequent target localization. Both of these networks have to be initialized in the first frame. Afterward, when a new frame is to be processed, a region of interest is centered at the last target location, and the whole process repeats. (*source*: [63])

metric space for tracking, and [66] were among the first ones to successfully demonstrate it.

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 [67] solely focused on Siamese trackers and their fundamentals. At the time of writing these lines, this paper already has 2 international citations.

Authors of [66] 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* image within a larger *search* image (see Fig. 3.20). The model got the name `SiamFC`. We mentioned this to make the comparison easier because a lot of follow-up work has been done, producing models such as `SA-Siam` [68], `SiamRPN` [28], `SiamMask` [25], `SiamMask-E` [24], 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 this identical transformation γ to both inputs, search image x and exemplar image z, and then combine the result as

$$f(x, z) = g(\gamma(x), \gamma(z)). \tag{3.21}$$

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 use of the Siamese architecture spawned a lot of follow-up work, and we will describe some that serve our purposes. A team of authors in [68] 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).

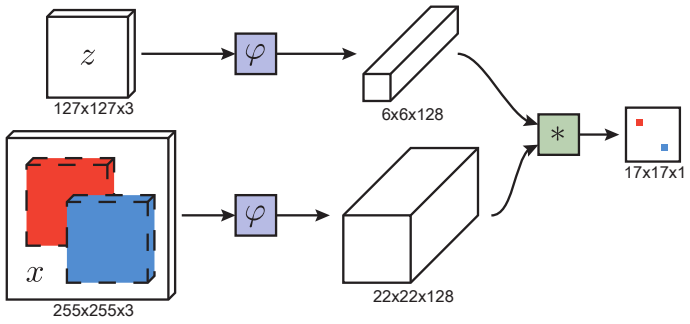


Fig. 3.20: The fully-convolutional Siamese architecture produces a scalar-valued score map, the size of which depends on the size of the search image. The similarity function is computed for all sub-windows within the search image and stored in a 2D score map, rather than just a pure 1D embedding vector. 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. Best viewed in color. (*source*: [66])

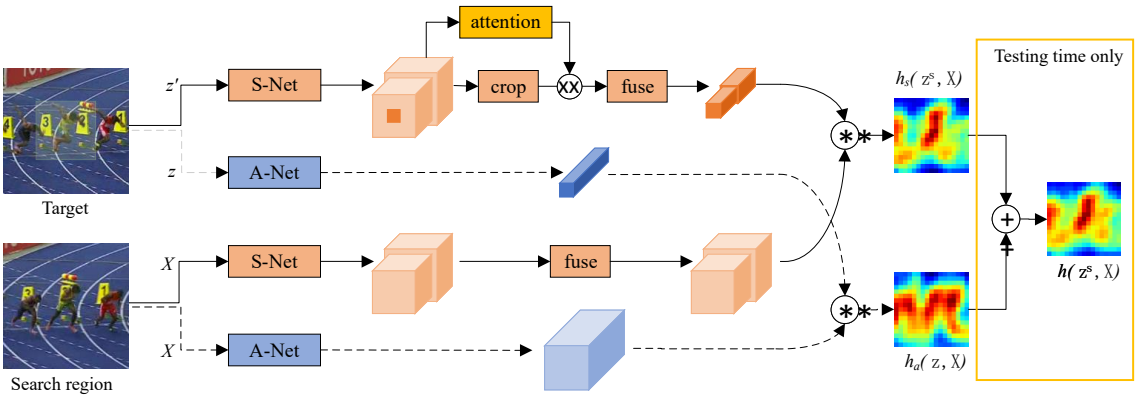


Fig. 3.21: The architecture of the `SA-Siam` network. The `A-Net` represents the appearance network and the `S-Net` represents the semantic network. For the reason that this work builds on `SiamFC` model [66], structures connected by dotted lines are exactly the same as in the `SiamFC` model. The last two convolutional layers provide features that are extracted afterward. The attention model determines the weight of each feature channel by simultaneous consideration of exemplar and context information. The fusion module is trivial and uses just 1×1 convolutions. As shown on the right, combining branches is allowed only when testing. (*source*: [68])

They also suitably commented that the key to designing a high-performance tracker is to harness expressive features that are simultaneously discriminative and generalized. In light of this, they developed a model consisting of a semantic and an appearance branch (see Fig. 3.21), with each branch being represented by a standard similarity-learning Siamese network (as in `SiamFC` [66]). 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 rationale behind such a decision 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 the hindsight, there are more important aspects of Siamese trackers to address in order 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 (see Fig. 3.21) cropped

from the initial (target, 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 (`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

$$h_a(z, X) = corr(f_a(z), f_a(X)), \tag{3.22}$$

where $corr(\cdot)$ 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 $i = 1, \ldots, N$, with N being the no. of chosen training samples. The parameters θ_a of the `A-Net` model are optimized from scratch by minimizing the logistic loss function $\mathcal{L}(\cdot)$:

$$\underset{\theta_a}{arg\,min} \frac{1}{N} \{\mathcal{L}(h_a(z_i, X_i; \theta_a), Y_i)\}. \tag{3.23}$$

Analogically, the semantic branch (`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 ($corr(\cdot)$) 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. Although the use of multilevel features in conjunction with the attention module delivers substantial progress for the semantic branch, it would be counterproductive for the appearance branch. Appearance features extracted from different convolutional layers lack sufficient differences in terms of expressiveness, as these features are dense whereas high-level semantic features are very sparse.

The attention module operates channel-wise. Assume the i^{th} channel from some convolutional feature map with spatial dimensions of 22×22 is being processed. The tracking target is contained in the center, covering a region of 6×6. The entire feature map is divided into a grid of 3×3 cells. Within each grid, a max-pooling operation is applied followed by a simple neural network consisting of just 1 hidden layer. The weight coefficient ζ_i for the particular i^{th} channel is generated by the sigmoid activation function (see

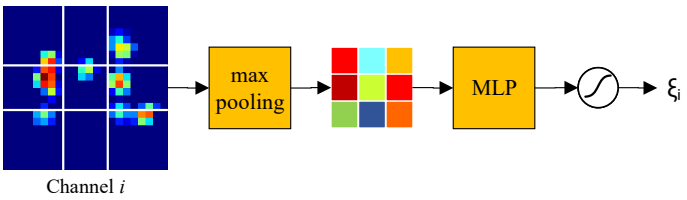


Fig. 3.22: The attention module of the `S-Net` network. (*source*: [68])

Fig. 3.22). The attention module 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. The response map for the semantic branch is produced as

$$h_s\left(z^s, X\right) = corr\left(g\left(\zeta \cdot f_s\left(z\right)\right), g\left(f_s\left(X\right)\right)\right), \tag{3.24}$$

where the no. of elements of ζ is equal to the no. of channels in $f_s\left(z\right)$, and the operator $\cdot$ indicates an element-wise multiplication.

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 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. Let (z_i^s, X_i) with their corresponding ground-truth response map Y_i, for $i = 1, \ldots, N$, be the N chosen training samples. The parameters θ_s of the `S-Net` model are fitted using the logistic loss function $\mathcal{L}\left(\cdot\right)$ (equal to Equation 3.23)

$$\underset{\theta_s}{arg\,min} \frac{1}{N} \left\{\mathcal{L}\left(h_a\left(z_i^s, X_i; \theta_a\right), Y_i\right)\right\}. \tag{3.25}$$

The inference phase involves computation of the overall heat map for which a weighted average of the two produced heat maps is used as shown below:

$$h\left(z, z^s, X\right) = \lambda h_a\left(z, X\right) + \left(1 - \lambda\right) h_s\left(z^s, X\right). \tag{3.26}$$

This introduces another hyperparameter λ (where $0 < \lambda < 1$) that can in practice be reliably estimated from the validation dataset.

The series of Siamese-based architectures for tracking continued with the idea of using the RPN [28] (see Section 3.2.3 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` [66]) and RPN as another subnetwork encompassing both classification and regression branch (see Fig. 3.23). 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) (see Fig. 3.24). 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.

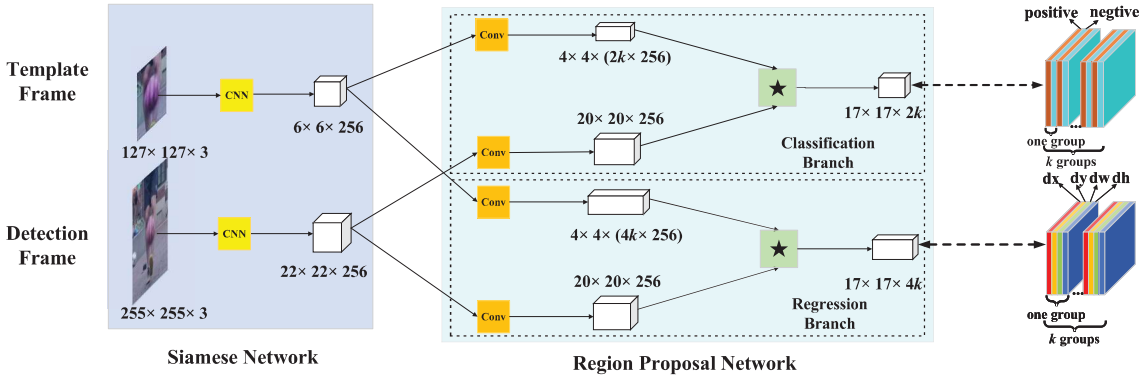


Fig. 3.23: The pipeline starts with the original `SiamFC` network followed by the RPN which has two branches: classification and regression. The output of the two branches is obtained using a pair-wise correlation. Foreground/background classification and the box regression are given by the $17 \times 17 \times 2k$ and $17 \times 17 \times 4k$ feature maps, respectively. (*source*: [28])

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. 3.23). 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

$$\begin{aligned} A^{cls}_{w\times h\times 2k} &= [\gamma(x)]_{cls} \star [\gamma(z)]_{cls}\,, \\ A^{reg}_{w\times h\times 4k} &= [\gamma(x)]_{reg} \star [\gamma(z)]_{reg}\,, \end{aligned} \tag{3.27}$$

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.

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, $\forall i = 1, \ldots, N$, it is given by

$$\underset{W}{arg\,min} \left\{ \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(\psi(x_i; W), y_i\right) \right\}. \tag{3.28}$$

One-shot learning aims to learn W when only a single template z is available. Discriminative one-shot learning tackles a major challenge of *learning to learn*, by finding a mechanism to integrate category information into the learner [69]. If we consider a meta-learning feed-forward function ω that maps $(z_i; W')$ to W, then the problem can be stated as

$$\underset{W'}{arg\,min} \left\{ \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(\psi\left(x_i; \omega\left(z_i, W'\right)\right), y_i\right) \right\}. \tag{3.29}$$

In this setting, this objective function can be re-written in terms of the Siamese subnetwork

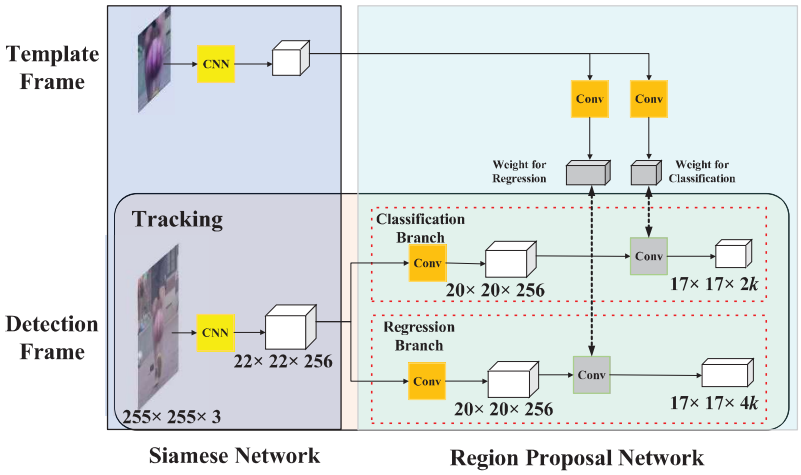


Fig. 3.24: Tracking is framed as one-shot detection here. First, the template branch predicts the weights of the kernels for the RPN using the first frame. Later on, only the detection branch is retained so the framework can be thought of as a local detection network. (*source*: [28])

feature extraction γ and region proposal subnetwork Ψ as

$$\underset{W}{arg\,min} \left\{ \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(\Psi\left(\gamma\left(x_i; W\right); \gamma\left(z_i; W\right)\right), y_i\right) \right\}. \tag{3.30}$$

The template branch provides training parameters to predict the kernel for the detection task, a typical example of the *learning to learn* process. The template branch, therefore, embeds necessary category information into the kernel that is subsequently utilized for detection (Fig. 3.24).

Later on, a fork of publications emerged with an endeavor to improve the tracking 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 [24]. Thus, datasets with rotated BBOXes provide tighter enclosed boxes. Additionally, the orientation information may be useful for solving different computer vision problems, such as action classification.

Inspiration from techniques for the problem of object segmentation yielded another approach where the tracking process was assisted with additional semi-supervised object segmentation [25]. 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. Notwithstanding its convenience, a single rectangle frequently fails to represent the object appropriately, hence the motivation to generate additional segmentation masks.

As always, the `SiamFC` [66] was employed as the fundamental building block. However, a notable alternation consisted of the use of a depth-wise cross-correlation layer instead of a simple cross-correlation layer. The latter one compresses all the information into one channel, impeding the potential to encode richer information about the target object. As a reminder, the original model used $6 \times 6 \times 128$ and $22 \times 22 \times 128$ tensors to produce a

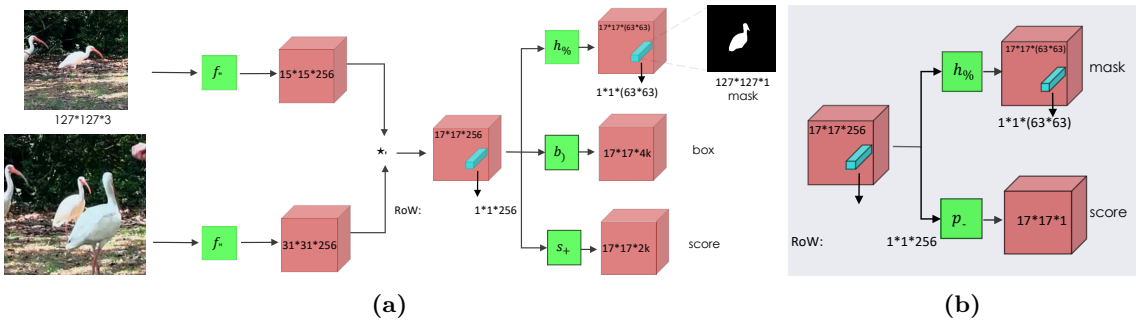


Fig. 3.25: A schematic illustration of the two `SiamMask` variants: **(a)** a three-branch architecture (full), **(b)** a two-branch architecture (head). Depth-wise correlation layers are adopted again. (*source*: [25])

$17 \times 17 \times 1$ response map (Fig. 3.20). Here, a multi-channel response maps are utilized (Fig. 3.25).

An incremental improvement of `SiamMask` model came when [24] 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 [25], 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 computer 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` [66] model) has been recently proposed, named `SiamCAR` [27]. 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 on hyperparameter tuning for successful tracking.

An indispensable part of localization are 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 of 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 [41]. This branch outputs a feature map where each point indicates the *centerness* score for the corresponding location. This concept was also utilized within the base architecture

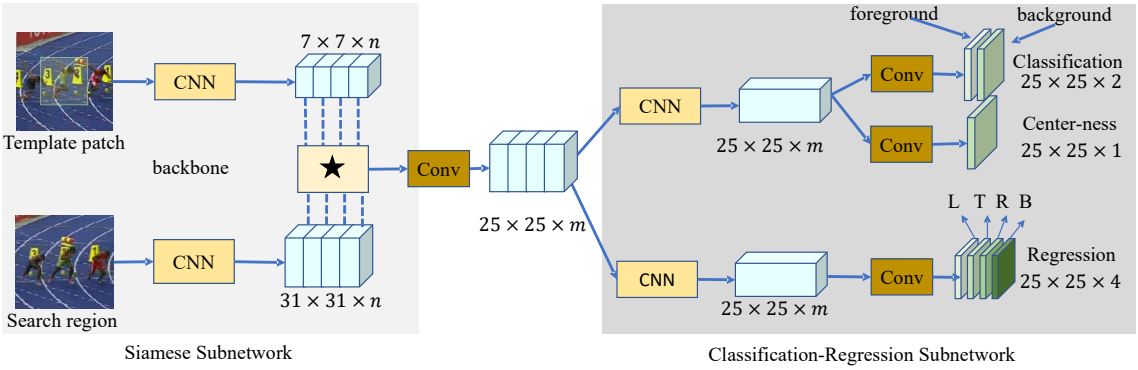


Fig. 3.26: `SiamCAR` architecture. The left side consists of the original `SiamFC` [66] model, with a simple amendment of using depth-wise correlation for multi-channel response map extraction. The right side depicts the subnetworks for foreground/background classification and BBOX regression. (*source*: [27])

we tested our experiments on. We can say that *centerness* is a very general concept, and practically it represents a weighting mechanism to penalize areas within the Region of Inteset (ROI) that most likely do not contain the target object.

Conclusions Made In the Survey Paper

Due to space limitations, in this short section we summarize the most important conclusions and observations that we made in our survey paper [67] when researching Siamese-based visual object trackers.

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. This treatise could be thought of as an in-depth analysis of the core principles on which Siamese trackers operate together with 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 thorough discusssions of Siamese trackers was published in 2018 [70]. Last but not the 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 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 [71],

custom sampling strategies for foreground/background discrimination [72], or conditional object re-detection [73]. Besides, there is 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-mention RPN head, *e.g.*, [28, 71, 74]. A comprehensive survey concerning deep visual object tracking by Marvasti-Zadeh *et al.* [75] also reached similar conclusion.

The utilization of cross-correlation has a great share of the leading performance in terms of their effectiveness. But the original single-channel formulation from [66] has been improved into a multi-channel cross-correlation that has been in use up to date. It was argued that a single channel did not capture sufficient information [27], thus multi-channel cross-correlation layers were used instead [74]. On top of that, nce multiple channels are present, we observed an emerging trend in using various attention mechanisms to aid the feature selection [76]. As we will see further, multi-channel cross-correlation was also exploited in our work from a practical standpoint.

Speaking of cross-correlation, its core principle of performing a "learned template matching" using the exemplar and the search region, it raises the question 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, [66, 77]. 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 loose its track.

Even though our discussed survey focused on SOT, there are emerging works where Siamese architectures were integrated into a MOT pipeline. We shall cover some important works in the next section.

