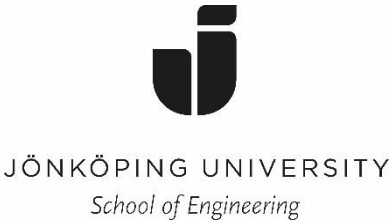

OSPREY: Person Re-Identification in the sport of Padel

Utilizing **O**ne-**S**hot **P**erson **Re**-identification with locall**y** aware transformers to improve tracking

PAPER WITHIN *Artificial Intelligence*
AUTHOR: *Måns J Svensson, Jim Hult*
TUTOR: *Patrick Gabrielsson*
JÖNKÖPING 06 2022

This exam work has been carried out at the School of Engineering in Jönköping in the subject area Artificial Intelligence. The work is a part of the two-year university diploma programme, of the Master of Science programme. The authors take full responsibility for opinions, conclusions and findings presented.

Examiner: Vladimir Tarasov

Supervisor: Patrick Gabrielsson

Scope: 30 credits (second cycle)

Date: 2022-06-14

Postadress:	Besöksadress:	Telefon:
Box 1026	Gjuterigatan 5	036-10 10 00 (vx)
551 11 Jönköping		

Abstract

This thesis is concerned with the topic of person re-identification. Many tracking algorithms today cannot keep track of players reentering the scene from different angles and times. Therefore, in this thesis, current literature is explored to gather information about the topic, and a current state-of-the-art model is tested. The person re-identification techniques will be applied to Padel games due to the collaboration with PadelPlay AB. The purpose of the thesis is to keep track of players during full matches of Padel with correct identities. To this, a current state-of-the-art model is applied to an existing tracking algorithm to enhance its capabilities.

Furthermore, the purpose is broken down into two research questions. Firstly, how well does an existing person re-id model perform on Padel matches when it comes to keeping a consistent and accurate id on all players. Secondly, how can this model be improved upon to perform better in the new domain, being the sport of Padel?

To be able to answer the research questions, a Padel dataset is created for benchmarking purposes. The state-of-the-art model is tested on the new dataset to see how it handles a new domain. Additionally, the same state-of-the-art model is retrained on the Padel dataset to answer the second research question.

The results show that the state-of-the-art model that is previously trained on the Market-1501 dataset is highly generalizable on the Padel dataset and performs closely to the new model that is purely trained on the Padel dataset. Although they perform alike, the new model trained on the Padel dataset is slightly better as seen through both the quantitative and qualitative evaluations. Furthermore, the application of re-identification technology to keep track of players yielded significantly higher results than conventional solutions such as YOLOv5 with Deepsort.

Keywords

Person Re-identification, Re-id, Transformer, Padel, Computer Vision, Transfer-learning, Deep Learning.

Acknowledgments

The authors would like to thank everyone involved in this thesis for their help. First and foremost, Patrick Gabrielsson the supervisor from Jönköping University has given invaluable help in making this thesis possible. Secondly, Fredric Lundberg and Jimmy Lindström have been there throughout the journey giving both technical and non-technical aid respectively. Additionally, Maria Hedblom and Filip Andersson have given good insights on many concepts regarding the master thesis.

Contents

1 Introduction

The sport of Padel, popular primarily in Latin America and Spain, has recently experienced a boom of interest and active players. The sport is starting to gain traction both in Europe and the Middle East with a new world tour cup called "Premier Padel" starting in Qatar. With the emerging professional scene, the need for recording games and capturing data from the players for further analytical use is increasing. This is done via computer vision to track the players during play. However, algorithms for keeping track of which identities are who, are not readily available in the area of sports analytics.

The World Padel Tour (WPT) utilizes a standard camera setup for Competitive Padel, in which the entire court is visible. The camera angle, in the standard camera setup, enables the players to be visible if they move outside the court, which regularly happens during professional play. However, as Padel has expanded, most games are played by amateurs for recreational purposes. Recreational courts are usually packed together in tight configurations to optimize space in sports complexes, preventing a standard camera setup as used in professional courts. One of the most common solutions for cameras, in this case, is placing them close to the back wall. The back wall which is made from tempered glass gives a good overview of the court. However, it comes with some downsides. Due to some inherent gameplay techniques within Padel, the camera sometimes loses track of the player from either walking into a camera dead spot, such as the corner, or leaving the court to catch a play from their opponents. When the players reemerge from the dead spots, the camera tracking fails to track them as the same identity and instead assumes it's a new player entering the court. This is a problem if there is an ongoing data-gathering operation tracking statistics for each player. The technique of person re-identification (re-id), a method commonly used in camera surveillance technology with deep learning to feature extract identities, can be useful for keeping track of the players leaving and reentering the court. This helps the data gathering operation during gameplay to stay consistent and track the right data to the right player, which is crucial.