3.7 Multiple Object Tracking

This section continuously expands the previously started discussion on single object trackers. Our research originally targeted SOT, especially Siamese single object trackers. The plan to incorporate multiple objects remained only as a hypothesis to explore later since it brings a whole new set of challenges to overcome. However, thanks to our comprehensive survey on Siamese tracking [67], we gained enough background knowledge to quickly absorb the newly emerging body of literature on a specific branch of multiple-object trackers that exploit Siamese architectures. We have to acknowledge that we did not compile a thorough SOTA overview of MOT for the following reason. 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 [59] through complicated graph formulations [78] to even graph-based convolutional neural networks [79]. Even though there are works that claim the use of Siamese neural networks in MOT, *e.g.* [80], their utilization is in terms of ReID within the tracking-by-detection philosophy, for which Siamese networks are widely adopted. By Siamese tracking we explicitly mean the type of trackers described in Section 3.6. Nevertheless, we did not necesarily need as much background knowledge in MOT to identify

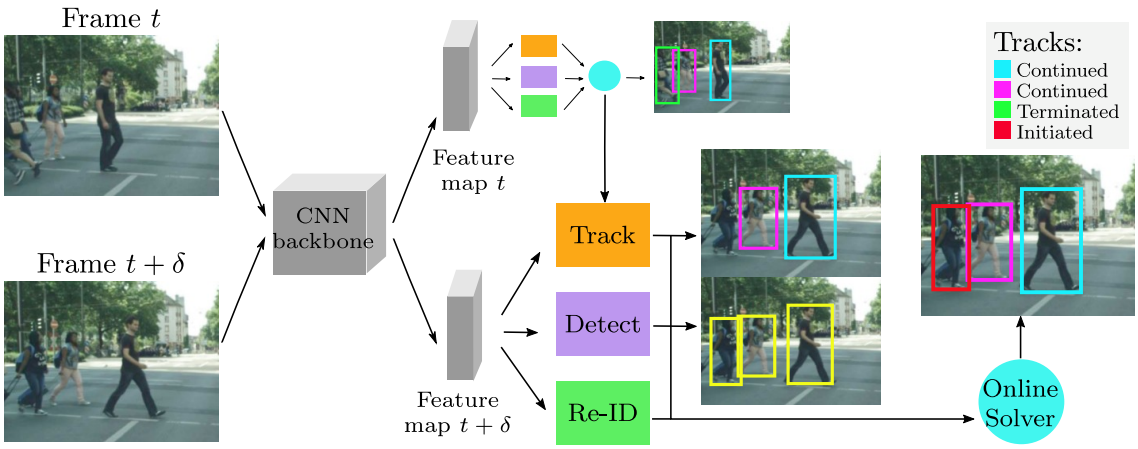


Fig. 3.27: Demonstration of how unification of the detection, tracking and ReID within a single architecture can be achieved. One important aspect that contributes to cost-effectiveness in terms of inference is that features required for all the mentioned tasks share the same backbone, which results in low computation and efficient runtime. (*source*: [81])

that Siamese-based MOT is a freshly rising subfield of trackers we should contemplate exploiting due to its direct applicability to traffic analysis. As we will demonstrate, all the best practices from Siamese SOT have found their use in MOT, too.

3.7.1 Siamese-based Multiple Object Tracking

Shuai *et al.* [81] proposed a Siamese-based framework that can simultaneously handle object tracking, detection, and ReID (see Fig. 3.27). 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, which is a great contribution. Although this tracking system follows an inference pipeline similar to other tracking-by-detection systems, the distinction is that it does so based on cues generated by a single network. One important remark made by the authors is that that tracking by use of ReID alone is not robust enough to short-term changes, especially in the presence of partial occlusions. Our experiments will provide corroborating evidence, too.

From our point of view one trivial, but at the same time, very effective extension of the often-mentioned `SiamFC` tracker utilized n exemplars to produce n response maps and, therefore, to perform tracking of n objects simultaneously [82], aptly dubbed as `SiamMT` (see Fig. 3.28). The authors mentioned the incentive to develop their tracker 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. We believe this paradigm of tracking is yet to uncover its full potential.

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. Lee *et al.* [83] combined Siamese similarity learning with Feature Pyramid Networks (FPNs) (discussed in Section 3.8.2). This tracker still follows the path of tracking-by-

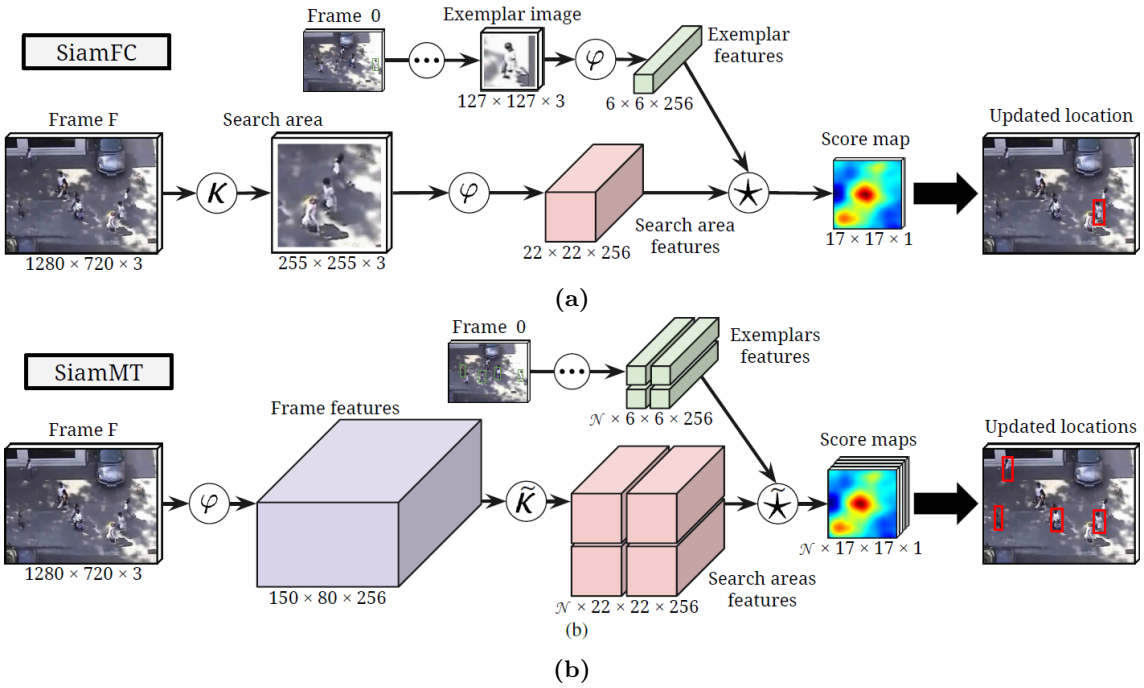


Fig. 3.28: The inference phase of the **(a)** `SiamFC` architecture to see the difference between the **(b)** `SiamMT` successor. The `SiamMT` framework first extracts features of the entire frame via the backbone φ. Subsequently, the obtained features belonging to distinct regions are cropped and resized 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 multiple-object response map indicating their predicted positions. (*source*: [82])

detection paradigm, in which the similarity metric between the current detections and existing tracks plays an essential role. In this work, a 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, in order 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 with working with `SiamMOT` [84] (Section 6.1) architecture, which we will introduce in great detail later on. It is a multi-object tracker that practically 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

The efficient capabilities of deep learning models of extracting robust features pertinent to the task at hand are of great significance. As we have observed in our survey of Siamese trackers [67], incremental improvements in feature extraction were often the major contribution of numerous works that granted the authors a competitive performance against other SOTA frameworks. With this in mind, we consider feature extraction a necessary part of any deep learning model design. As we will see further, the model with which we

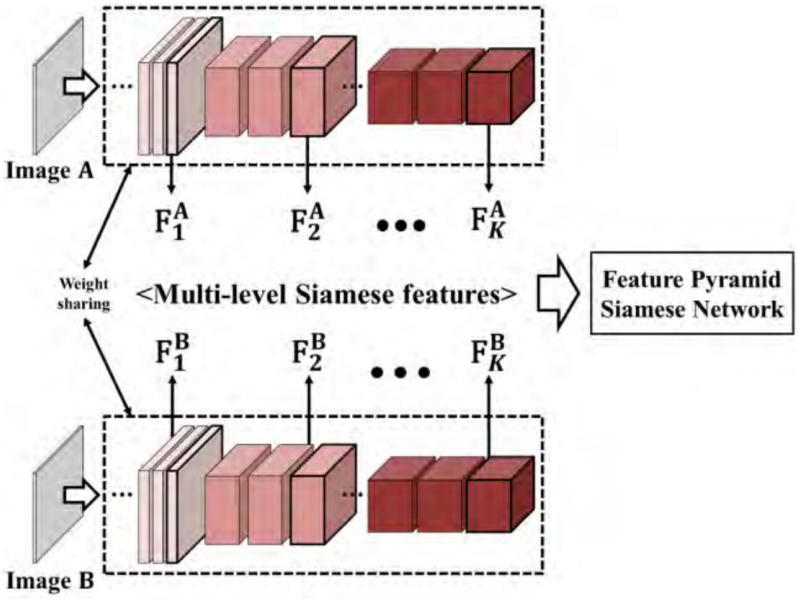


Fig. 3.29: A CNN serving as a backbone that adopts Similarity learning as an enhancement to feature aggregation using FPN. (*source*: [83])

performed the majority of our experiments exploited top feature extraction approaches. Thus, it is also suitable to provide a brief background to the most influential methods of feature extraction and feature fusion.

3.8.1 Residual Neural Networks

He *et al.* [85] aptly remarked that deeper neural networks are more difficult to train. In their work, the authors proposed a residual learning framework to facilitate easier training of neural networks that were significantly deeper than their previously used counterparts. 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 proposed architecture was dubbed as ResNet.

The foundation of ResNets is the adoption of skip connections that represent shortcuts to jump over some layers. Typically, such models are implemented using double or even triple layer skips containing nonlinearities (*e.g.*, Rectified Linear Unit (ReLU)) and batch normalization [86] in between.

The primary reason for adding skip connections was to avoid vanishing gradient problems. As demonstrated in Fig. 3.30, 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. By construction, a deeper network could at least learn an identity mapping, *i.e.*, to copy the previous value, and thus yield at least the same performance, but not worse.

Specifically, let $H(\mathbf{x})$ denote the desired underlying mapping. The stacked nonlinear layers are then expected to fit a different mapping $F(\mathbf{x}) = H(\mathbf{x}) - \mathbf{x}$. Thus, 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,

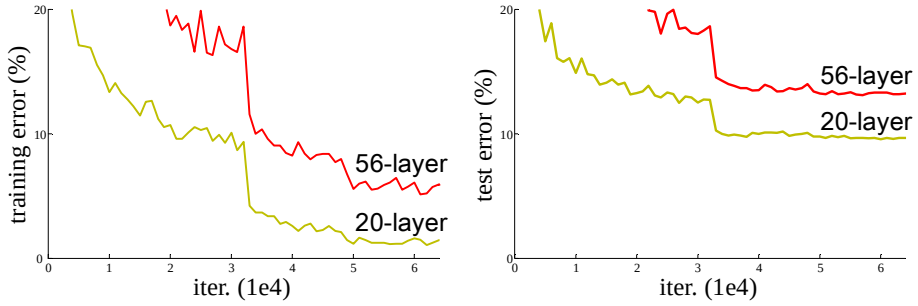


Fig. 3.30: The motivating behind the introduction of ResNets. The training error, and thus the test error as well, is greater for the deeper model than for the shallower model. Therefore, the inevitable conclusion is that in order to learn better networks, it takes more than just stacking more layers. (*source*: [85])

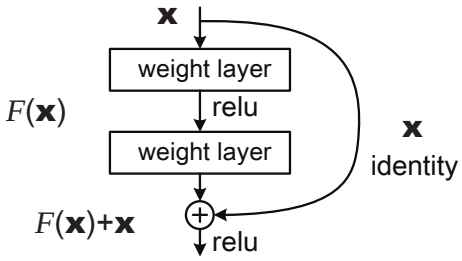


Fig. 3.31: A basic building block of residual learning demonstrating the mapping reformulation using skip connections. (*source*: [85])

unreferenced mapping. This formulation is visualized in Fig. 3.31.

3.8.2 Feature Pyramid Networks

When detecting objects visually at different image resolutions, pyramidal feature aggregation brings significant improvements. FPN [87] 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. 3.32 compares competing methods of feature aggregation by their core principles. Regarding the FPN itself, observe the two pathways in Fig. 3.32 **(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. Each lateral connection merges features of the same spatial size at each stage from the two pathways.

3.8.3 Deep Layer Aggregation

A successor of the previously discussed FPN is the DLA [88]. This architectural extension emphasizes the importance of feature aggregation across multiple levels to merge

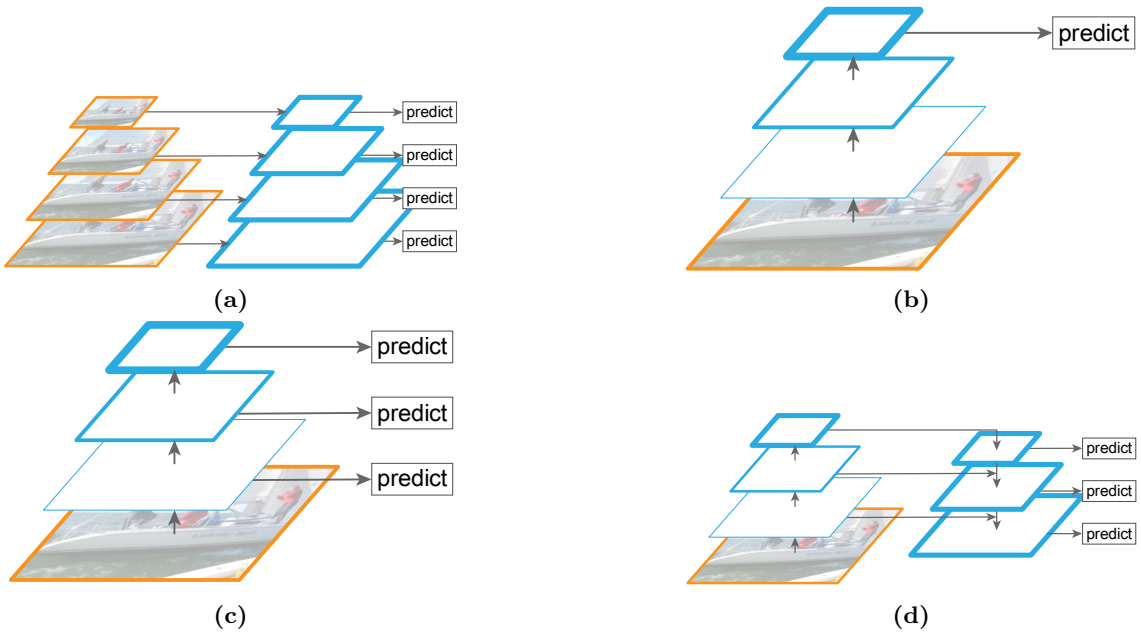


Fig. 3.32: A comparison of four traditional approaches to feature aggregation. **(a)** Computing features on distinct image scales (computationally expensive); **(b)** the use of single scale features only (fast, but not robust); **(c)** Reusing pyramidal feature hierarchy (fast and robust); **(d)** the proposed FPN - pyramidal feature aggregation in both directions (practically fast as previous methods but considerably more accurate). (*source*: [87])

information from different stages of input processing (see Fig. 3.33). Experimentally, this technique shows significant improvements in both memory usage and performance over frequently employed baselines such as ResNet [85] or DenseNet [89]. 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) (see Fig. 3.34 for further ideas). 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.

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

$$I\left(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_n\right) = \begin{cases} \mathbf{x}_1 \text{ if } n = 1 \\ I\left(A\left(\mathbf{x}_1, \mathbf{x}_2\right), \mathbf{x}_3, \ldots, \mathbf{x}_n\right) \text{ otherwise} \end{cases}, \tag{3.31}$$

where A is the aggregation node.

Hierarchical Deep Aggregation

HDA focuses on feature merging across all modules as well as channels. The 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

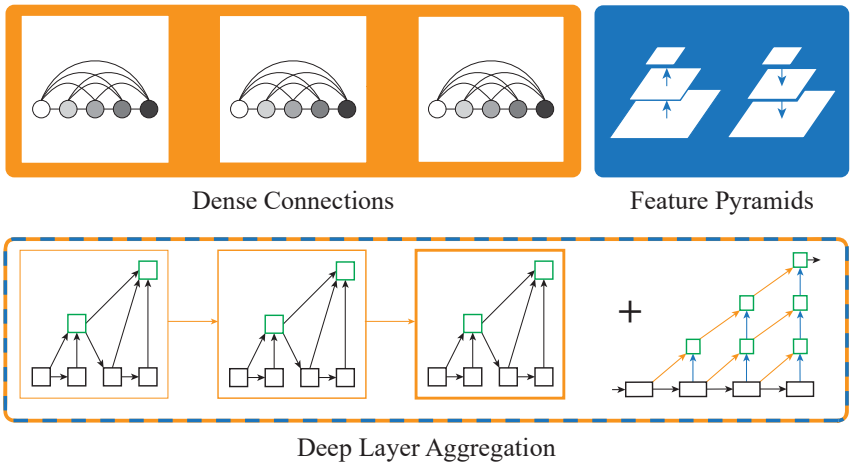


Fig. 3.33: A demonstration of unification of semantic and spatial information. The Deep Layer Aggregation (DLA) architecture extends densely connected networks, *i.e.*, Densely Connected Convolutional Networks (DenseNets), and FPNs. This extension builds on the idea of skip connections for enhanced feature fusion. (*source*: [88])

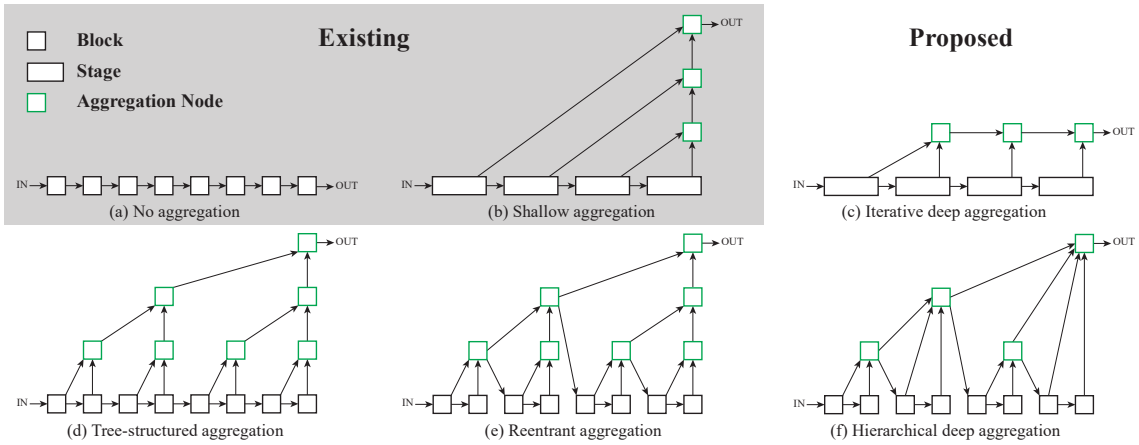


Fig. 3.34: Different approaches to feature aggregation. **(a)** No aggregation; **(b)** Shallow aggregation using skip connections; **(c)** Reordering of skip connections; **(d)** shallowst parts are aggregated the most; **(e)**, **(f)** further refining for deeper aggregation by routing intermediate aggregations back into the network. (*source*: [88])

can be formulated as

$$T_n(\mathbf{x}) = A\left(R^n_{n-1}(\mathbf{x}), R^n_{n-2}(\mathbf{x}), \ldots, R^n_1(\mathbf{x}), L^n_1(\mathbf{x}), L^n_2(\mathbf{x})\right), \tag{3.32}$$

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

$$\begin{aligned} L^n_1(\mathbf{x}) &= B\left(R^n_1(\mathbf{x})\right), \\ L^n_2(\mathbf{x}) &= B\left(L^n_1(\mathbf{x})\right) \end{aligned} \tag{3.33}$$

and

$$R^n_m(\mathbf{x}) = \begin{cases} T_m(\mathbf{x}) & \text{if m = n - 1} \\ T_m\left(R^n_{m+1}(\mathbf{x})\right) & \text{otherwise} \end{cases}, \tag{3.34}$$

where B represents some convolutional block.

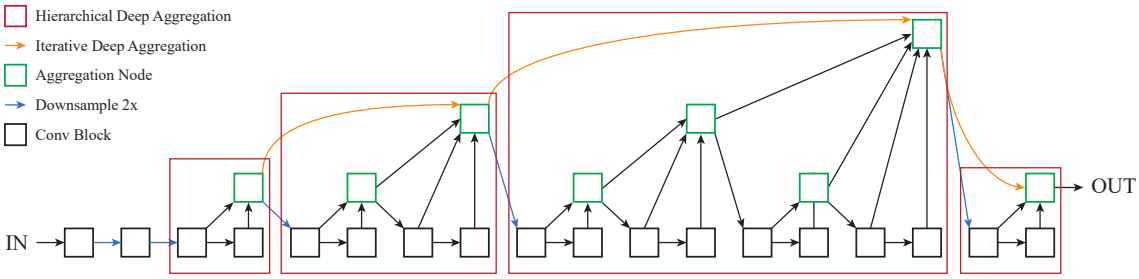


Fig. 3.35: DLA promotes enhanced feature extraction in spatial and semantic spectrum from the underlying network. In this combined approach, iterative connections progressively deepen and spatially refine the feature representation by joining neighboring stages. Simultaneously, hierarchical connections cross these stages with a tree-like structure to aid gradient propagation through the model. (*source*: [88])

Combination of Approaches

The two approaches above are independent as well as compatible enough to facilitate combining the two for even richer feature aggregation as shown in Fig. 3.35.

Chapter 4

Overview of Relevant Datasets

4.1 Object Detection Datasets

4.1.1 MS-COCO

The `MS-COCO` dataset [90] (web: [91]) 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 (see Fig. 4.1).

4.1.2 KITTI Object Detection

The `KITTI Object Detection` dataset [92] (web: [93]) 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 occlusion, this dataset provides an occlusion tag that represents the level of occlusion, from full visibility to large occlusion (see Fig. 4.2).

4.2 Object Re-identification Datasets

In this section, we focus on re-identifying vehicles. Faces and their ReID have been researched more than vehicles, which have brought various datasets in that area. Nonetheless, several important datasets for our work exist and we will shortly review their properties.

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*: [91])

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*: [92])

4.2.1 CompCars

The `CompCars` dataset [94] (web: [95]) consists of data from two scenarios: web-nature and surveillance-nature (see Fig. 4.3). The data in the web-nature category encompass 163 cars with 1 716 car models. In total, there are 136 726 images that capture the entire car, whereas 27 618 of additional images provide only a view of car parts. The full car images are labels with BBOXes as well as viewpoints. The attributes each car model is labeled with are the following: maximum speed, displacement, number of doors, number of seats, and type of car. On the other hand, the surveillance-nature data contains 50 000 images of just the front view. This dataset is freely available for research purposes and can be downloaded without the need to ask for permission. Apart from ReID, the dataset is well prepared for computer vision tasks such as fine-grained classification, attribute prediction, car model verification (see Fig. 4.4).

4.2.2 PKU VehicleID

The `VehicleID` dataset [96] (web: [97]) 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 (see Fig. 4.5). 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.

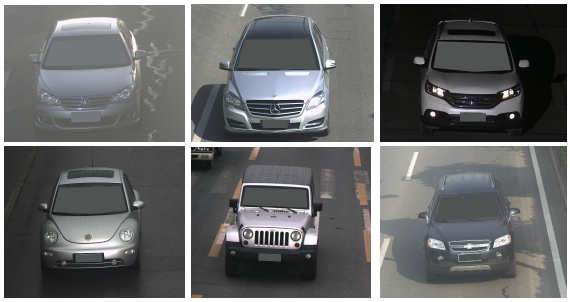


Fig. 4.3: Sample images of the surveillance-nature data from the `CompCars` dataset. The images have considerable appearance variations due to the varying conditions of light, weather, traffic, etc. (*source*: [94])

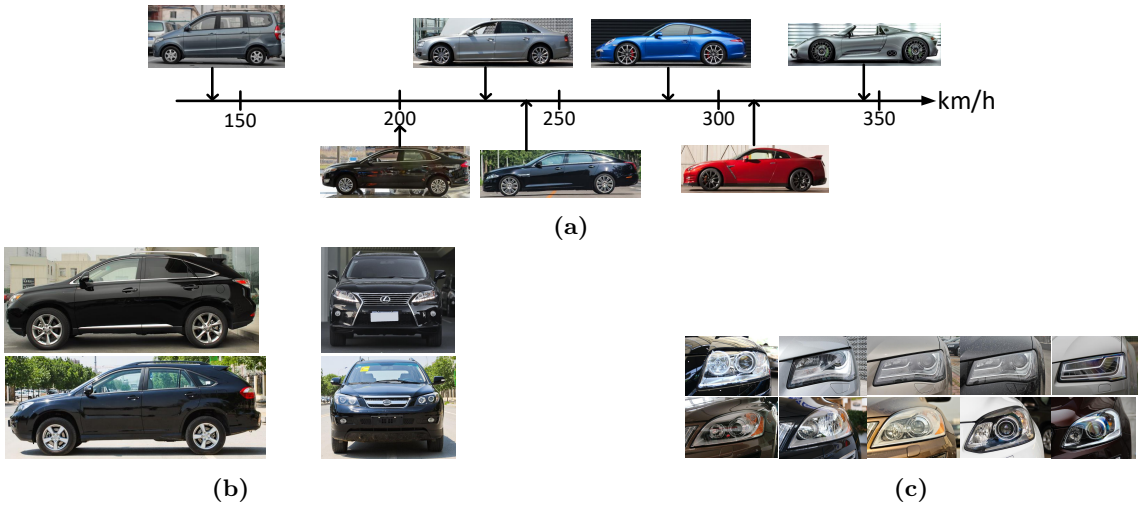


Fig. 4.4: Samples depicting various attributes of the `CompCars` dataset. **(a)** shows the possibility to predict the maximum speed of a car, **(b)** shows different views of the same car, **(c)** shows the evolution of headlights of two different car models (years: 2006 - 2014). (*source*: [94])

4.2.3 VeRI-776

A large-scale benchmark dataset named `VeRI-776` (web: [98]) for vehicle ReID in the real-world urban surveillance scenario [99] (see Fig. 4.6). 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. However, one has to send an official request to retrieve a copy. 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

Fig. 4.5: 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*: [96])

Fig. 4.6: The key properties of the `VeRI-776` dataset. Individual vehicles offer rich within-class differences in different viewpoints. At the same time, different but similar vehicles may have trivial inter-class differences. Moreover, the license plates as the unique ID are at disposal for a vehicle search. Additional contextual information can assist in vehicle searches in the city. (*source*: [99])

of vehicles, and the distances between neighboring cameras.

4.3 Visual Object Tracking Datasets

4.3.1 OTB 2013 and 2015

The OTB dataset [4] is widely adopted for performance evaluation of a SOT algorithms. The dataset contains up to 100 sequences each of which 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 [100] (web: [101]). Our personal opinion, based on the comprehensive survey regarding single object tracking that we composed [67], is that 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

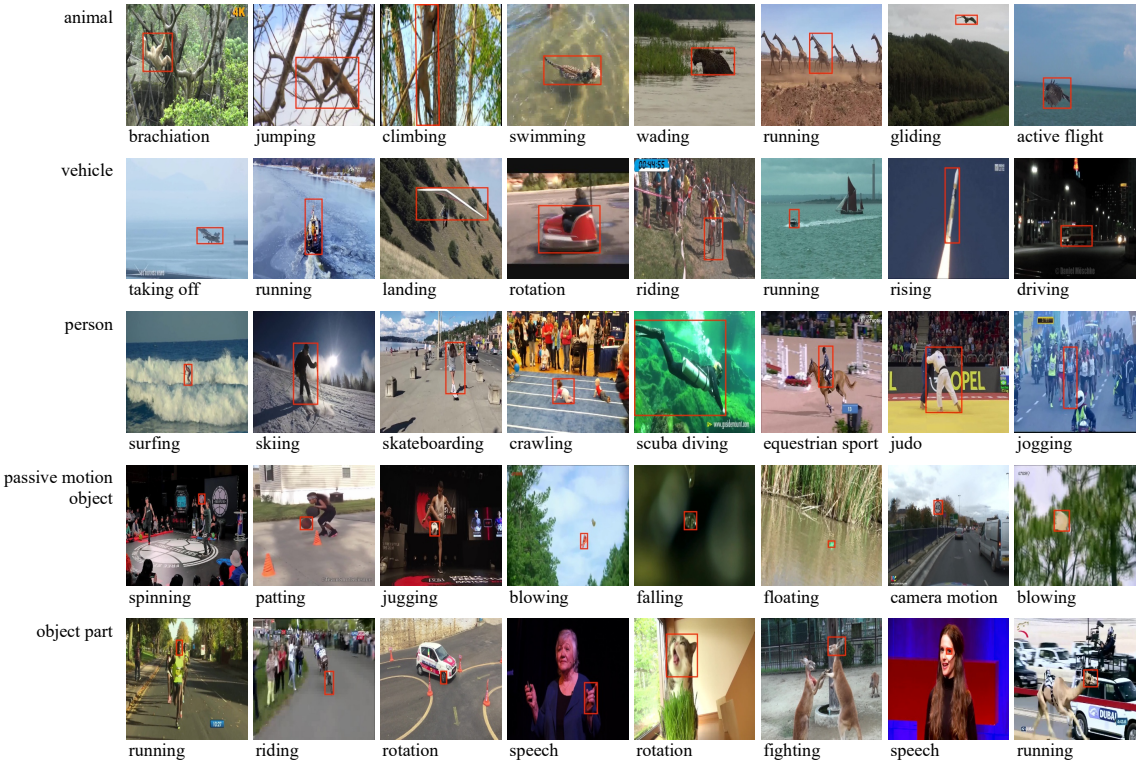


Fig. 4.7: Screenshots of some representative videos collected and annotated in the `GOT-10k` dataset. (*source*: [101])

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 (see Fig. 4.7). 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.

4.3.3 VOT 2015-2019

This prominent dataset (web: [102]) is a benchmark in visual object tracking [103]. 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.

Fig. 4.8: A sample of the only two classes ("car" and "pedestrian") that are evaluated in the benchmark using the `KITTI Object Tracking` dataset. (*source*: [104])

4.3.4 KITTI Object Tracking

This object tracking benchmark [92] (web: [104]) consists of 21 training sequences and 29 test sequences. Even though there have been labeled 8 different classes, only the classes "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` [57] (web: [105]). 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 (see Fig. 4.9). 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. Among other things, we treat this benchmark as the base of our experiments even thanks to the existence of an online leaderboard, which provides an opportunity to compare our solution with others using the same metrics.

`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` [92].

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

Fig. 4.9: 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*: [105])

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.

usage. In the beginning, there is a section describing some ignored regions (see Fig. 4.11). 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 4.1 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.

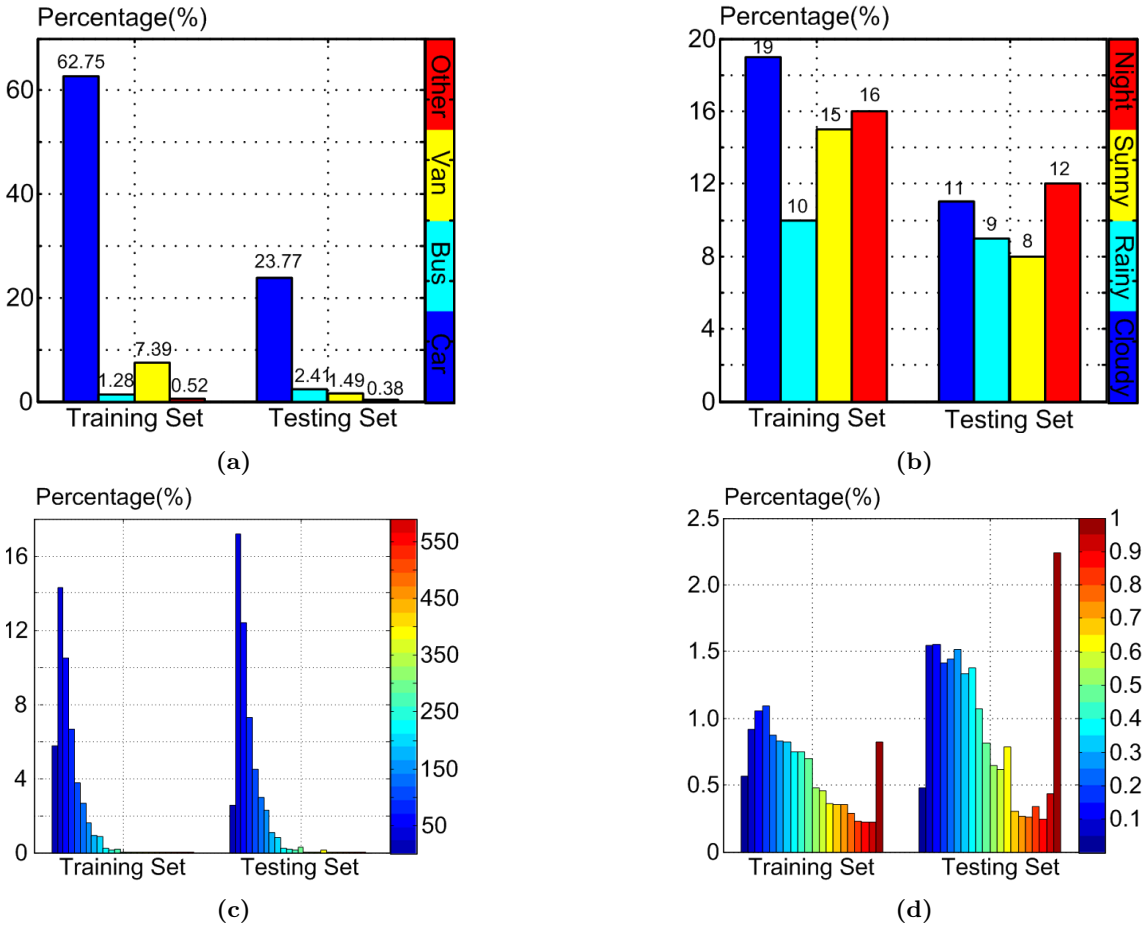


Fig. 4.10: Summary statistics of the `UA-DETRAC` dataset. **(a)** shows the distribution of vehicle categories, one of *car*, *bus*, *van* or *other*; **(b)** shows the varying weather conditions belonging to either *night*, *sunny*, *rainy* or *cloudy*; **(c)** depicts the change in scale given by the 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*: [105])

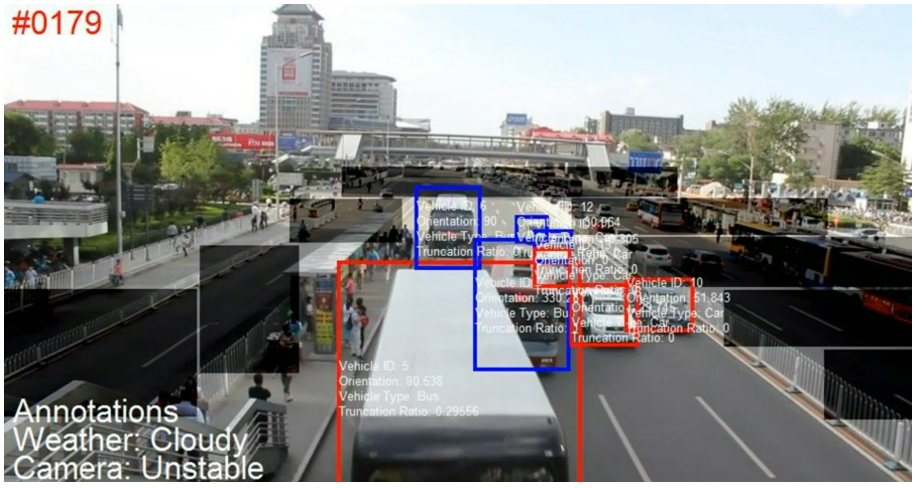


Fig. 4.11: A demonstration of a possible distribution of ignored regions in the `UA-DETRAC` dataset. (*source*: [105])

Chapter 5

Developed Homography Ranking Method

In this methodology-focused chapter, we describe one of our scientific contributions. We start with our attempts that did not manifest into primary advances in the field of object tracking per se, yet they were significant enough that they earned a journal publication in the end.

5.1 Introduction

This section is dedicated to one of our experiments that were not completely related to the VOT itself, yet we achieved an original scientific contribution in this area when exploring certain solutions that could potentially be applied to object tracking, especially traffic analysis. Even though we did not set out for homography-based object tracking (explained later) 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. We wrote up the whole research process in a paper called **Homography Ranking Based on Multiple Groups of Point Correspondences** [106], published in journal **Sensors** (web: [107]), under the category *Physical Sensors*. In what follows, we provide a concise report of our research. For more information, we suggest the reader use the aforementioned article, even though a great deal of information is duplicated here. Moreover, our preliminary discussion of this possibility with the initial proposal of the solution can be found in our first paper dubbed **Foundations for homography estimation in presence of redundant point correspondencies** [108] published in **Mathematics in Science and Technologies** (web: [109]) conference.

5.2 Motivation

One of the fundamental tasks of computer vision is dealing 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 [110, 111]. 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

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

The transformation above may facilitate mapping between two views of the same plane. It is a known fact that in the pinhole camera model, the homography relates any two distinct images of the same planar surface to each other [112, 113]. In particular, a single vector $\mathbf{u}^T = [u_x, u_y, 1]$, which represents a warped keypoint in homogeneous coordinates, is mapped onto its counterpart, 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 its most general form, the homography serves the purpose of mapping between various perspectives. However, we restricted our focus to producing a view where perspective distortion is absent. In other words, the objective 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 when taking the picture.

Homography is frequently adopted for text document rectification to generate a fronto-parallel view [114, 115], image stitching [116, 117], video stabilization [118], extracting metric information from 2D images [119], pose estimation [120], and for various traffic-related applications, *e.g.*, ground-plane detection [121], and bird's-eye view projection [122].