2 Background

2.1 Person re-identification

The act of identifying someone could be described as ascertaining the identity of someone or something unfamiliar or unknown. This is something humans do all the time without thinking about it. Every time we see a new person or object, we try to identify who or what it is. Re-identifying someone or something is the same process, but we are instead able to connect it to a previous instance. As a subject person re-identification comes quite intuitively to us because we, as humans, do it all the time by recognizing and remembering identities from previous encounters.

"Our eyes and brains are trained to detect, localize, identify, and later re-identify objects and people in the real world. Re-ID implies that a person that has been previously seen is identified in their next appearance using a unique descriptor of the person." (Bedagkar-Gala & Shah, 2014, pp. 270). "Humans are able to extract such a descriptor based on the person's face, height and built, clothing, hair color, hair style, walking pattern, etc. A person's face is the most unique and reliable feature that humans use to identify people." (Bedagkar-Gala & Shah, 2014, pp. 271) state that the face is the most unique feature to re-id someone. Therefore, unlocking phones with your face works so well. But this requires the camera to have a clear visual of the face in high detail.

Numerous, but similar, definitions of person re-id are found in the research literature (Gheissari et al., 2006, pp. 1528), (Bedagkar-Gala & Shah, 2014, pp. 270), (Ye et al., 2021, pp. 2872). However, one particular definition of re-id stands out and will be used as the definition in this thesis:

According to (Nambiar et al., 2020, pp. 2), one of the earliest definitions of person re-id comes from (Plantinga, 1961) stating "To re-identify a particular, then, is to identify it as (numerically) the same particular as one encountered on a previous occasion."

The definition above exemplifies the main concern with person re-identification in research. Most common scenarios utilize multiple cameras working together trying to re-id a person that has been seen in another place or time before. Therefore, the quote by (Plantinga, 1961) applies foremost to this case while also enveloping the quotes from more recent literature without concerning different cameras or camera networks.

2.2 Computer Vision

Computer Vision is a field of artificial intelligence (AI) that enables computers and systems to derive meaningful information from images, videos, and other forms of visual input. To put it into perspective, AI constitutes the brain of the computer and makes it able to think, computer vision is the eyes that enable them to see (IBM, n.d.).

The computer vision field had its humble beginnings in the 1950s and managed to successfully distinguish between typed and handwritten text in the 1970s. However, today the applications for computer vision have grown exponentially and the hardware market for this field is expected to reach $48,6 billion by 2022 (Marr, 2019).

Computer vision along with deep learning requires a lot of data. To train a model to recognize mugs, vast amounts of images containing mugs, or objects resembling mugs, would have to be fed to the model during training, for the model to learn how to discriminate between mug- and non-mug objects (IBM, n.d.).

Models are algorithms trained on-or introduced to, data. However, artificial neural networks do not only consist of a single algorithm but multiple. Artificial neural networks, for example, uses several algorithms throughout the process of training. While conducting supervised learning for a neural network backpropagation is used. Backpropagation, or reverse mode automatic differentiation in computational graphs as its also known as, calculates the gradient descent of the error function for each respective weight in the network.

Often deep learning and convolutional neural networks (CNN) are used. The "deep" in deep learning means that there is more than one hidden layer in the network if the network only has one hidden layer it is typically called a shallow network. Deep learning is a subset of machine learning networks. Much like the mammalian brain, neural nets consist of vast networks of interconnected neurons, organized in hierarchical layers communicating with each other from where they learn from large amounts of data. The networks learn from repeatedly performing tasks and gradually learn to improve the outcome with the help of deep layers who are progressively learning (IBM, n.d.).