Our goal was to explore 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.* [123, 124, 125]. To provide more detail, Bose *et al.* [123] presented a fully automated technique for both affine and metric rectification of a given ground plane (up to a scale factor) 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 [8], so the rationale was to estimate the motion model in an

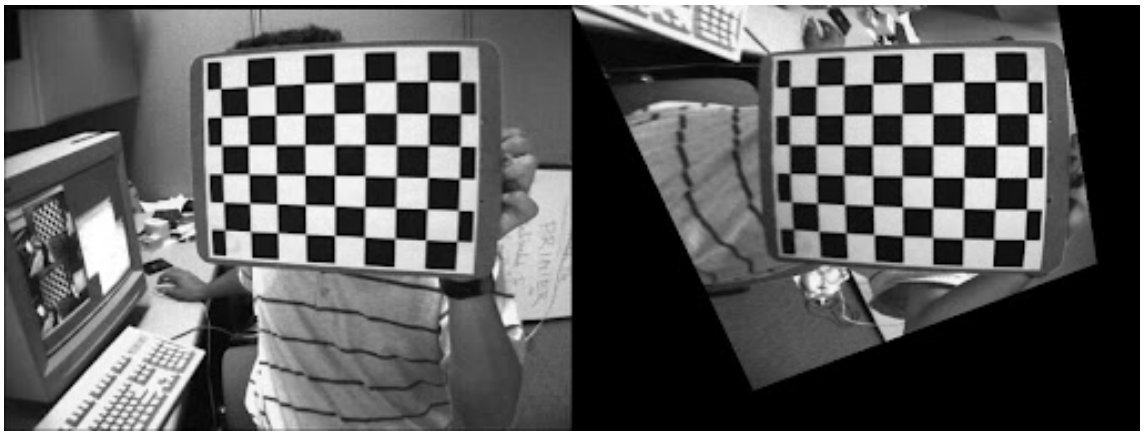


Fig. 5.1: An example of a chessboard marker present in a scene that may be used to establish a point correspondence that would serve for a homography transformation. The rectified, fronto-parallel view demonstrates the desired effect that points present on the "ground" plane (in this case, the chessboard) are properly projected, whereas other points suffer from substantial distortion. (*source*: [126])

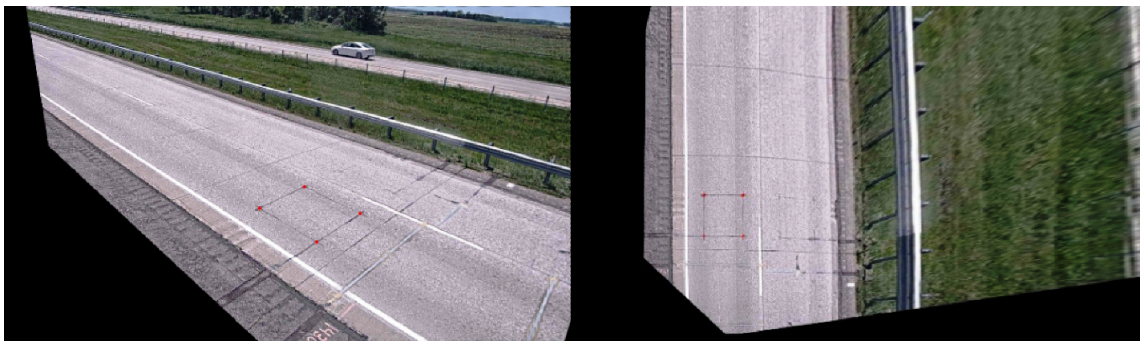


Fig. 5.2: An example of a virtual square marker present on a road that may be used to establish a point correspondence, and thus the homography transformation, too. The obtained view allows for many applications such as speed and size measurements that would otherwise be a lot more problematic in a perspectively deformed view. (*source*: [123])

orthogonal projection, rather than a perspectively distorted one.

A common approach to estimate the homography is to use a set of at least four 2D point correspondences [113]. 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 [127] (see Fig. 5.1). 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 (see Fig. 5.2). 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 suitable cost function [128, 129]. Subsequently, an outlier removal becomes an important step in the processing pipeline, for which effective and robust algorithms

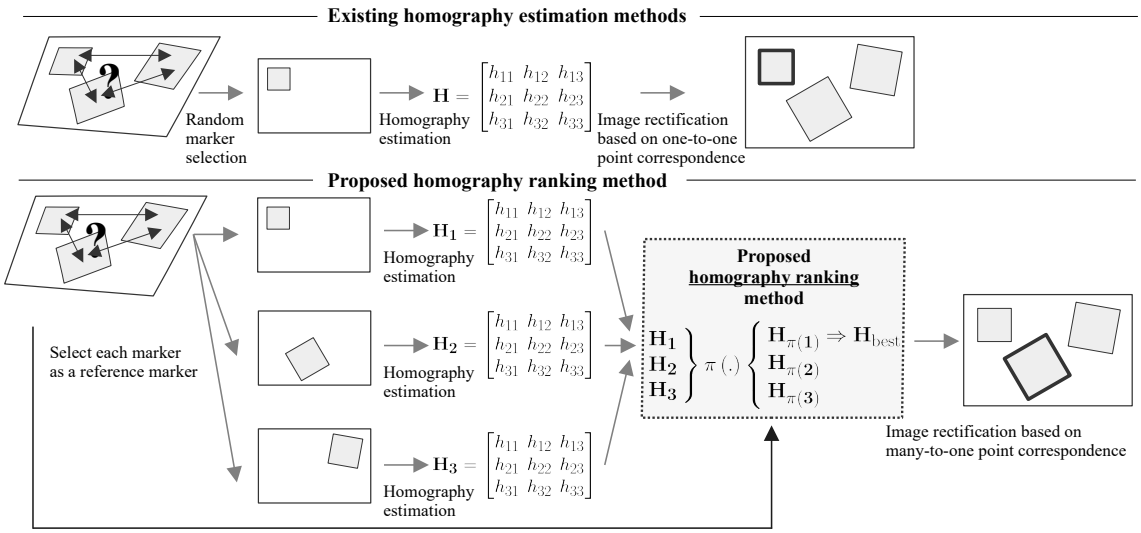


Fig. 5.3: A fundamental difference between existing homography estimation methods and our proposed method for homography ranking. If there are multiple markers while the information about their relative positions in the world is absent, the existing approaches can only estimate isolated homographies 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".

such as RANSAC [130] are usually employed.

A real-world application of generating a bird's-eye view over a road (see Fig. 5.4) from a video recording when we could not use a large marker to cover a sufficient portion of the road (see Fig. 5.5) 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. 5.2). 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 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 (see

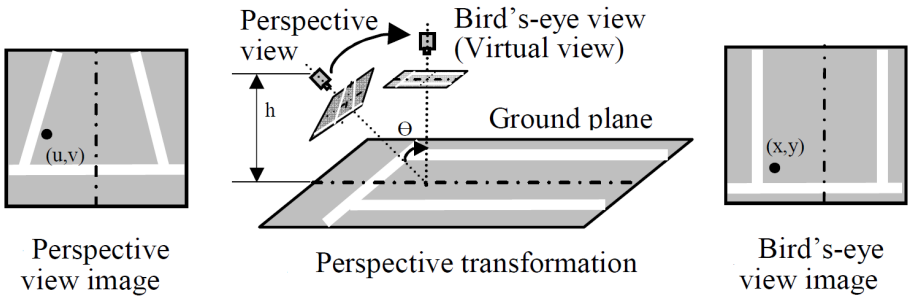


Fig. 5.4: A demonstration of the process of rectification when obtaining a fronto-parallel view over the road using homography. (*source*: [122])

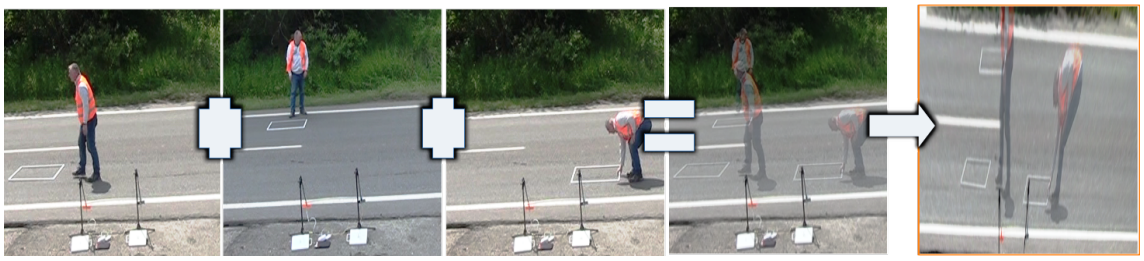


Fig. 5.5: A motivating real-world example for the proposed method. We can see different frames captured during a video recording that show various positions of the same marker. The picture after the "equality" sign is a merge of the previous frames for better illustration. Due to the use of a static 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?

Fig. 5.3). 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, so 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.

5.3 Preliminaries

This section contains a description of our custom terminology. Despite the existence of standard conventions for naming certain aspects of our problem, we nevertheless had to coin a few more terms for clarity.

We define a marker as an object with a known, easy-to-detect shape. Such object can be either naturally occurring or artificially placed on the planar surface of the scene we want to remove perspective distortion from, *i.e.*, to produce a bird's-eye view. 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

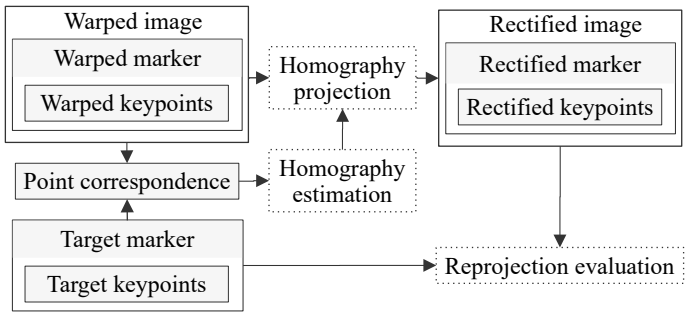


Fig. 5.6: Visualization of relationships within our established terminology. This diagram also depicts the hierarchical dependence between individual terms. In addition, the dotted elements represent processes with arrows denoting their input and output.

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 (see Fig. 5.6).

Unless stated otherwise, a **similarity transformation** denotes a limited affine transformation with 4 DoF which encompasses translation, rotation and uniform scaling (Equation 5.5). 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

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

Analogivally, 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

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

with the point correspondence relationship formulated as

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

5.4 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 (see Fig. 5.7). 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. This particular aspect was the primary source of confusion during the review process of our paper.

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 [130] 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 markers 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 on solutions that we heavily rely upon. As already highlighted, we did not contribute to this problem in terms of improving the homography estimation itself. We provided a way to rank the resulting homographies according to the devised score function. 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. This suggestion was often brought up during the review process by multiple independent reviewers.

To summarize, 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.

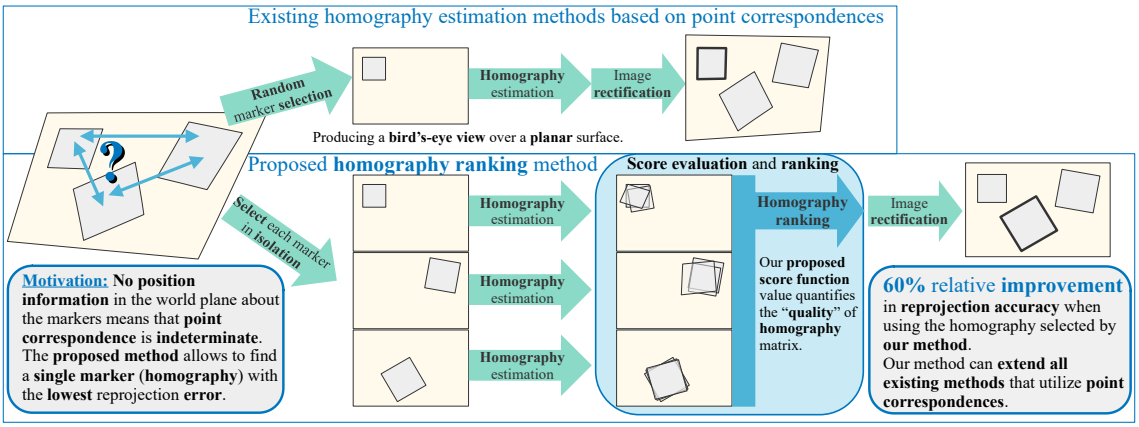


Fig. 5.7: Here we present the graphical abstract from our paper. The basic idea is that existing approaches may only estimate an isolated homography for each marker and cannot determine which homography achieves the best reprojection over the entire image. Therefore, we proposed a method to rank isolated homographies obtained from multiple distinct markers to select the best homography. This method extends existing approaches in the post-processing stage, provided that the point correspondences are available and the markers differ only by similarity transformation after rectification. We demonstrated the robustness of our method using a synthetic dataset and showed an approximately 60% relative improvement over the random selection strategy based on the homography estimation from the `OpenCV` library.

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. Thus, its adoption is restricted solely for the removal of perspective distortion.

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. 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 (see Fig. 5.8). 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 5.5). Then, 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 5.7. The objective is to minimize the computed quantity. The optimal similarity matrices are just auxiliary and redundant after the algorithm ends.

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

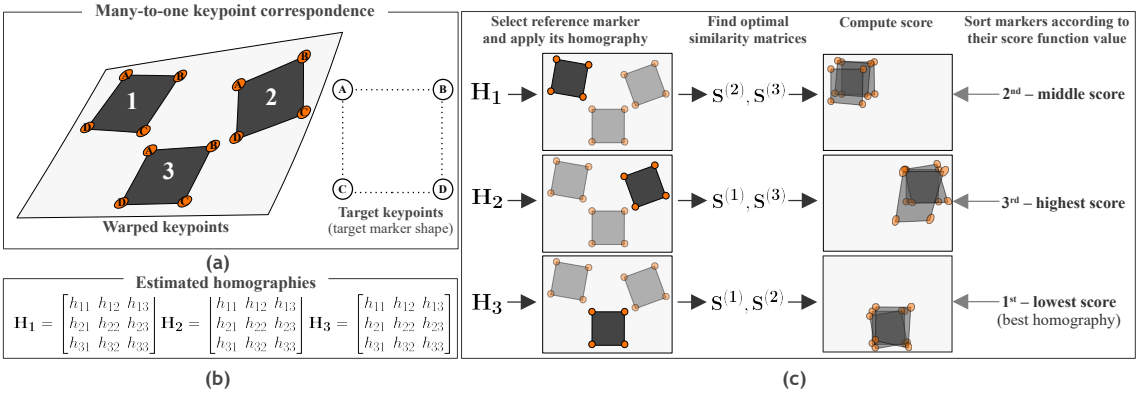


Fig. 5.8: A system diagram depicting the underlying principles behind our method. **(a)** The input consists of a many-to-one point correspondence specified by multiple similar markers together with the information about the ground-truth shape (up to an arbitrary positive scale) of the target marker. **(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 (sorting in ascending order) the homographies. The homography that ends up ranked first is considered (predicted) to the "best" candidate for achieving the minimal reprojection error over the whole image.

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

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

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

$$\mathbf{R}^{(i)}_{2\times 2} = \begin{bmatrix} s^{(i)} \cdot \cos\left(\theta^{(i)}\right) & -s^{(i)} \cdot \sin\left(\theta^{(i)}\right) \\ s^{(i)} \cdot \sin\left(\theta^{(i)}\right) & s^{(i)} \cdot \cos\left(\theta^{(i)}\right) \end{bmatrix}, \quad \mathbf{T}^{(i)}_{2\times 1} = \begin{bmatrix} t^{(i)}_x \\ t^{(i)}_y \end{bmatrix}. \tag{5.6}$$

This transformation (besides the identity) involves 4 DoF: a single rotation angle $\theta^{(i)}$, two x and y translation coefficients $t^{(i)}_x$, $t^{(i)}_y$, 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 [131]. The application of homography that rectifies an image generates a frontal plane that is related to the ground-truth plane by a similarity transformation [112, 132]. Thus, we do not include the shear and we only support uniform scaling. This choice can be easily mathematically justified and can be found in the appendix section of our paper [106].

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 per-

spective 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

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

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

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

In what follows, we describe the proposed Algorithm 1 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 $\left(\mathbf{W}^{(i)}, \mathbf{T}\right)$, $i = 1, \dots, m$ 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. To measure this objective, the score function defined in Equation 5.7 comes into place. 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 of markers, thus $\Theta\left(m\left(m-1\right) + m\log_2\left(m\right)\right) \simeq \Theta\left(m^2\right)$.

It is important to remark that the two functions used to compute the homography and similarity matrices in the pseudocode above may stand for arbitrary methods that produce the required transformations.

The score function 5.7 is just a proxy for the reprojection error computed over the whole image. Since we utilize only a small subset of points from the entire image, which may be subjected to noise, the assumption that the "best" homography is the one our method ranks as first may not hold in every case. In very few cases, the marker that achieves the lowest score function value does reconstruct the remaining markers the best, but not the overall image. Nevertheless, the conducted experiments show that our method consistently preserves its performance under various conditions.

Algorithm 1 Homography Ranking

$\bar{\boldsymbol{H}} \leftarrow \textbf{array}\,[m]$ ▷ output array of homographies
$\boldsymbol{s} \leftarrow \textbf{array}\,[m]$ ▷ array of scores
for $i \leftarrow 1, \ldots, m$ **do**
  $\bar{\boldsymbol{H}}\,[i] \leftarrow \textsc{homography}(\mathbf{W}^{(i)}, \mathbf{T})$ ▷ retrieve or estimate perspective
  $\bar{\mathbf{S}}^{(i)} \leftarrow \mathbf{I}_{3\times 3}$
  $\bar{\mathcal{S}} \leftarrow \left\{\bar{\mathbf{S}}^{(i)}\right\}$ ▷ set of similarity matrices
  for all $j : \{1, \ldots, m\} - \{i\}$ **do**
    $\bar{\mathbf{S}}^{(j)} \leftarrow \textsc{similarity}(\bar{\boldsymbol{H}}\,[i] \cdot \mathbf{W}^{(j)}, \mathbf{T})$
    $\bar{\mathcal{S}} \leftarrow \bar{\mathcal{S}} \cup \bar{\mathbf{S}}^{(j)}$
  end for
  $\boldsymbol{s}\,[i] \leftarrow \mathcal{F}\left(\bar{\boldsymbol{H}}\,[i], \bar{\mathcal{S}}\right)$ ▷ evaluate score function 5.7
end for
$\boldsymbol{\omega} \leftarrow \textsc{argsort}(\boldsymbol{s})$ ▷ indirect sort
return $\bar{\boldsymbol{H}}, \boldsymbol{\omega}$

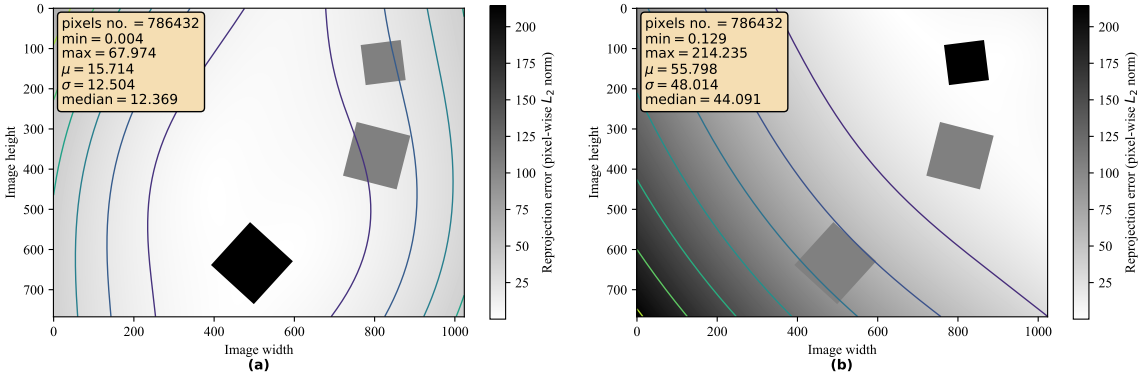


Fig. 5.9: Distribution of pixel-wise reprojection error. The heat map along with the corresponding contours demonstrate the varying distance between the ground truth and rectified pixel position after removing the perspective distortion. The bold square represents the reference marker. We show the result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

5.5 Experiments and Discussion

Fig. 5.9 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. Simply put, the transformation achieves the best accuracy in the vicinity of the chosen marker and steadily decreases for more distant pixels. 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.

The evaluation of the proposed homography ranking algorithm involved various condi-

tions. 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.

All the test scenarios indicated the following trend. On average, the homography 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. The change of shape and number of markers had the greatest impact.

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 [133]. Each homography was estimated by the `findHomography()` function which internally employs DLT [134] algorithm for $k = 4$ and RANSAC [130] 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.

5.5.1 Dataset Creation

Our synthetic dataset was created to simulate the presence of markers in the scene subjected to perspective distortion. Our experiments were based on 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 5.1 contains description of the generated test scenarios. To create test instances (within test scenarios), we employed the procedures described below (see Fig. 5.10).

Our dataset easily allows complete reproducibility of the reported results. Thanks to the synthetic nature of our data, fixing the seed for the used pseudo-random generator was sufficient. The source code for running the experiments is freely available on our Github repository [135].

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

In order to justify our use case, we demonstrated the effect of similarity transformations before applying the perspective transformation. 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.

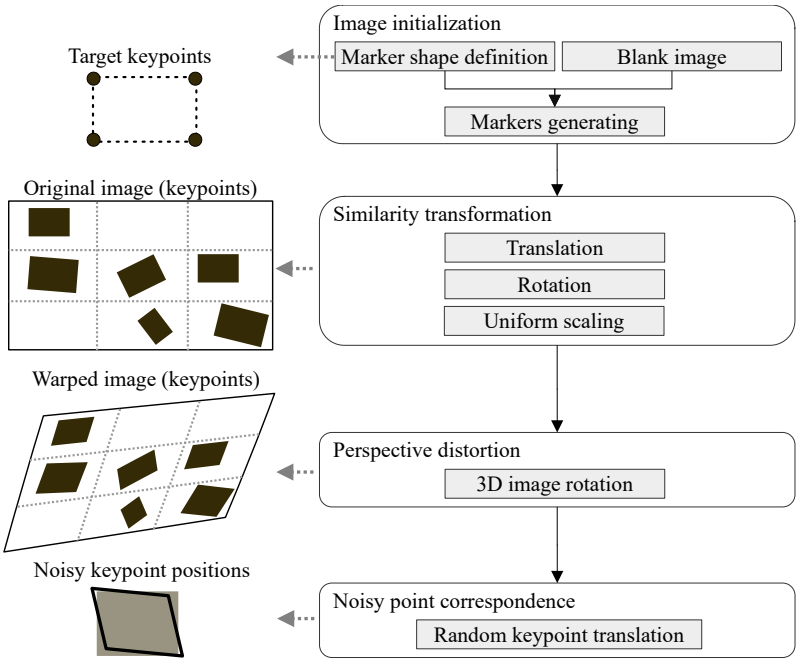


Fig. 5.10: Description of how each one of t test instances in a specific test scenario is created. The input is a blank $w \times h$ image over which m markers are initialized in a uniform grid, which produces the original marker keypoints. Depending on the test scenario, a particular subset of similarity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

5.5.2 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 $i = 1, \ldots, h, j = 1, \ldots, w$. We then compute the 2D reprojection error grid (a $h \times w$ matrix) for the given homography $\mathbf{H}$ as

$$\boldsymbol{\xi}_{wh} = \begin{bmatrix} e(\mathbf{w}_{1,1}, \mathbf{g}_{1,1}) & \ldots & e(\mathbf{w}_{1,w}, \mathbf{g}_{1,w}) \\ \ldots & \ldots & \ldots \\ e(\mathbf{w}_{h,1}, \mathbf{g}_{h,1}) & \ldots & e(\mathbf{w}_{h,w}, \mathbf{g}_{h,w}) \end{bmatrix}, \tag{5.9}$$

where

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

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

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

Evaluation Algorithm

On the input, there are m markers (Section 5.5.1) 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` [133] 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 5.11 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 5.7. 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 2. For simplicity, we show an evaluation of just a single instance.

5.5.3 Experimental Results

Fig. 5.9 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 representing 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.

Algorithm 2 Evaluation Algorithm

$\bar{\boldsymbol{H}}, \boldsymbol{\omega} \leftarrow$ RANKHOMOGRAPHIES( ) ▷ Algorithm 1
$e_b \leftarrow 0$ ▷ baseline
$\boldsymbol{e} \leftarrow$ **array** $[m]$ ▷ reprojection errors
$\boldsymbol{p} \leftarrow$ **array** $[m]$ ▷ relative improvements
for $i \leftarrow 1, \ldots, m$ **do**
    $\boldsymbol{e}[i] \leftarrow \xi_{\text{reproj}}$ ▷ Equation 5.11
    $e_b \leftarrow e_b + \boldsymbol{e}[i]$
end for
$e_b \leftarrow e_b / m$ ▷ mean reprojection error
for $i \leftarrow 1, \ldots, m$ **do**
    $k \leftarrow \boldsymbol{\omega}[i]$ ▷ position of i-th best homography
    $\boldsymbol{p}[i] \leftarrow (e_b - \boldsymbol{e}[k]) / e_b$ ▷ relative improvement
end for
return $\boldsymbol{p}$

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. 5.11, Fig. 5.12, Fig. 5.13 and in Fig. 5.14. 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` [133] method. We evaluated relative improvements for the sake of interpretability. For better comprehension, we suggest to see Table 5.1. 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 5.1. 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. 5.11 demonstrates that the relative improvement 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. 5.12, 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.

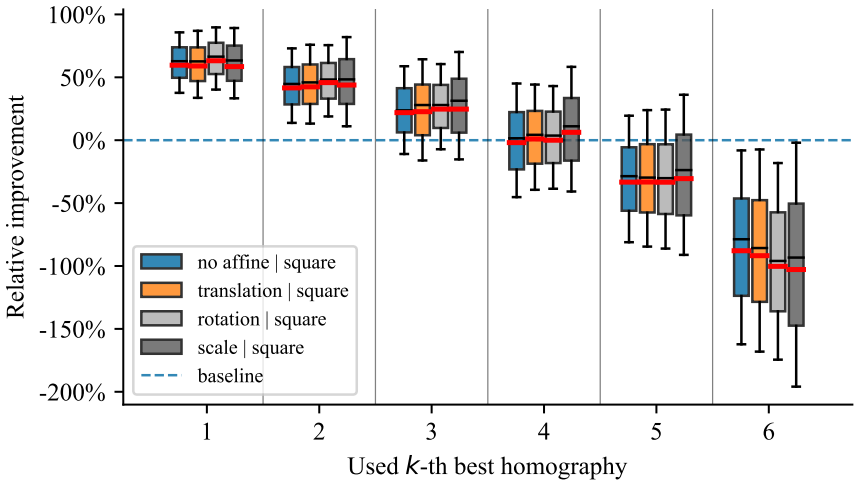


Fig. 5.11: Influence of similarity transformation on the reprojection error.

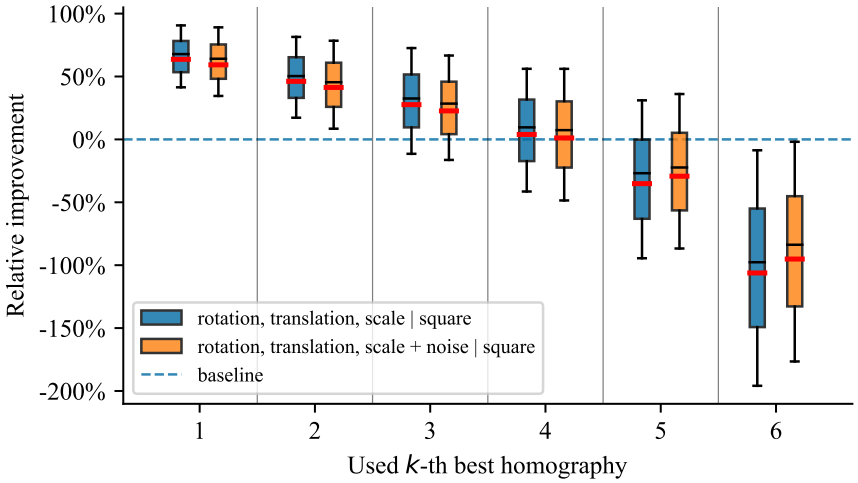


Fig. 5.12: Influence of noise applied to the warped keypoints representing a noisy point correspondence.

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. 5.13 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. 5.14 shows that the greater the set of markers, the better the relative improvement of our method. 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.

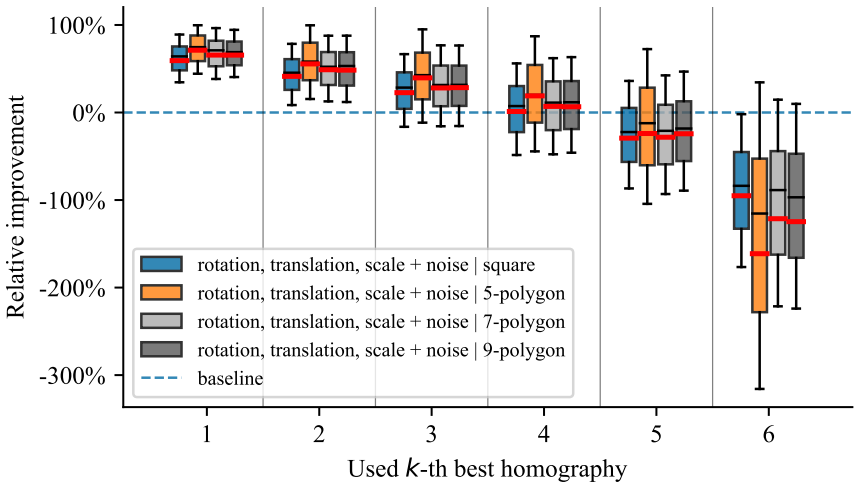


Fig. 5.13: Results for different marker shapes.

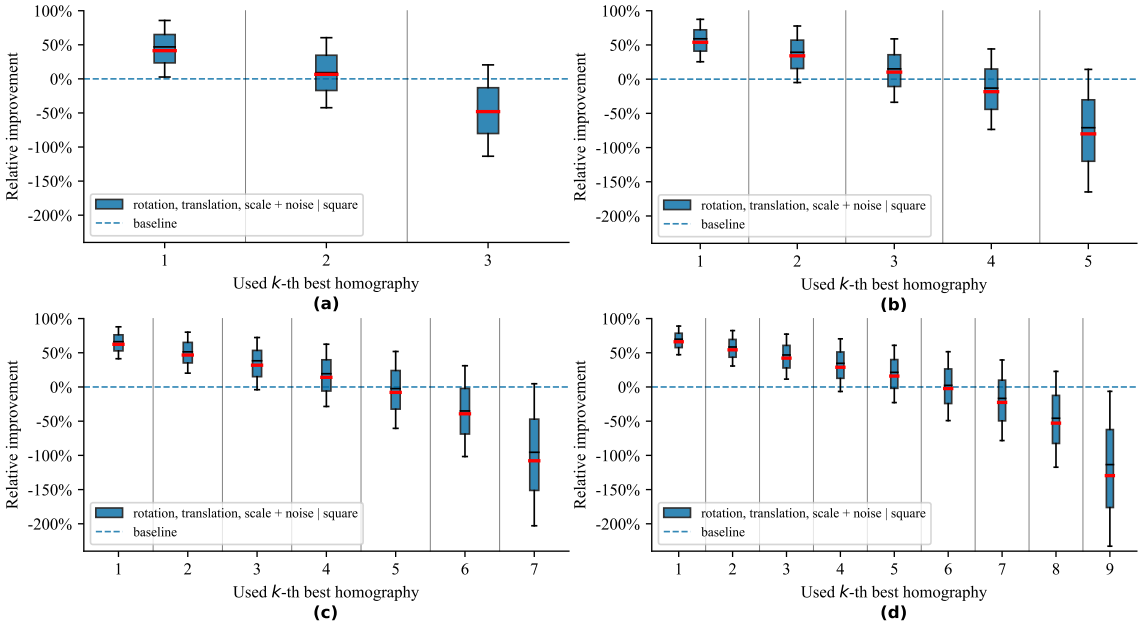


Fig. 5.14: Influence of different number of markers on reprojection error. We experimented with **(a)** three, **(b)** five, **(c)** seven, and **(d)** nine markers.

5.6 Conclusion

In this homography-related subpart of our object tracking research, 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.

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 (*e.g.*, SIFT [**?**]) 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.

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. 5.11 - 5.14.

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

Generally speaking, all the improvements at individual ranking positions steadily decreased, reaching 0% improvement at around $^2/_3\ m$, 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` [133] reprojection error under the assumption of no prior preference of specific markers, hence the random marker selection.

A practical 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 very efficient, as it scales well with a quadratic complexity $\Theta\left(m^2\right)$ 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 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 Siamese Visual Object Tracking

In this chapter, we dive into our contributions to the field of Siamese-based VOT. This entire chapter follows a specific pattern. It consists of five 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 with a general description of a so-called `SiamMOT` tracker (Section 6.1). After that, we comment on why we think this model is suitable for traffic analysis (Section 6.2). Then, we propose our first enhancement to the `SiamMOT` tracker based on top of external neural network aimed at object ReID (Section 6.3). In the next section, we describe our experiments where we tried to incorporate learning metric embeddings into the end-to-end training pipeline of the `SiamMOT` tracker itself (Section 6.4). 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 of the `SiamMOT` tracker (Section 6.5).

6.1 Siamese Multi-Object Tracking Framework

This section is dedicated to the most important tracker we have encountered during our research, called `SiamMOT` [84]. Its importance stems from the fact that the majority of our experiments adopted this model. Further elaboration upon the underlying incentives for such a choice will be introduced in Section 6.2.

6.1.1 General Description

The authors of [84] tracker focused on improving online MOT. As far as their methodology was concerned, they employed a region-based approach [48] 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 [48] 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 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 [136]. 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 [137] 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 and long-lasting 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. 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 [138] was primarily used. It is a widely known library aimed at building deep neural network models while exploiting automatic differentiation.

6.1.2 Model Architecture

The two key aspects of the the `SiamMOT` architecture are `Faster R-CNN` [48] 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. 6.1, the input consists of two frames, namely $\boldsymbol{I}^t$ and $\boldsymbol{I}^{t+\delta}$, accompanied by a set of detected object instances $\boldsymbol{R}^t = \{R_1^t, R_2^t, \ldots, R_i^t, \ldots\}$ at time t. During the inference, the detection head produces a set of detected object instances $\boldsymbol{R}^{t+\delta}$ whilst the tracker's task is to propage the detections $\boldsymbol{R}^t$ to time $t+\delta$, and thus yielding the tracker output denoted as $\tilde{\boldsymbol{R}}^{t+\delta}$. 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 associaton of the tracker output $\tilde{R}_i^{t+\delta}$ with detections $\tilde{R}_i^{t+\delta}$ 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 $\boldsymbol{I}^{t+\delta}$ while exploiting a contextual

window spanning a fixed neighborhood of the object's location (*i.e.*, R_i^t) at frame $\boldsymbol{I}^t$. In order to define this step more formally, consider the following dependency:

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

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 frame over a specific search region that should be, by definition, greater than the exemplar region. Thus, the feature map $\boldsymbol{f}_{R_i}^t$ is extracted over the region R_i^t contained in the frame $\boldsymbol{I}^t$. Analogically, the feature map $\boldsymbol{f}_{S_i}^{t+\delta}$ is extracted over the search region $S_i^{t+\delta}$ delineated in the frame $\boldsymbol{I}^{t+\delta}$. The region $S_i^{t+\delta}$ is computed by simple 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. 6.1 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 6.1 is applied multiple times, *i.e.*, for each object detected in frame t, signified by $R_i^t \in \boldsymbol{R}^t$. However, from implementation's perspective, all these operations can run in parallel and thus the backbone features are computed only once, making the online tracking inference very efficient.

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 $\boldsymbol{R}^t$ and $\boldsymbol{R}^{t+\delta}$. Despite its efficacy, there are still issues to be addressed. The association will fail due to the following reasons:

1. if $\tilde{\boldsymbol{R}}^{t+\delta}$ does not match to the correct object instance in $\boldsymbol{R}^{t+\delta}$,
2. or if $v_i^{t+\delta}$ is low (below a specific threshold) for a visible object (person, vehicle, etc.) 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 [67].

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. It only served for having a baseline to overcome during evaluation.

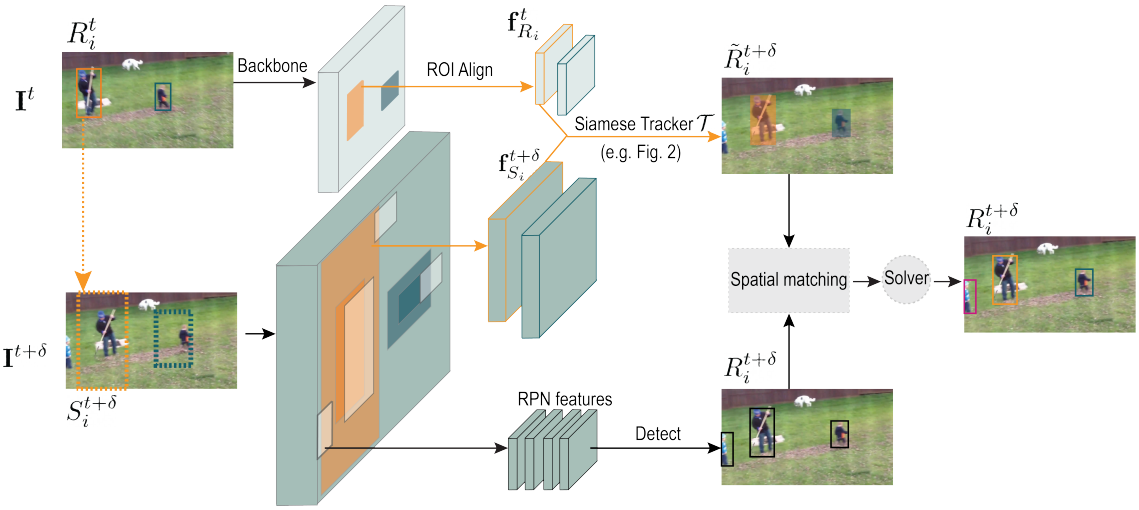


Fig. 6.1: The base architecture of the `SiamMOT` model. This tracker detects and associates object instances simultaneously. The Siamese tracker situated in the top branch serves the purpose of predicting the motion of objects across frames, thus facilitating the temporal linking of objects in an online 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. We have to emphasize that the spatial matching and solver blocks are only used during the inference and are not differentiable. Note that the feature map corresponding to the frame $\boldsymbol{I}^t$ 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*: [84])

Explicit Motion Modeling

The most fundamental aspect of Siamese trackers is the cross-correlation operator (Section 3.22 on page 43) to generate a pixel-level 2D response map. In `SiamMOT`, this operation correlates each location of the search feature map (belonging to the search region) $\boldsymbol{f}_{S_i}^{t+\delta}$ with the exemplar (target) feature map $\boldsymbol{f}_{R_i}^t$ to produce a response map

$$\boldsymbol{r}_i = \boldsymbol{f}_{S_i}^{t+\delta} \star \boldsymbol{f}_{R_i}^t. \tag{6.2}$$

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

Inspired by the Fully Convolutional One-Stage Object Detector (FCOS) [41] visual object detector, this tracker adopts fully-convolutional network ψ to facilitate instance detection using the response map $\boldsymbol{r}_i$. 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 $\boldsymbol{v}_i$. Every pixel of $\boldsymbol{v}_i$ is used as an indicator of the likelihood that this pixels falls within the location of the target object. Besides, a dense location map $\boldsymbol{p}_i$ 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

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

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

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

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

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

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 (Section **??** on page **??**).

Loss Function

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

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

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 [139, 140]. To address the class-balance problem in an effective way, the focal loss for classification [141] given by the term l_{focal} is employed, too. All ground-truth values are marked by the $*$ character. So,

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

and

$$\boldsymbol{p}_i^*(x, y) = \left[x - x_0^*, y - y_0^*, x_1^* - x, y_1^* - y\right], \tag{6.8}$$

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 [41] tracker, the regression loss l_{reg} is additionally modulated by computing the "centerness" for every location (see Fig. 6.2). This concept describes the amount of deviation of the pixel's location from the object center. The reason for adding this score was to

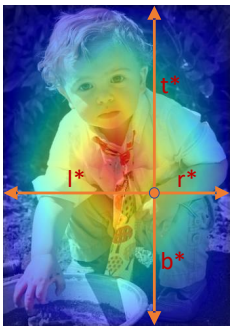


Fig. 6.2: Red, blue, and other colors denote 1, 0 and the values between them, respectively. Centerness is calculated as given by Equation 6.9 and decays from 1 to 0 as the location deviates from the center of the object. Best viewed in color. (*source*: [41])

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

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

6.1.3 Training and Inference Phases

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

$$\mathcal{L} = l_{rpn} + l_{detect} + l_{motion}, \tag{6.10}$$

where the l_{rpn} as well as the l_{detect} are standard RPN [48] and detection-subnetwork [142] losses, respectively. The l_{motion} loss is used to train the Siamese tracker.