CNN's name comes from the mathematical linear operation between matrixes called convolution (Albawi et al., 2018). However, CNN's do not perform convolutions according to the mathematical definition. The original definition originates from signal theory where one signal first must be mirrored. The mathematical operation cross-correlation is done if the signal is not mirrored. The reason convolutions are not implemented in the convolutional layer is that it is not needed. The CNN learns to use a true convolutional mathematical operation if it is needed by altering the weights, which constitute the filter in a convolutional layer. CNN's can avoid problems with spatial information. For example, in face detection software CNNs do not need to pay attention to where the face is located in the image, this method is called translation invariance (Albawi et al., 2018). The only concern of the model is to detect them regardless of position in the given image (Albawi et al., 2018). With this spatial obstacle out of the way, scientists have been able to create larger models that were not possible before (Albawi et al., 2018). For computer vision, CNN blocks are mostly used as feature extractors. Feature extraction is the process of creating derived values from measured data (features). These derived values are intended to be non-redundant and informative to further help the algorithm later when training and help it generalize on the data.

2.3 Sport Analytics

Sports analytics is concerned with collecting relevant and historical statistics of an individual or a team to gain a competitive advantage. By analyzing this data, the team, coaches, and supporting staff can make educated decisions both during and after games. Teams have utilized this data to employ strategies such as minimax and risk-taking and also used it in marketing and management of teams in terms of player trading and advertisement (Stekler et al., 2010). With the rise of sports betting, the field of sports analytics has become interesting to fans around the globe (Stekler et al., 2010). Sports betting markets are constituted differently from sport to sport (Stekler et al., 2010). In baseball and football, the bet involves picking the winner, and the market quotes odds that a certain team or individual will win (Stekler et al., 2010).

Current solutions such as YOLOv5 with Deepsort (explained later in chapter Deepsort and Kalman filter) struggle with keeping a consistent id of the player throughout an entire game of Padel. This is cumbersome when collecting data on each player according to the id they have. A simple mistake where an id is lost or swapped causes all the previous data to no longer be of use since it cannot be connected to a player. This is what a person re-id intends to fix, to re-id a person so that the old id is given back to the same player.

2.4 Padel

Padel is a racquet sport. Padel is played in doubles on a court that is roughly 75% the size of a tennis court. The court is enclosed with caged sidewalls and tempered glass walls on the player side. The main difference between Padel and tennis is that the balls can be played off the walls similarly to squash. The racquets are in the shape of an enlarged table tennis racquet and are solid instead of strung.

Over the recent years the sport of Padel, which has been popular primarily in Latin America and Spain, has gained traction in many parts of the world and especially in Europe. Sweden has seen near exponential growth since 2014. Where the amount of Padel courts have seen an average of +99% yearly average growth since 2014 as seen in Figure 1.