For better understanding, we suggest the reader follow the diagram in Fig. 6.3. At inference, the well-established IoU-based NMS operation (Section 3.2.1 on page 21) 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 (thus, in total, it is used three times during the inference).

The spatial matching and the object identity association happen within the non-trainable module that is activated only during the inference. This algorithm is adequately referred to in the original paper 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. 6.4, 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

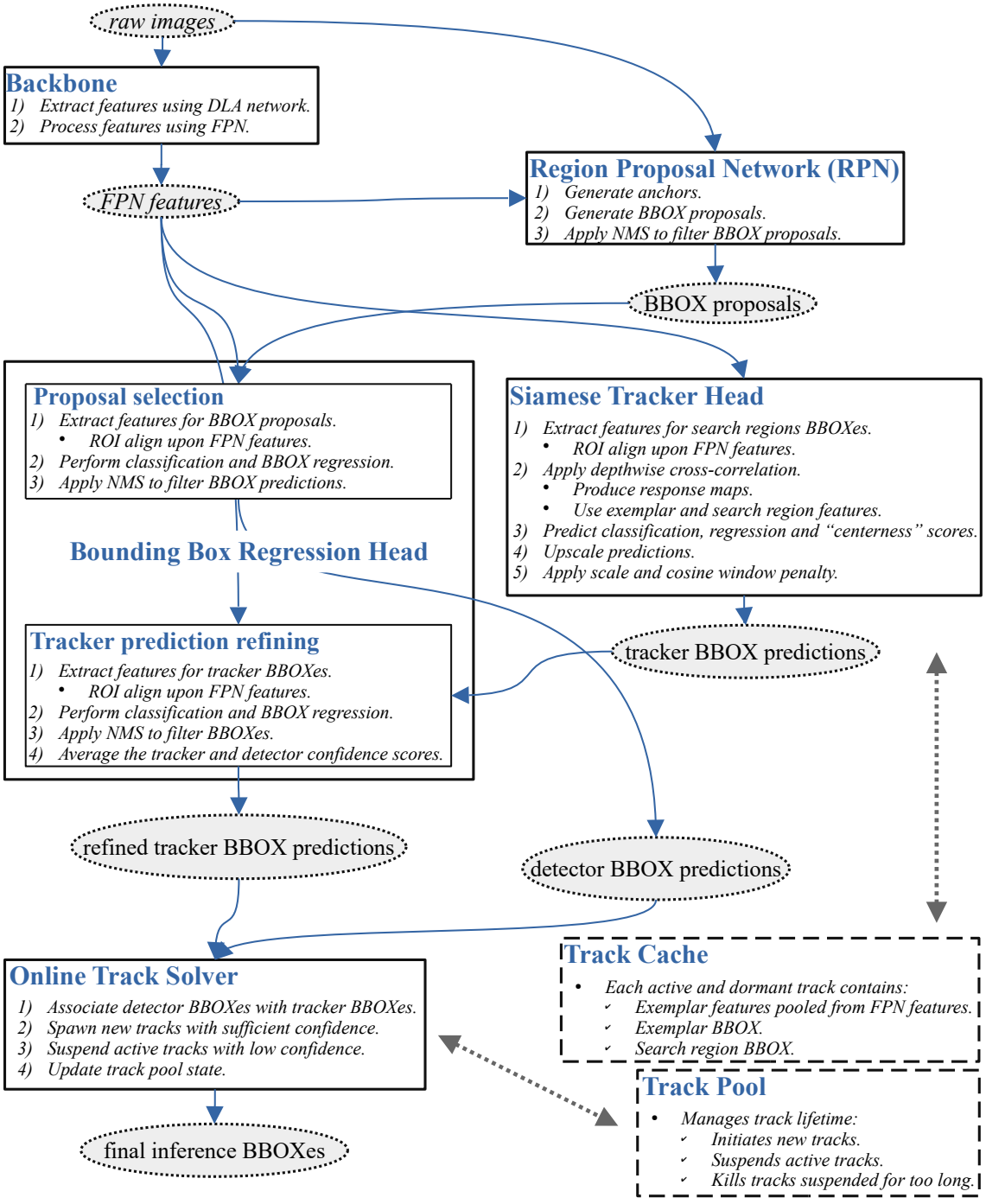


Fig. 6.3: Visualization of the inference pipeline in the `SiamMOT` architecture. The entire framework efficiently reuses as much information as possible, making it fast and accurate. Backbone features that are a result of intricate DLA and FPN processing are fed into the detector and the tracker. Please note that the predictions from the tracker are once again refined using the detector head. Two aforementioned heads function on top of backbone features through the lens of ROI align operations. During the inference phase, the "online solver" works only with the final BBOXes produced by the tracker and the object detector. It utilizes a simple caching mechanism to store the backbone features belonging to active or dormant objects. The decision-making regarding initialization, suspension, and complete removal of the track is performed within the "track pool" module.

specific threshold α, otherwise, this trajectory becomes dormant,

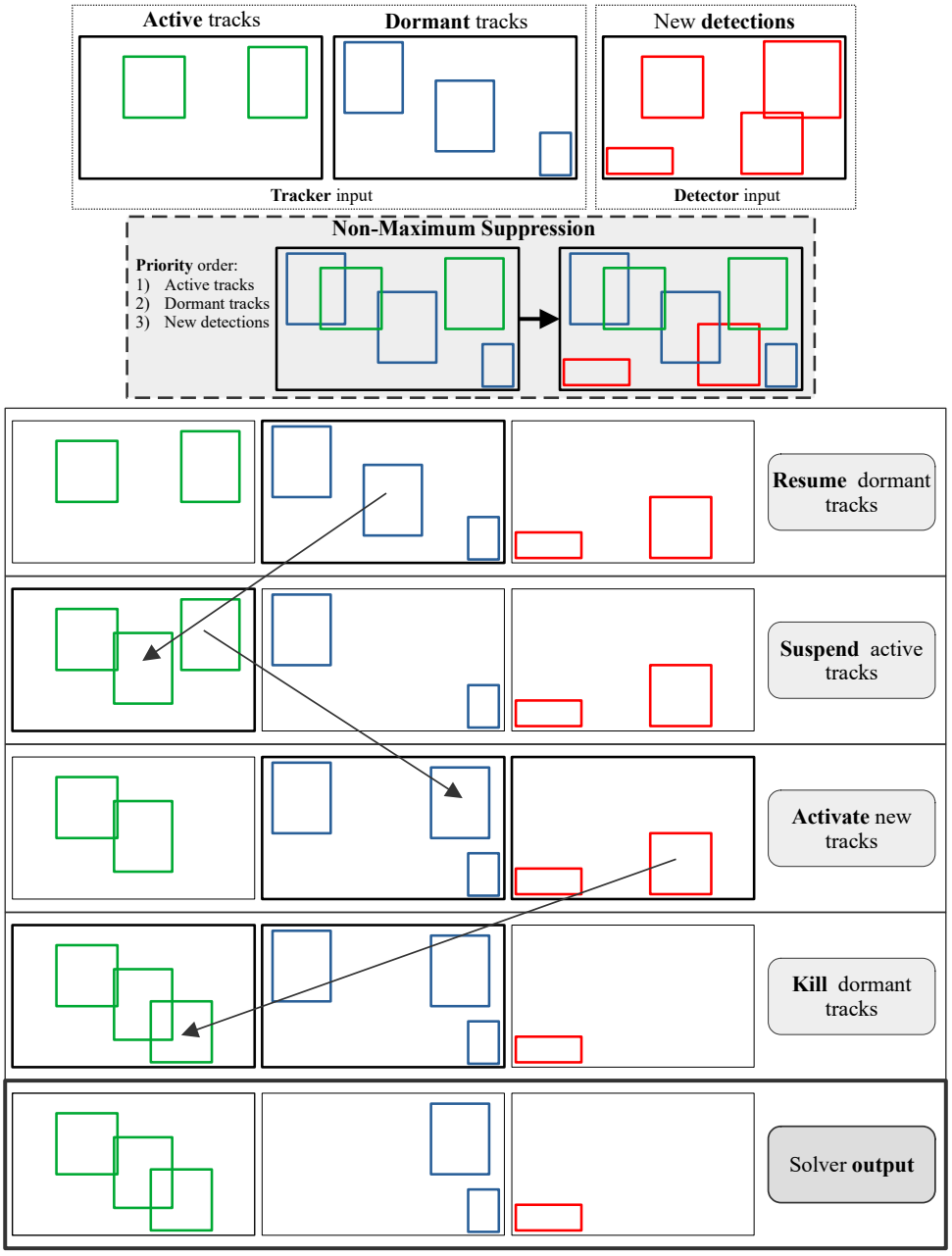


Fig. 6.4: Here we present the process which the tracker, as well as the detector predictions for the current frame, undergo to derive the final tracker output. At first, all the BBOXes are merged upon which a NMS algorithm is performed. The NMS algorithm preserves the priority that active tracks are 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.

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 also attempts to tackle short-term occlusion. It is one of the key incentives that led us to consider this architecture for our experiments, which is the model's ability to handle occlusion efficiently and trivially. 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.

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 [10, 143, 144, 145]. 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.

6.1.4 Implementation Details

6.1.5 Relevant Implementation-Related Remarks

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 GPU VRAM available to use sufficiently large minibatches. As far as our development experience with the `SiamMOT` model is concerned, we used the `NVidia` [146] `RTX 2080Ti` graphics cards which provides 4352 CUDA cores together with 11GB of GDDR6 memory. For training, we often exploited an existing backbone model pre-trained on ImageNet [36] dataset provided by the authors. Even though there is the entire model available pre-trained on `MS-COCO` [90] 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

Since the majority of our experiments involved expanding the model by adding additional heads, we often struggled with the amount of available GPU VRAM in order to preserve the reasonable size of minibatches. To this end, we also experimented with gradient

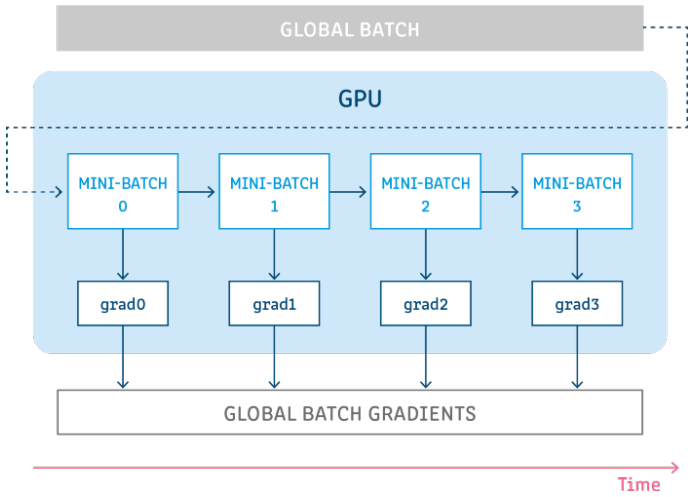


Fig. 6.5: Gradient accumulation is a mechanism that allows splitting the minibatch of samples used for training a neural network into several even smaller mini-batches of samples that will be processed sequentially. (*source*: [147])

accumulation. This technique allows the user to postpone the model weight updating after more minibatches have been processed (see Fig. 6.5). Therefore, the programmer may simulate using larger minibatches 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 variable values 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

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

When using gradient accumulation, the update step is modified as

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

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 [86] 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 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 works of [148] and [149].

(a) (b)

Fig. 6.6: Road scenes we encountered when working on the *Interreg SK-CZ* project.

6.2 Siamese Multi-Object Tracking and Traffic Analysis

6.2.1 Motivation

6.2.2 Dataset Selection

6.2.3 Implementation Remarks

We chose

6.3 Siamese Multi-Object Tracking and Re-identification

6.3.1 Motivation

The use of ReID has been emphasized numerous times during our preliminary research for this thesis. We conjectured that once full occlusion ensues, *i.e.*, the object completely disappears, the ReID mechanism could be adopted to recover the lost track since the object appears at the scene as a new one, thus a corresponding reasoning whether to assing a new or previously track identified is necessary. The presence of occlusion percolates traffic scenes all the time, 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. Under such circumstances, occlusion was inevitable. Such a setup inexorably lead 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 re-appeared for a miniscule amount of time at the other side of frame. The task of maintaining the same object identifier in such situations was burdensome and often impossible with high degree of precision.

6.3.2 Preliminary 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. Retrospectively, our experiments showed detrimental effects upon the tracker accuracy. In this section, we will elaborate further on why it is so in order to learn from such an

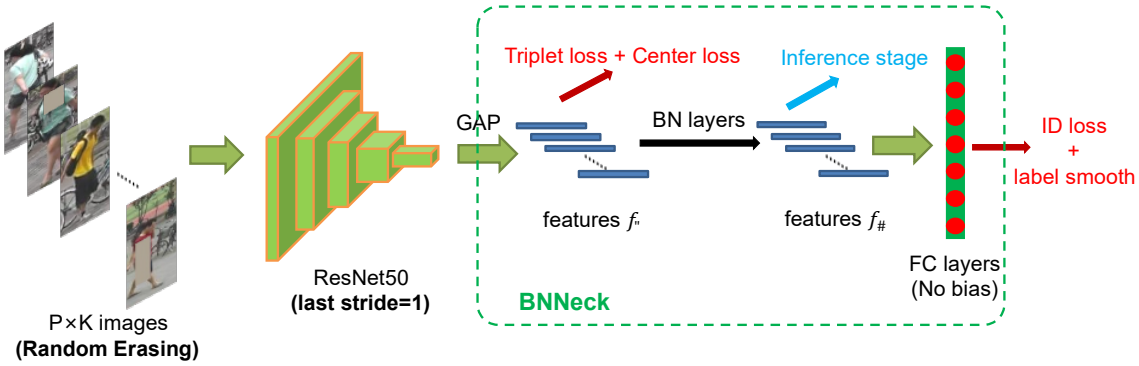


Fig. 6.7: A object ReID baseline which we used for our experiments. (*source*: [150])

experience as much as possible. Our surprise stems primarily from the fact that our thesis goals largely revolved around the application of ReID in object tracking. Therefore, finding that such approach does not yield the desired outcome raises multiple questions, namely the 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, specifically, the `SiamMOT` tracker?

3. Is the target use case in terms of datasets adequate to showcase the potential of ReID applied in the `SiamMOT` tracker?

In our dicussion, we will answer these questions. Despite our negative outcome, we still consider the performed research to be a contribution to the MOT area.

6.3.3 Proposed ReID-Enhanced Architecture

We adopted the object ReID architecture published in [150] by Luo *et al.*. The authors proposed a simple yet very robust framework for person ReID. We adopted this architecture (see Fig. 6.7) for vehicle ReID due to its simplicity accompanied with SOTA performance at the time of publishing.

6.3.4 Training Phase

6.3.5 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.

6.3.6 Experimental Evaluation

6.3.7 Discussion

6.4 Siamese Multi-Object Tracking and Feature Embedding

6.4.1 Motivation

One of our experiments involved an end-to-end training of the whole `SiamMOT` model together with a custom head supposed to create embeddings based on ROI-pooled backbone-extracted features for the object BBOX. The motivation 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 subsequent 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 [67], we noticed one work where the exemplar features were projected using global average pooling operation into an embedding space consisting of fewer dimensions [72]. The embedding vector was produced using the feature tensor that represents the kernel for the cross-correlation operation in 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 average pooling operation along the channel dimmension 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.* [72], these embedding vectors were exploited for 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 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.

6.4.2 Feature Embedding Head Architecture

As far as the vector embedding computation was concerned, we attached the embedding head (Table 6.1) into the backbone features but after the ROI-pooling operation. This ensured fixed tensor shapes and allowed us to process the very same feature that the object detector and Siamese tracker utilized, too. Simply put, for every proposal made for

layer	tensor shape	parameters no.
input	$[B, 256, 15, 15]$	0
conv 3×3	$[B, 256, 13, 13]$	589 824
ReLU	$[B, 256, 13, 13]$	0
conv 3×3	$[B, 512, 11, 11]$	1 179 648
ReLU	$[B, 512, 11, 11]$	0
flatten	$[B, 61952]$	0
linear	$[B, 1024]$	63 439 872
l_2-normalize	$[B, 1024]$	0
	total	65 209 344

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 at the end, we avoided learning biases throughout the whole network.

a particular frame, we looked at the delineated BBOX through the lens of ROI-pooling to extract backbone 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 negative mining.

6.4.3 Training Phase

The training phase was altered by adding another loss function to the sum of already existing three loss functions from the original model. In particular, the general `SiamMOT` loss function defined in Equation 6.10 was reformulated as

$$\mathcal{L} = l_{rpn} + l_{detect} + l_{motion} + l_{emb}. \tag{6.13}$$

The l_{emb} loss incorporated triplet loss (Equation 3.9 on page 26). As we discussed in Section 3.3 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 encountered harder and harder triplets to generate sufficient learning signal. To this end, we went for the semi-hard triplet mining strategy (Equation 3.11 on page 29). We implemented the entire mining algorithm followed by the loss computation in a GPU-only fashion for fast execution and easy intergration into the pipeline.

6.4.4 Inference Phase

Feature-based Non-Maximum Suppression

Salscheider [151] 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. At 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 3.2.1.). Nevertheless, here we repeat the same definitions for clarity. Let $\mathcal{B} = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\}$ be a set of n region proposals described by n BBOXes. Scores for each detection are contained in a set $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$, 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 $\mathcal{E} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$. Let $\mathcal{B}_{nms}$ be the set of filtered proposal instances from the set $\mathcal{B}$ produced using the 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. Now, the reasonining is different. 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 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($\mathcal{B}$, $\mathcal{S}$, $\mathcal{E}$, τ_{lower}, τ_{upper}, δ)
2: $\quad\mathcal{B}_{fnms} \leftarrow \emptyset$ ▷ initialize the output (filtered) set of region proposals
3: $\quad$**while** $\mathcal{B} \neq \emptyset$ **do** ▷ loop until all the proposals are processed
4: $\quad\quad m \leftarrow \underset{i \in \{1,2,\ldots,|S|\}}{arg\,max}\ \mathcal{S}$ ▷ find an index of a proposal with the highest score
5: $\quad\quad\mathcal{B} \leftarrow \mathcal{B} - \mathbf{b}_m$, $\mathcal{S} \leftarrow \mathcal{S} - s_m$, $\mathcal{E} \leftarrow \mathcal{E} - \mathbf{e}_m$ ▷ remove the proposal

$\mathcal{B}_{fnms} \leftarrow \mathcal{B}_{fnms} \cup \mathbf{b}_m$ ▷ save the proposal with the highest score
for $i \leftarrow 1$ to $|\mathcal{B}|$ **do** ▷ iterate through remaining proposals
if IOU$(\mathbf{b}_m, \mathbf{b}_i) \geq \tau_{\text{lower}}$ **then** ▷ above the lower-bound threshold
if IOU$(\mathbf{b}_m, \mathbf{b}_i) \geq \tau_{\text{upper}}$ **then** ▷ above the upper-bound threshold
$\mathcal{B} \leftarrow \mathcal{B} - \mathbf{b}_i$, $\mathcal{S} \leftarrow \mathcal{S} - s_i$, $\mathcal{E} \leftarrow \mathcal{E} - \mathbf{e}_i$ ▷ remove the proposal
else
if SIMILARITY$(\mathbf{e}_m, \mathbf{e}_i) \geq \delta$ **then** ▷ similarity above threshold
$\mathcal{B} \leftarrow \mathcal{B} - \mathbf{b}_i$, $\mathcal{S} \leftarrow \mathcal{S} - s_i$, $\mathcal{E} \leftarrow \mathcal{E} - \mathbf{e}_i$ ▷ remove the proposal
end if
end if
end if
end for
end while
return $\mathcal{B}_{fnms}$
end function

6.4.5 Experimental Evaluation

6.4.6 Discussion

We conjecture that our inability to improve the tracker performance was not particularly caused by the feature embedding itself. There is a recently published work Lu *et al.* [152], who introduced their `RetinaTrack` tracker. This framework exploited the base visual object detector called a `RetinaNet` [141] and then added principially the same head as we did for the purpose of producing feature embeddings that could be used for ReID. However, there are obvious differences between the two trackers in terms of how the inference phase is executed, and that is where we see the root cause of our failure.