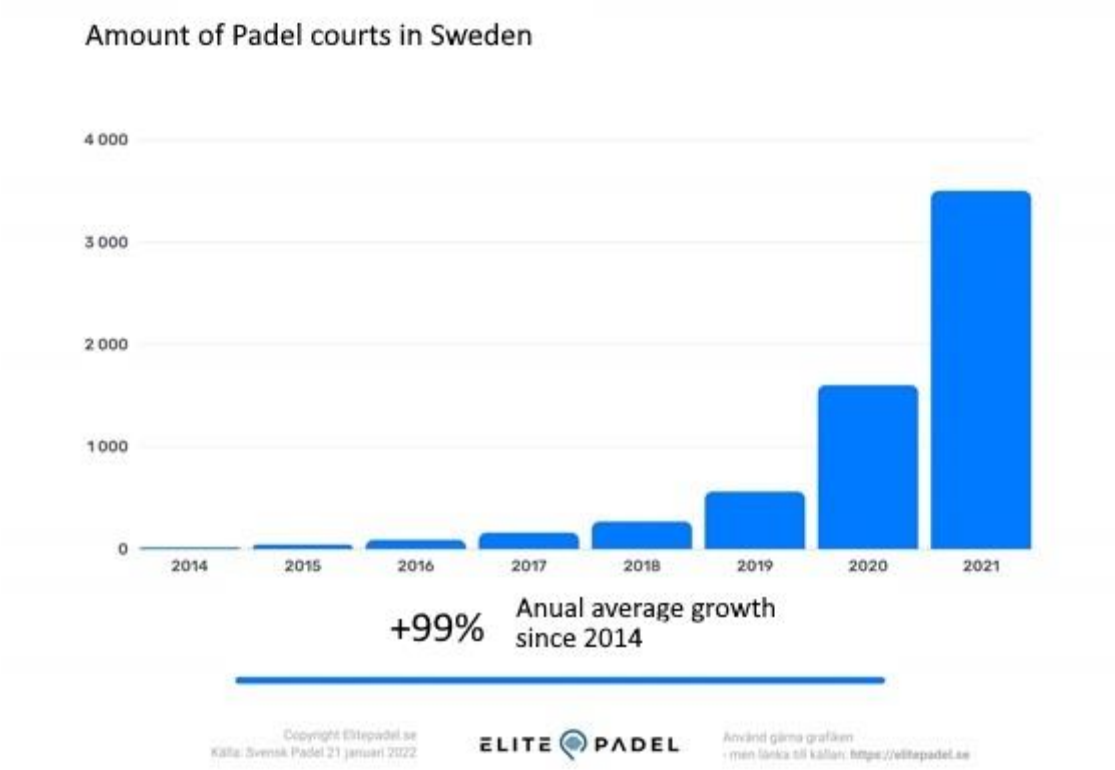

Figure 1 Yearly average growth of the number of courts in Sweden since 2014. (ElitePadel, 2022)

2.5 Purpose and research questions

From the exposition in the previous chapter, it is obvious that person re-identification is a highly relevant and important problem to solve is tracking the correct identities of people in videos or a live camera feed. This problem is especially relevant in sports, and particularly in the game of Padel where person re-id is lacking.

To address the shortcomings described above, the purpose of this thesis is therefore to apply person re-id technology to an already working object detection model which in this case is YOLOv5. YOLOv5 is chosen because previous internships at PadelPlay have resulted in this up-and-running object detection model. This will then be applied to Padel videos to attempt to keep track of each player during a match despite them leaving the frame and coming back multiple times throughout the game. This technology could also be generally useful within other sports and player tracking where the players are not always in the frame of recordings. This technology could also prove very beneficial to the area of sports analytics to record accurate data from players.

Thus, the purpose is:

To investigate the applicability of existing state-of-the-art person re-id methods in the game of Padel, and how the methods can be improved to yield highly consistent and accurate performance in tracking all players through complete games of Padel.

The goal is that each player should have a correct id throughout the entire match even if they make sudden movements or reappear into the frame after being gone for a while. Sometimes, the players might even change appearance features after disappearing from the frame. Therefore, to thoroughly evaluate the purpose it will be broken down into smaller research questions that will be investigated. Thus, the study's first research question is:

1. How well does an existing person re-id model perform on Padel matches when it comes to keeping a consistent and accurate id on all players?

In a paper by (Zou et al., 2020) they highlight the problem of re-id models having difficulties performing in new domains. This can be due to the fact that there exist large domain gaps between different datasets (Ye et al., 2022). Therefore, the second research question is:

2. How can this model be improved upon to perform better in the new domain; Padel courts?

This project is in collaboration with PadelPlay AB. The objective of this thesis is to investigate the ability to track players and identify them on footage of Padel games. This will help PadelPlay AB to further develop the company and its services and increase the company's value to its customers. By investigating this topic, the company can, consequently, track players and their whereabouts on the court and track statistics of the players. These statistics can then, later, be presented to customers.

2.6 Delimitations

This work will not cover the process of capturing nor evaluating player statistics or conducting any form of sports analytics. This study is concerned precisely with person re-id.

End-to-end learning refers to training a specific complex learning system that is represented by a single model. Some non-end-to-end re-id models want the players to be located beforehand, either manually or by another object detection model. The nature of the system set up in this thesis utilizing YOLOv5 means that the results of this thesis will not consist of an end-to-end system. This thesis has a working object detection model that has been trained on locating objects/players. Therefore, the technicalities of this model or any prerequisites will not be covered in detail and are assumed to be working. Furthermore, no end-to-end person re-id models will be tested.

2.7 Outline

The structure of the report is presented in this chapter along with the project plan that the authors adhered to during the process of creating this thesis. In the following chapter 3, Theoretical background, the theoretical background of the field and subject is introduced and used as a basis for the rest of the thesis. In chapter 4, Method and implementation, the methodology and technical setup used within this thesis will be presented along with the basis from the theoretical background chapter. After the methodology chapter, the findings and analysis from the thesis are presented in chapter 5, Findings and analysis. In this chapter data collected is presented and analysis is conducted. In the final chapter 6, Discussion and conclusions, the whole thesis is tied up. The method is discussed along with the findings. Conclusions are drawn from the knowledge gained in the previous chapter. Finally, in the subchapter Future work, additional points are proposed on what work can be done to further explore and improve this field of study.

The project plan used in this study is depicted in Figure 2.

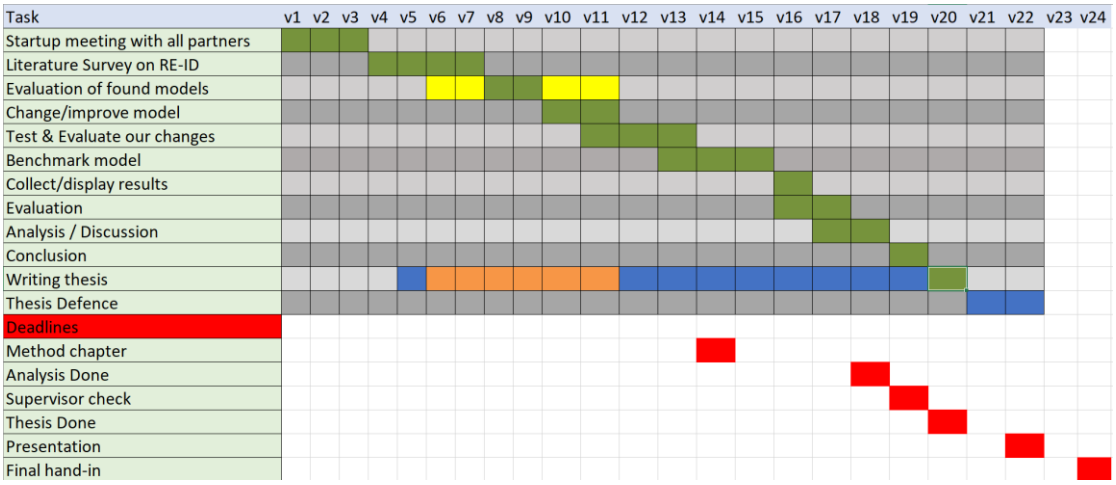

Figure 2 The time plan followed during the duration of this project

3 Theoretical background

3.1 Process

Large amounts of research have been done within the field of person re-identification. To begin the search for relevant literature, the first goal was to see whether there had been any surveys or reviews done in the specific field. This was done by using the literature database Primo. Through Primo, the first chosen keywords were: any field contains "person re-identification review", followed by the setting: peer review and date span: 2000 to 2022 and English.

The second task was to search top-tier journals and conferences to get a view of what the newest and best research was at the moment. There exist some web pages that rank the best journals and conferences. CORE Rankings Portal (Padgham et al., n.d.) is a portal called Core. It has two different ranking databases, one for journals and one for conferences. These databases include a large amount of literature that has been ranked from A*, A, B, or C where A* is the highest grade. In the journal database, 6% of the literature is in A*, A has 10% of the literature, B has 30%, C has 43%, and the remaining other literature is 8%. As for the conference database, it has a 2021 summary that includes 805 ranked venues. The A* includes 7,2% of these 805 venues, A has 16,02%, B has 37,27%, Australasian B has 1,61%, C has 36,15%, Australasian C has 1,74% and the remaining other consist of 165.

Through this Core database, relevant keywords were searched such as "computer vision" and "artificial intelligence". The result would then be a list of either journals or conferences displayed where the title and grade are shown along with some other information. From this, the most relevant sources would be chosen depending on the titles and the rank.

Some other used ranking databases include Google Scholar and Scimago Journal & Country Rank. Google Scholar (Google Scholar, n.d.) has a quite simple page that ranks the top 20 papers in each field, which in this study's case is Engineering & Computer Science -> Computer Vision & Pattern Recognition. The rank for each publication depends on its h5-index and h5-median. The h5-index is for articles published in the last 5 years and h5-median is for the median number of citations for the articles that make up the h5-index.

Scimago Journal & Country Rank is a publicly available portal that includes the journals and country scientific indicators. For the search in this study, the choice was to search within the database of Computer Vision and Pattern Recognition. They have a large amount of information on each journal and conference, with a special SJR ranking. The SJR (SCImago Journal Rank indicator) "is a measure of a journal's impact, influence, or prestige. It expresses the average number of weighted citations received in the selected year by the documents published in the journal in the thee previous years" ("Journal Rankings," n.d.).

Through all these ranking platforms, the found literature could be checked. Depending on how well each piece of literature was ranked, the trust in what they wrote increased. If some literature was not on any database, it could not be as trusted, and on the contrary, if the literature was found to be included in a journal of A* quality, it could be well trusted.