6.5 Siamese Multi-Object Tracking and Attention

6.5.1 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 red light, but viewed under an angle somewhere in the range of $30-60$ degrees (see Fig. 6.8). 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` [25] 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

Fig. 6.8: An example of a situation where multiple vehicles are standing still on a cross-road. In this scenario, even though a slight degree of occlusion is necessary, 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.

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. 6.8, 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` [24] which altered the original formulation of predicting the rotated BBOX by use of ellipse fitting for even better accuracy.

However, as stated by the authors as well, 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 [153], especially spatial attention, which we found effective during our survey research [67]. 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 [154]. In what follows, we shall discuss these two methods (Section 6.5.2 and Section 6.5.3) as a foundation for our subsequent experiments that yielded a positive outcome.

6.5.2 Attention

An attention mechanism was first introduced by Vaswani *et al.* [153]. 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 will then map 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

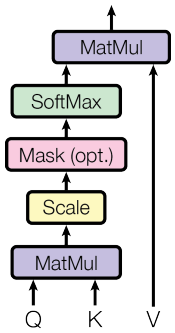


Fig. 6.9: An example of the input transformation by the scaled dot-product attention module. The pair of queries and keys is used to produce the probability distribution over the individual values for the final weighted sum. (*source*: [153])

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.

In abstract terms, 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 two other mappings.

For a more concrete demonstration, we shall 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 (see Fig. 6.9). 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 $\mathbf{Q}$. Analogically, keys and values can be also 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

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

The two most prominent variants of attention are the additive attention [155] 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 [153]. These ideas will be explored later.

6.5.3 Deformable Convolutional Neural Networks

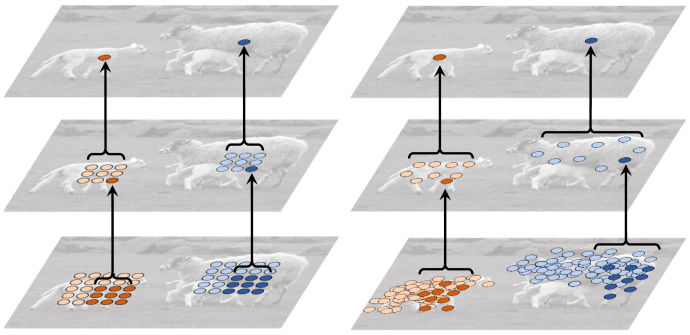
Deformable Convolutional Neural Networks (DCNNs) [154] are gaining popularity and are being applied to numerous sophisticated computer vision tasks, *e.g.*, object segmentation (dense predictions) and object detection (semi-dense predictions). Since object tracking revolves around the same requirements in terms of pixel-wise precision, we contemplated using this advancement, too.

Although CNNs (Section 3.1.2 on page 19) 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 in practice.

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) [156]. It warps the feature map via a global parametric transformation such as affine transformation. In the real of convolutional operations, there is the atrous convolution operation [157] 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 [158] 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 (see Fig. 6.10) comes into place, discussed next.

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

$$\mathcal{R} = \{(-1,-1), (-1,0), \ldots, (0,1), (1,1)\}. \tag{6.15}$$



(a) standard convolution (b) deformable convolution

Fig. 6.10: 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*: [154])

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

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

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 $\{\Delta\mathbf{p}_n \mid n = 1, \ldots, |\mathcal{R}|\}$ (see Fig. 6.11). Thus, Equation 6.16 is reformulated as

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

However, one 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

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

in which $\mathbf{q}$ enumerates all integral locations and $G\left(\cdot\right)$ 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 (see Fig. 6.12).

The original paper [154], where DCNNs were introduced showed, that learning dense spatial transformation in using deep learning by use of CNNs or sophisticated vision tasks 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.

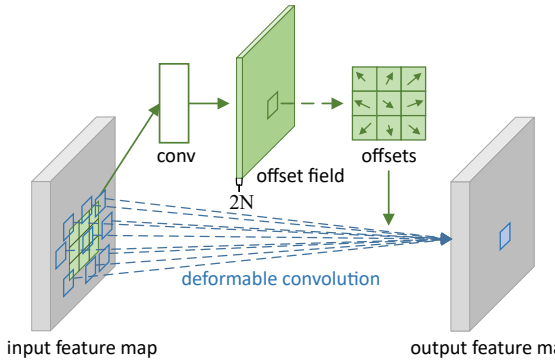


Fig. 6.11: Illustration of a 3×3 deformable convolution operation. Unlike the standard convolution operation used in neural networks, this one employs one additional step of predicting variable offsets instead of using a fixed rectangular grid. (*source*: [154])

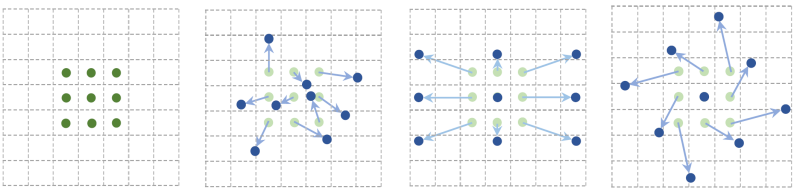


Fig. 6.12: Deformable convolution is effective at learning appropriate sampling locations reflecting the underlying transformation. **(a)** shows the regular sampling grid of a standard convolution; **(b)** is an example of irregularly deformed sampling region; **(c)** and **(d)** represent an expected pattern corresponding to scaling and rotation operations, respectively. (*source*: [154])

6.5.4 Modulated Deformable Convolutional Neural Networks

The original paper by Zhu *et al.* [159] aptly described their contribution in the headline that contained the words "mode deformable, better results". Because of this, we shall 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 6.17. Then, the modified equation becomes

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

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 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 $sigmoid\,(0) = 1/2$ at the beginning.

6.5.5 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 domains. Such experiments resulted in slight improvements for the reasons outlined in the motivation section. To support that our proposals were based on properly identified causes, we found a recently published work that demonstrated the effectiveness and potential use of our ideas, too.

Yu *et al.* [160] 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 well.

Considering their contribution and promising outcomes for the single object tracking demonstrated on the `SiamRPN` framework (first discussed in Section 3.6.3 on page 40), we decided to implement their proposed module into the `SiamMOT` tracker as described in their paper. However, with the prospect of greater improvement, we adopted DCNNs, instead of pure DCNNs. The rationale behind this is the following. 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 template (exemplar) and search branch independently. This operation can be easily achieved since exemplar and search tensors only differ in width and height, which does not pose a problem to convolution at all (fully-convolutional networks exploit this property). The following description of the self-attention computation conforms to the established attention principles regarding “queries,

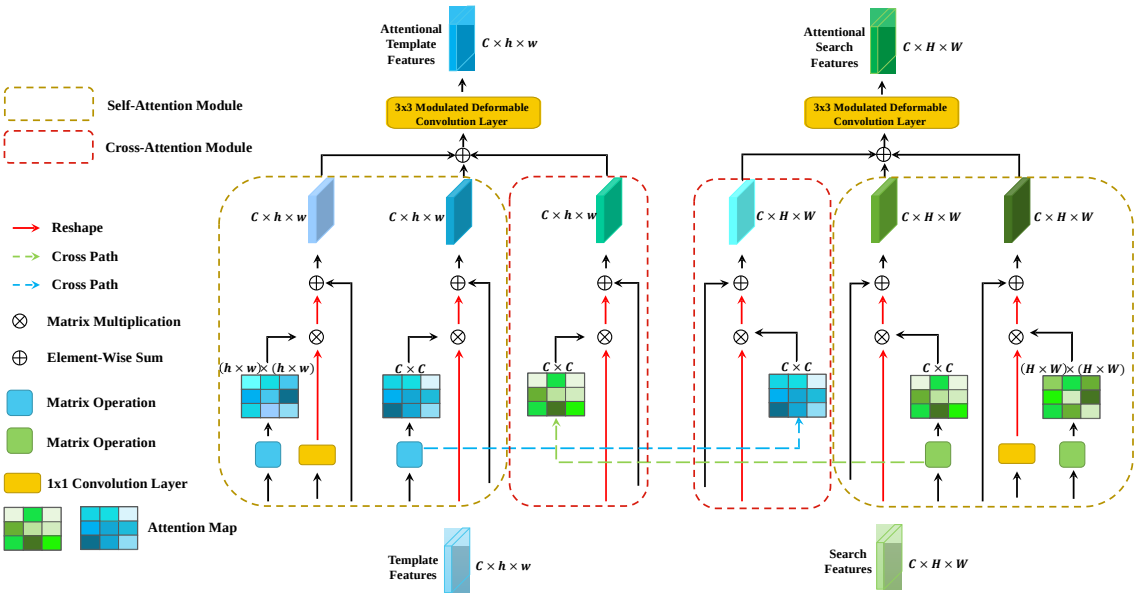


Fig. 6.13: The DSA extension introduces two sub-modules for both template (exemplar) and search branch. The self-attention is further divided into two operations, namely the spatial and channel attention. The very same attention network is used to process both features independently. Notice how the channel 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. In this figure, we modified the deformable convolution operation to include the modulated version, which is our modification compared to the original formulation. (*source*: [160])

keys and values" introduced in Section 6.5.2. For better understanding of the computation, see the diagram in Fig. 6.13.

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, first 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 $\boldsymbol{A}_S^S \in \mathbb{R}^{N \times N}$ is then produced via matrix multiplication and a column-wise softmax operation as

$$\boldsymbol{A}_S^S = softmax_{col}\left(\bar{\mathbf{Q}}^T \bar{\mathbf{K}}\right). \tag{6.20}$$

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 and keys as well as computed the values. We may proceeed further to the weighted selection from the values and to incorporate the attention into the features as follows

$$\bar{\boldsymbol{X}}_S^S = \alpha \bar{\mathbf{V}} \boldsymbol{A}_S^S + \bar{\mathbf{X}}, \tag{6.21}$$

where α is a learnable scalar parameter, and $\bar{\boldsymbol{X}}_S^S \in \mathbb{R}^{C\times N}$. The outputs $\bar{\boldsymbol{X}}_S^S$ are then reshaped back to the original size, specifically $\boldsymbol{X}_S^S \in \mathbb{R}^{C\times H\times W}$. From our experience, the parameter α is very useful for training stabilization.

The corresponding channel self-attention $\boldsymbol{A}_C^S$ and the channel-wise attentional features $\boldsymbol{X}_C^S$ 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, $\boldsymbol{X}_S^S$ and $\boldsymbol{X}_C^S$, respectively.

Cross-Attention

Let $\mathbf{Z} \in \mathbb{R}^{C\times h}w$, $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$ 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 target 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 $\boldsymbol{A}^C \in \mathbb{R}^{C\times C}$ as

$$\boldsymbol{A}^C = softmax_{row}\left(\bar{\mathbf{Z}}\bar{\mathbf{Z}}^T\right). \tag{6.22}$$

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

$$\bar{\boldsymbol{X}}^C = \gamma \boldsymbol{A}^C \bar{\mathbf{X}} + \bar{\mathbf{X}}, \tag{6.23}$$

where γ is a learnable scalar parameter. The merged features $\bar{\boldsymbol{X}}^C$, once again, have to be reshaped, so the features $\boldsymbol{X}^C \in \mathbb{R}^{C\times H\times W}$ 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 (see Fig. 6.13).

Once the attention is applied to tracking, the corresponding response map is altered as expected. Discriminative power of the tracker is enhanced by appropriate suppression of the semantic background. As Fig. 6.14 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 deformable convolution, in our case, modulated deformable convolution. The resulting features the shape of which is identical to the input shape are used to compute the response

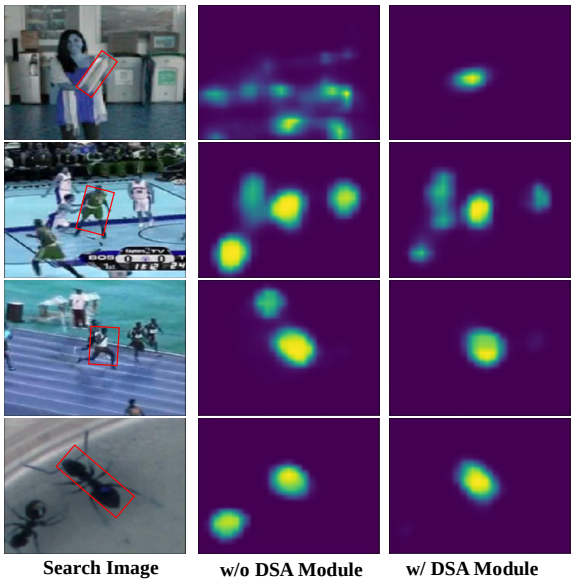


Fig. 6.14: Visualization of response (confidence) maps. The first column represents the search image, the second column represents the activation levels without the DSA module, whereas the third column clearly demonstrates the improved target-background discirminability in the computed attentional features. (*source*: [160])

map. As a result, this extension can be easily integrated into an existing pipeline thanks to its ability to preserve tensor shapes.

6.5.6 Experimental Evaluation

6.5.7 Discussion

Chapter 7

Discussion and Conclusion

7.1 Discussion

7.2 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 [67] cover such approaches explicitly in one section, our most relevant contribution to object tracking exploits the attention as well.

Bibliography

[1] David A. Forsyth and Jean Ponce. *Computer Vision - A Modern Approach, Second Edition.* Pitman, 2012. ISBN 978-0-273-76414-4. 1–791 pp.

[2] Anand Jalal and Vrijendra Singh. The State-of-the-Art in Visual Object Tracking. *Informatica (Slovenia)*, 36:227–248, 01 2012. ISSN 03505596.

[3] VOT challenge. https://www.webvotchallenge.net/. Accessed: 2020-04-20.

[4] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Object Tracking Benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37:1, 09 2015.

[5] MOT challenge. https://webmotchallenge.net/. Accessed: 2020-04-20.

[6] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564–577, 2003.

[7] Christopher Richard Wren, Ali Azarbayejani, Trevor Darrell, and Alex Paul Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on pattern analysis and machine intelligence*, 19(7):780–785, 1997.

[8] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering, Transactions of the ASME*, 82(1):35–45, 1960. ISSN 1528901X.

[9] H. W. Kuhn and Bryn Yaw. The Hungarian method for the assignment problem. *Naval Res. Logist. Quart*, pages 83–97, 1955.

[10] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. *Proceedings - International Conference on Image Processing, ICIP*, 2016-Augus:3464–3468, 2016. ISBN 9781467399616. ISSN 15224880.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 2:1097–1105, 2012. ISBN 9781627480031. ISSN 10495258.

[12] Zheng Tang, Milind Naphade, Ming Yu Liu, Xiaodong Yang, Stan Birchfield, et al. Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June: 8789–8798, 2019. ISBN 9781728132938. ISSN 10636919.

[13] Laura Leal-Taixé, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, et al. Tracking the Trackers: An Analysis of the State of the Art in Multiple Object Tracking, 2017.

[14] Chelsea Finn, Tianhe Yu, Tianhao Zhang, Pieter Abbeel, and Sergey Levine. One-Shot Visual Imitation Learning via Meta-Learning. *CoRR*, abs/1709.04905, 2017. URL `http://arxiv.org/abs/1709.04905`.

[15] Xue Bin Peng, Angjoo Kanazawa, Jitendra Malik, Pieter Abbeel, and Sergey Levine. SFV: Reinforcement Learning of Physical Skills from Videos. *ACM Trans. Graph.*, 37(6), November 2018.

[16] Pan Jiyan and Hu Bo. Robust occlusion handling in object tracking. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (June 2007), 2007. ISBN 1424411807. ISSN 10636919.

[17] Pierre F Gabriel, Jacques G Verly, Justus H Piater, and André Genon. The state of the art in multiple object tracking under occlusion in video sequences. In *Advanced Concepts for Intelligent Vision Systems*, pages 166–173, 2003.

[18] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June:815–823, 2015. ISBN 9781467369640. ISSN 10636919.

[19] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. DeepFace: Closing the gap to human-level performance in face verification. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014. ISBN 9781479951178. ISSN 10636919.

[20] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2:1735–1742, 2006. ISBN 0769525970. ISSN 10636919.

[21] Ratnesh Kuma, Edwin Weill, Farzin Aghdasi, and Parthasarathy Sriram. Vehicle Re-identification: An Efficient Baseline Using Triplet Embedding. *Proceedings of the International Joint Conference on Neural Networks*, 2019-July, 2019. ISBN 9781728119854.

[22] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[23] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In Defense of the Triplet Loss for Person Re-Identification, 2017.

[24] Bao Xin Chen and John K. Tsotsos. Fast Visual Object Tracking with Rotated Bounding Boxes, 2019.

[25] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip H.S. Torr. Fast online object tracking and segmentation: A unifying approach. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019-June:1328–1338, 2019. ISBN 9781728132938. ISSN 10636919.

[26] Jimei Yang, Brian Price, Scott Cohen, Honglak Lee, and Ming Hsuan Yang. Object contour detection with a fully convolutional encoder-decoder network. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem:193–202, 2016. ISBN 9781467388504. ISSN 10636919.

[27] Dongyan Guo, Jun Wang, Ying Cui, Zhenhua Wang, and Shengyong Chen. SiamCAR: Siamese Fully Convolutional Classification and Regression for Visual Tracking, 2019.

[28] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High Performance Visual Tracking with Siamese Region Proposal Network. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 8971–8980, 2018. ISBN 9781538664209. ISSN 10636919.

[29] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, et al. SSD: Single shot multibox detector. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9905 LNCS:21–37, 2016. ISBN 9783319464473. ISSN 16113349.

[30] Ran Tao, Efstratios Gavves, and Arnold W.M. Smeulders. Siamese instance search for tracking. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-December:1420–1429, 2016. ISBN 9781467388504. ISSN 10636919.

[31] Frank F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65 6:386–408, 1958.

[32] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[33] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning Internal Representations by Error Propagation. In David E. Rumelhart and James L. Mcclelland, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*, pages 318–362. MIT Press, Cambridge, MA, 1986.

[34] Fully Connected Neural Network Architecture. https://ch.mathworks.com/fr/solutions/deep-learning/convolutional-neural-network.html. Accessed: 2020-04-20.

[35] François Chollet. *Deep Learning with Python* . Manning, November 2017. ISBN 9781617294433.

[36] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, et al. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[37] Convolutional Neural Network Architecture. https://ch.mathworks.com/fr/solutions/deep-learning/convolutional-neural-network.html. Accessed: 2020-04-20.

[38] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem:779–788, 2016. ISBN 9781467388504. ISSN 10636919.