When the search began for relevant literature began, an excel document was set up where all the found literature would be logged in. The title of each document would be marked along with if it is peer-reviewed or not, the publication date, a summary of the paper, a link to the publication site, and a database ranking of the journal/conference is included if there is one. Each search would get its place in the excel sheet to make it clear where it came from. E.g., the "International Journal of Computer Vision" has its row and underneath it comes all the relevant issues published in the last years.

3.2 Literature findings

There exist extensive amounts of previous literature. As previously mentioned, the goal was to first find literature reviews within the field. (Bedagkar-Gala & Shah, 2014) released a detailed study where they give a good introduction to the field of person re-id and then continue to talk about the challenges and tasks of re-id. The focus of the paper is on large multi-camera networks such as those found in public areas e.g., shopping malls. They categorize the person re-id into 3 main problems:

- System-level
- Descriptor
- Correspondence

The first one is System-level. The difficulty with person re-id, in this case, has to do with the change in appearance that people might have. This can be during the same day when observing the same person from two different angles at two different points in time. To make person re-id work correctly, the persons first need to be detected and localized accurately. The second main problem, descriptor issues, comes after the fact that persons have been correctly localized and detected. After that, the person's visual descriptors need to be learned. The simplest and easiest descriptor is appearance which is characterized by features like color and texture. "However, these descriptors are hardly unique and prone to variations" (Bedagkar-Gala & Shah, 2014). Illumination, pose, view angle, and scale change are all prone to drastic changes in multi-camera settings. There exist two different types of re-ids. Either short or long-period re-id. In short-period re-id, the time taken between photos is only a few minutes or hours apart. In this case, appearance features work because they are more likely to be similar. However, in long-period re-id, the time taken between photos can be from days to months. In this case, the appearance features are less likely to succeed. The third main problem this paper discusses is the correspondence issues. Here they discuss the "uncertainty attributed to the possible lack of prior known spatio-temporal relationships between cameras" (Bedagkar-Gala & Shah, 2014). This is when the appearance of the person changes

dramatically due to objects like bags, unzipped jackets, etc. They also discuss that there are two different scenarios for person re-id: open set re-id and closed set re-id. The open set is when the tracking is done over several cameras where the gallery of photos evolves. Closed-set re-id is when a single camera is present, and the gallery size is fixed. A gallery is a collection of identities either in the form of feature vectors that has been extracted from images from known identities or images of known identities. In this scenario, the person that is to be detected is assumed to be in the gallery. The paper then continues to discuss the current works in the field of person re-id. They finish off the paper by discussing that most of the work on person re-id leverages clothing appearance-based features designed for short-period re-id and is evaluated in closed set re-id scenarios. The issue of long-period re-id is entirely unexplored and open set re-id is not completely tackled.

A very recent article by (Ye et al., 2022) made a very thorough survey of different deep learning approaches to person re-identification. It makes the difference from the previously summarized paper by talking more about the topic person re-id itself. (Ye et al., 2022) gives a good introduction to the topic of person re-id. They explain that, given a person-of-interest query, the goal of person re-id is to determine if this person has appeared in another place at a distinct time captured by a different camera, or even the same camera but at a different time. This query can be represented by an image (Almazan et al., 2018), (L. Zheng et al., 2015), (Martinel et al., 2019), a video sequence (T. Wang et al., 2014), (L. Zheng et al., 2016), or even a text-description (Ye et al., 2015), (S. Li et al., 2017). They use the rationale of "demand of public safety" and an increasing amount of surveillance cameras to state that person re-id is crucial in intelligent surveillance systems with high research impact and practical importance, which will be further elaborated on in chapter 3.5 Ethics. (Ye et al., 2022) move on by stating some challenges that come with person re-id such as different viewpoints (Karanam et al., 2015), (Bak et al., 2014), varying low-image resolutions (X. Li et al., 2015), (Y. Wang et al., 2018), illumination changes (Y. Huang et al., 2019), unconstrained poses (Cho & Yoon, 2016), (Zhao et al., 2017), (Sarfraz et al., 2018), occlusions (H. Huang et al., 2018), (Hou et al., 2019), heterogeneous modalities (S. Li et al., 2017), (A. Wu et al., 2017), complex camera environments, background clutter (Song et al., 2021), unreliable bounding box generations. With these challenges, along with several more, person re-id remains an unresolved problem. (Ye et al., 2022) continue by stating that some earlier research has focused on handcrafted feature construction with body structures (Gray & Tao, 2008), (Farenzena et al., 2010), (Yang et al., 2014), (Liao et al., 2015), (Matsukawa et al., 2016) or distance metric learning (Kostinger et al., 2012), (W.-S. Zheng et al., 2011), (Xiong et al., 2014), (Hirzer et al., 2012), (Liao & Li, 2015), (Yu et al., 2020). There exist previous surveys in this area that have summarized deep learning techniques. The paper by (Ye et al., 2022) does three things differently. First, this paper provides existing deep learning methods by showing the pros and cons. Secondly, the paper presents a design with a new powerful baseline: "AGW: Attention Generalized mean pooling with Weighted triplet loss". As well as a new evaluation metric "mINP: mean Inverse Negative Penalty". Their AGW model performs better than many state-of-the-art models. Finally, they discuss several important research directions as well as under-