[39] C. Cortes and V. Vapnik. Support Vector Networks. *Machine Learning*, pages 273–297, 1995.

[40] R K McConnell. Method of and apparatus for pattern recognition. 1 1986.

[41] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-Octob: 9626–9635, 2019. ISBN 9781728148038. ISSN 15505499.

[42] Jan Hosang, Rodrigo Benenson, and Bernt Schiele. Learning non-maximum suppression. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-January: 6469–6477, 2017. ISBN 9781538604571.

[43] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S. Davis. Soft-NMS - Improving Object Detection with One Line of Code. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-October:5562–5570, 2017. ISBN 9781538610329. ISSN 15505499.

[44] Joseph Redmon and Ali Farhadi. YOLO9000: Better, faster, stronger. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-January:6517–6525, 2017. ISBN 9781538604571.

[45] Joseph Redmon and Ali Farhadi. YOLOv3: An Incremental Improvement, 2018.

[46] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. YOLOv4: Optimal Speed and Accuracy of Object Detection, 2020.

[47] Alexander Wong, Mahmoud Famouri, Mohammad Javad Shafiee, Francis Li, Brendan Chwyl, et al. YOLO Nano: a Highly Compact You Only Look Once Convolutional Neural Network for Object Detection. *2019 Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS Edition (EMC2-NIPS)*, pages 22–25, 2019.

[48] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017. ISSN 01628828.

[49] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, et al. Speed/accuracy trade-offs for modern convolutional object detectors. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:3296–3305, 2017. ISBN 9781538604571.

[50] Gregory R. Koch. Siamese Neural Networks for One-Shot Image Recognition. 2015.

[51] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep Face Recognition. (Section 3): 41.1–41.12, 2015.

[52] R. Manmatha, Chao Yuan Wu, Alexander J. Smola, and Philipp Krahenbuhl. Sampling Matters in Deep Embedding Learning. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-Octob:2859–2867, 2017. ISBN 9781538610329. ISSN 15505499.

[53] Jesse Davis and Mark H. Goadrich. The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd international conference on Machine learning*, 2006.

[54] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

[55] Gerard Salton and Michael McGill. *Introduction to modern information retrieval*. McGraw-Hill, New York, NY, 1983.

[56] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008(1):246309, May 18, 2008. ISSN 1687-5281. URL `https://doi.org/10.1155/2008/246309`.

[57] Longyin Wen, Dawei Du, Zhaowei Cai, Zhen Lei, Ming-Ching Chang, et al. UA-DETRAC: A New Benchmark and Protocol for Multi-Object Detection and Tracking. *Computer Vision and Image Understanding*, 2020.

[58] webpymotmetrics. `https://github.com/cheind/py-motmetrics`. Accessed: 2020-04-20.

[59] James R. Munkres. Algorithms for the Assignment and Transportation Problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38, March 1957.

[60] David Held, Sebastian Thrun, and Silvio Savarese. Learning to Track at 100 FPS with Deep. *Computer Vision ECCV 2016 Lecture Notes in Computer Science*, pages 749–765, 2016. URL `http://davheld.github.io/GOTURN/GOTURN.html`.

[61] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, pages 1–10, 2014.

[62] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, et al. Going deeper with convolutions. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June:1–9, 2015. ISBN 9781467369640. ISSN 10636919.

[63] Lijun Wang, Wanli Ouyang, Xiaogang Wang, and Huchuan Lu. Visual tracking with fully convolutional networks. *Proceedings of the IEEE International Conference on Computer Vision*, 2015 Inter: 3119–3127, 2015. ISBN 9781467383912. ISSN 15505499.

[64] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pages 1–14, 2015.

[65] Andrew Y Ng and Michael I Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems*, pages 841–848, 2002.

[66] Luca Bertinetto, Jack Valmadre, João F. Henriques, Andrea Vedaldi, and Philip H.S. Torr. Fully-convolutional siamese networks for object tracking. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9914 LNCS: 850–865, 2016. ISBN 9783319488806. ISSN 16113349.

[67] Milan Ondraovi and Peter Tarábek. Siamese Visual Object Tracking: A Survey. *IEEE Access*, 9: 110149–110172, 2021.

[68] Anfeng He, Chong Luo, Xinmei Tian, and Wenjun Zeng. A Twofold Siamese Network for Real-Time Object Tracking. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 4834–4843, 2018. ISBN 9781538664209. ISSN 10636919.

[69] Luca Bertinetto, João F. Henriques, Jack Valmadre, Philip H.S. Torr, and Andrea Vedaldi. Learning feed-forward one-shot learners. *Advances in Neural Information Processing Systems*, (Nips):523–531, 2016. ISSN 10495258.

[70] Roman Pflugfelder. An in-depth analysis of visual tracking with siamese neural networks, 2018.

[71] Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, and Weiming Hu. Distractor-aware siamese networks for visual object tracking, 2018.

[72] Daqun Li and Yi Yu. Foreground information guidance for siamese visual tracking. *IEEE Access*, 8: 55905–55914, 2020.

[73] Daqun Li, Yi Yu, and Xiaolin Chen. Object tracking framework with siamese network and re-detection mechanism. *EURASIP Journal on Wireless Communications and Networking*, 2019(1): 261, Nov 29, 2019. ISSN 1687-1499. URL `https://doi.org/10.1186/s13638-019-1579-x`.

[74] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks, 2018.

[75] Seyed Mojtaba Marvasti-Zadeh, Li Cheng, Hossein Ghanei-Yakhdan, and Shohreh Kasaei. Deep learning for visual tracking: A comprehensive survey. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–26, 2021.

[76] Qiang Wang, Zhu Teng, Junliang Xing, Jin Gao, Weiming Hu, and Stephen Maybank. Learning attentions: Residual attentional siamese network for high performance online visual tracking. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4854–4863, 2018.

[77] Z. Liang and J. Shen. Local semantic siamese networks for fast tracking. *IEEE Transactions on Image Processing*, 29:3351–3364, 2020.

[78] Hwann-Tzong Chen, Horng-Horng Lin, and Tyng-Luh Liu. Multi-object tracking using dynamical graph matching. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 2, pages II–II, 2001.

[79] Ioannis Papakis, Abhijit Sarkar, and Anuj Karpatne. Gcnnmatch: Graph convolutional neural networks for multi-object tracking via sinkhorn normalization, 2021.

[80] Bonan Cuan, Khalid Idrissi, and Christonhe Garcia. Deep siamese network for multiple object tracking. In *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, 2018.

[81] Bing Shuai, Andrew G Berneshawi, Davide Modolo, and Joseph Tighe. Multi-object tracking with siamese track-rcnn. *arXiv preprint arXiv:2004.07786*, 2020.

[82] Lorenzo Vaquero, Manuel Mucientes, and Víctor M. Brea. Siammt: Real-time arbitrary multi-object tracking. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 707–714, 2021.

[83] Sangyun Lee and Euntai Kim. Multiple object tracking via feature pyramid siamese networks. *IEEE Access*, 7:8181–8194, 2019.

[84] Bing Shuai, Andrew Berneshawi, Xinyu Li, Davide Modolo, and Joseph Tighe. Siammot: Siamese multi-object tracking, 2021.

[85] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[86] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015.

[87] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection, 2017.

[88] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation, 2019.

[89] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2018.

[90] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, et al. Microsoft COCO: Common objects in context. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8693 LNCS(PART 5):740–755, 2014. ISSN 16113349.

[91] MS COCO. `https://cocodataset.org/`. Accessed: 2020-04-20.

[92] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[93] KITTI Object Detection. http://www.cvlibs.net/datasets/kitti/eval_object.php?obj_benchmark=2d. Accessed: 2020-04-20.

[94] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June-2015:3973–3981, 2015. ISBN 9781467369640. ISSN 10636919.

[95] CompCars. http://ai.stanford.edu/~jkrause/cars/car_dataset.html. Accessed: 2020-04-20.

[96] Hongye Liu, Yonghong Tian, Yaowei Wang, Lu Pang, and Tiejun Huang. Deep Relative Distance Learning: Tell the Difference Between Similar Vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2167–2175, 2016.

[97] PKU VehicleID. https://www.pkuml.org/resources/pku-vehicleid.html. Accessed: 2020-04-20.

[98] VeRI-776. https://vehiclereid.github.io/VeRi/. Accessed: 2020-04-20.

[99] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma. PROVID: Progressive and Multimodal Vehicle Reidentification for Large-Scale Urban Surveillance. *IEEE Transactions on Multimedia*, 20(3):645–658, 2018. ISSN 15209210.

[100] Lianghua Huang, Xin Zhao, and Kaiqi Huang. GOT-10k: A Large High-Diversity Benchmark for Generic Object Tracking in the Wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1562–1577, 2021.

[101] GOT-10k. http://got-10k.aitestunion.com/. Accessed: 2020-04-20.

[102] VOT 2019. https://www.webvotchallenge.net/vot2019/dataset.html. Accessed: 2020-04-20.

[103] Matej Kristan, Jiri Matas, Ales Leonardis, Michael Felsberg, Roman Pflugfelder, et al. The Seventh Visual Object Tracking VOT2019 Challenge Results, 2019.

[104] KITTI Object Tracking. http://www.cvlibs.net/datasets/kitti/eval_tracking.php. Accessed: 2020-04-20.

[105] UA-DETRAC dataset. https://detrac-db.rit.albany.edu/. Accessed: 2020-04-20.

[106] Milan Ondrašovič and Peter Tarábek. Homography ranking based on multiple groups of point correspondences. *Sensors*, 21(17), 2021. ISSN 1424-8220. URL https://www.mdpi.com/1424-8220/21/17/5752.

[107] Sensors. https://www.mdpi.com/journal/sensors. Accessed: 2020-04-20.

[108] Milan Ondrašovič and Peter Tarábek. Foundations for homography estimation in presence of redundant point correspondencies. In *Mathematics in science and technologies - proceedings of the MIST conference 2020*, number 1. vydanie, pages 52–57, 2020.

[109] MiST. https://mist-klak.webnode.sk/. Accessed: 2020-04-20.

[110] A Geetha Kiran and S Murali. Automatic rectification of perspective distortion from a single image using plane homography. *J. Comput. Sci. Appl*, 3(5):47–58, 2013.

[111] Alexandre Bousaid, Theodoros Theodoridis, Samia Nefti-Meziani, and Steve Davis. Perspective distortion modeling for image measurements. *IEEE Access*, 8:15322–15331, 2020.

[112] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, USA, 2 edition, 2003. ISBN 0521540518.

[113] Richard I Hartley. In defense of the eight-point algorithm. *IEEE Transactions on pattern analysis and machine intelligence*, 19(6):580–593, 1997.

[114] Shijian Lu, Ben M Chen, and Chi Chung Ko. Perspective rectification of document images using fuzzy set and morphological operations. *Image and Vision Computing*, 23(5):541–553, 2005.

[115] Ligang Miao and Silong Peng. Perspective rectification of document images based on morphology. In *2006 International Conference on Computational Intelligence and Security*, volume 2, pages 1805–1808. IEEE, 2006.

[116] Ebtsam Adel, Mohammed Elmogy, and Hazem Elbakry. Image stitching based on feature extraction techniques: a survey. *International Journal of Computer Applications*, 99(6):1–8, 2014.

[117] Junhong Gao, Seon Joo Kim, and Michael S Brown. Constructing image panoramas using dual-homography warping. In *CVPR 2011*, pages 49–56. IEEE, 2011.

[118] W. X. Liu and T. Chin. Smooth globally warp locally: Video stabilization using homography fields. In *2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE, 2015.

[119] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334, 2000.

[120] Damien Mariyanayagam, Pierre Gurdjos, Sylvie Chambon, Florent Brunet, and Vincent Charvillat. Pose estimation of a single circle using default intrinsic calibration. *CoRR*, abs/1804.04922, 2018. URL `http://arxiv.org/abs/1804.04922`.

[121] Jon Arróspide, Luis Salgado, Marcos Nieto, and Raúl Mohedano. Homography-based ground plane detection using a single on-board camera. *IET Intelligent Transport Systems*, 4(2):149–160, 2010.

[122] Lin-Bo Luo, In-Sung Koh, Kyeong-Yuk Min, Jun Wang, and Jong-Wha Chong. Low-cost implementation of bird's-eye view system for camera-on-vehicle. In *2010 Digest of Technical Papers International Conference on Consumer Electronics (ICCE)*, pages 311–312. IEEE, 2010.

[123] Biswajit Bose and Eric Grimson. Ground plane rectification by tracking moving objects. In *IEEE International Workshop on Visual Surveillance and PETS*, 2004.

[124] Miaohui Zhang, Yandong Hou, and Zhentao Hu. Accurate object tracking based on homography matrix. In *2012 International Conference on Computer Science and Service System*, pages 2310–2312, 2012.

[125] Christopher Mei, Selim Benhimane, Ezio Malis, and Patrick Rives. Efficient homography-based tracking and 3-d reconstruction for single-viewpoint sensors. *Robotics, IEEE Transactions on*, 24: 1352–1364, 01 2009.

[126] Homography - basic concepts. `http://man.hubwiz.com/docset/OpenCV.docset/Contents/Resources/Documents/d9/dab/tutorial_homography.html`. Accessed: 2020-04-20.

[127] Yueqiang Zhang, Langming Zhou, Haibo Liu, and Yang Shang. A flexible online camera calibration using line segments. *Journal of Sensors*, 2016, Jan 06, 2016. ISSN 1687-725X. URL `https://doi.org/10.1155/2016/2802343`.

[128] Valentín Osuna-Enciso, Erik Cuevas, Diego Oliva, Virgilio Zúñiga, Marco Pérez-Cisneros, and Daniel Zaldívar. A multiobjective approach to homography estimation. *Computational Intelligence and Neuroscience*, 2016:3629174, Dec 28, 2015. ISSN 1687-5265. URL `https://doi.org/10.1155/2016/3629174`.

[129] Wei Mou, Han Wang, Gerald Seet, and Lubing Zhou. Robust homography estimation based on non-linear least squares optimization. In *2013 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 372–377. IEEE, 2013.

[130] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381395, June 1981. ISSN 0001-0782. URL `https://doi.org/10.1145/358669.358692`.

[131] Daniel Barath and Levente Hajder. Novel ways to estimate homography from local affine transformations. In Nadia Magnenat-Thalmann, Paul Richard, Lars Linsen, Alexandru C. Telea, Sebastiano Battiato, Francisco H. Imai, and José Braz, editors, *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 3*, pages 434–445, 2016. URL `https://doi.org/10.5220/0005674904320443`.

[132] Graham Beck et al. *Planar Homography Estimation from Traffic Streams via Energy Functional Minimization*. PhD thesis, Johns Hopkins University, 2016.

[133] Gary Bradski and Adrian Kaehler. *Learning OpenCV: Computer vision with the OpenCV library.* " O'Reilly Media, Inc.", 2008.

[134] Y.I Abdel-Aziz, H.M. Karara, and Michael Hauck. Direct linear transformation from comparator coordinates into object space coordinates in close-range photogrammetry*. *Photogrammetric Engineering & Remote Sensing*, 81(2):103–107, 2015. ISSN 0099-1112. URL `https://www.sciencedirect.com/science/article/pii/S0099111215303086`.

[135] Homography Ranking. https://github.com/mondrasovic/homography_ranking. Accessed: 2020-04-20.

[136] SiamMOT - GitHub (original project). https://github.com/amazon-research/siam-mot. Accessed: 2020-04-20.

[137] SiamMOT - GitHub (forked project). https://github.com/mondrasovic/siam-mot. Accessed: 2020-04-20.

[138] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

[139] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Atom: Accurate tracking by overlap maximization, 2019.

[140] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. UnitBox: An advanced object detection network. *MM 2016 - Proceedings of the 2016 ACM Multimedia Conference*, pages 516–520, 2016. ISBN 9781450336031.

[141] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2018.

[142] Ross Girshick. Fast r-cnn, 2015.

[143] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric, 2017.

[144] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points, 2020.

[145] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019. URL `http://dx.doi.org/10.1109/ICCV.2019.00103`.

[146] webnvidia. https://www.webnvidia.com/en-us/. Accessed: 2020-04-20.

[147] Gradient Accumulation. https://ml-cheatsheet.readthedocs.io/en/latest/nn_concepts.html. Accessed: 2020-04-20.

[148] Yuxin Wu and Kaiming He. Group normalization, 2018.

[149] Xiao-Yun Zhou, Jiacheng Sun, Nanyang Ye, Xu Lan, Qijun Luo, Bo-Lin Lai, Pedro Esperanca, Guang-Zhong Yang, and Zhenguo Li. Batch group normalization, 2020.

[150] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification, 2019.

[151] Niels Ole Salscheider. Featurenms: Non-maximum suppression by learning feature embeddings, 2020.

[152] Zhichao Lu, Vivek Rathod, Ronny Votel, and Jonathan Huang. Retinatrack: Online single stage joint detection and tracking, 2020.

[153] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[154] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 764–773, 2017.

[155] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2016.

[156] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks, 2016.

[157] M. Holschneider, R. Kronland-Martinet, J. Morlet, and Ph. Tchamitchian. A real-time algorithm for signal analysis with the help of the wavelet transform. In Jean-Michel Combes, Alexander Grossmann, and Philippe Tchamitchian, editors, *Wavelets*, pages 286–297. Springer Berlin Heidelberg, Berlin, Heidelberg, 1990. ISBN 978-3-642-75988-8.

[158] Yunho Jeon and Junmo Kim. Active convolution: Learning the shape of convolution for image classification, 2017.

[159] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results, 2018.

[160] Yuechen Yu, Yilei Xiong, Weilin Huang, and Matthew R. Scott. Deformable siamese attention networks for visual object tracking, 2021.

Appendix

List of Author's Publications

Home Conferences

AFD001 **Foundations for homography estimation in presence of redundant point correspondencies / Milan Ondraovi, Peter Tarábek**.
In: **Mathematics in science and technologies**: proceedings of the MIST conference 2020: proceedings of the MIST conference 2020 / Katarína Bachratá, Katarína Jasenáková and Monika Smieková. - 1. vyd. - [S.l.] : [s.n.], 2020. - 70 s. [print]. - ISBN 9798648566026. - s. 52-57 [print].
[Ondraovi Milan (50%) - Tarábek Peter (50%)]

AFD002 **Object position estimation from a single moving camera / Milan Ondraovi, Peter Tarábek, Ondrej uch**.
In: **Information and digital technologies 2021**: proceedings of the international conference: proceedings of the international conference / [bez zostavovatea]. - 1. vyd. - Danvers : Institute of Electrical and Electronics Engineers, 2021. - 370 s. - ISBN 978-1-6654-3692-2. - s. 31-37. Zaradené v: SCOPUS
[Ondraovi Milan (50%) - Tarábek Peter (40%) - uch Ondrej (10%)]

International Impacted Journals

ADC001 **Homography ranking based on multiple groups of point correspondences / Milan Ondraovi** and **Peter Tarábek**.
In: **Sensors**. - Bazilej: Multidisciplinary Digital Publishing Institute. - [online, print]. - ISSN 1424-3210. - Ro. 21, . 17 (2021), s. [1-17] [online, print]. Zaradené v: Current Content Connect ; SCOPUS ; Web of Science Core Collection
[Ondraovi Milan (50%) - Tarábek Peter (50%)]

ADC002 **Siamese visual object tracking: a survey / Milan Ondraovi** 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). - Ro. 9 (2021), s. 110149-110172 [online]. Zaradené v: Current Content Connect ; SCOPUS ; Web of Science Core Collection
[Ondraovi Milan (50%) - Tarábek Peter (50%)]

There are already **two 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).