investigated open issues to narrow the gap between the closed-world and open-world applications. When building a person re-id system, (Ye et al., 2022) writes that it generally requires 5 main steps:

1) Step 1: *Raw Data Collection*
2) Step 2: *Bounding Box Generation*
3) Step 3: *Training Data Annotation*
4) Step 4: *Model Training*
5) Step 5: *Pedestrian Retrieval*

Along with all these steps, the paper makes a comparison between the closed-world and open-world settings (Ye et al., 2022):

1) In step 1: in a closed-world setting, the persons are represented by images/videos. In the open-world setting, data might not be the same, where the data can be infrared images, sketches, depth images, or text descriptions.
2) In step 2: when generating bounding boxes, in a closed-world setting the assumption can be made that these have been pre-generated, whereas, in an open-world setting, they need to be handled.
3) In step 3: in a closed-world setting, the assumption can be made that there is enough labeled data for supervised training. In an open-world setting, the same assumption cannot be made, which necessitates the use of semi-/unsupervised models.
4) In step 4: closed-world person re-id models usually assume that all the annotations are correct with clean labels. However, "annotation noise is usually unavoidable due to annotation error (i.e., label noise) or imperfect detection/tracking results".
5) In step 5: In the closed world setting and the person retrieval stage, it is assumed that the query identity must exist in the gallery set. However, if the query does not exist in the gallery it needs to be added to the gallery. Therefore, the open set exists.

In (Z. Li et al., 2021), the authors highlight that the key to success within person re-id is to extract discriminative personal features. Therefore, they propose various part-level feature extracting methods.

Personal feature extraction can be divided into two aspects: global-based feature extraction and locally-based feature extraction. Globally based features focus more on the overall information about the person; however, it ignores the spatial structure of a person. Therefore, in recent years, the primary focus of much research has been on extracting locally based features. (Z. Li et al., 2021)

Wei et al. also mention that the key to extracting locally based features is that every part should be located accurately (Wei et al., 2017). To utilize locally based features some techniques are needed, such as pedestrian external information, a pose estimator, and datasets for the pose estimator. The pose estimator and its external datasets outside of the re-id model without a doubt contribute to increased training complexity. (Z. Li et al., 2021)

Due to the increased training complexity, the community has moved away from pose estimation and moved toward partition strategies. In the re-id community, person images are usually divided into six parts: head, upper body, lower body, upper legs, lower legs, and feet. Unfortunately, problems arise during the training of the models when parts of the bodies are different, this can be caused by misalignment of the person images. Therefore, (Z. Li et al., 2021) propose a network with Part Prediction Alignment (PPA) to extract part-level features for re-id. While global features contain the spatial information of an image and not a person's spatial structure cues, local features are concerned with a person's spatial structure. (Z. Li et al., 2021) therefore, believe that they can combine the best of both approaches where the features will be discriminative and therefore achieve higher overall accuracy. To tackle this, it is proposed to use a teacher-student network to extract global-local features for re-id. A teacher-student network works by using an already-trained network (teacher) as a backbone to train a new network (student). The advantage of this is that it is possible to switch out the teacher network and train a new student network with many different already trained teacher networks. (Z. Li et al., 2021)

The intuitive choice is made to not delve deeper into gain analysis since it usually assumes some natural way of walking of the persons as you might to casually on the street. This rarely happens in Padel since the players move around in many different ways when playing the game.

Deep networks require more data to avoid overfitting. Training deep networks on large datasets takes a long time, especially on traditional CPU cores. With the rise of the availability of general-purpose computing on GPUs deeper networks and more data could therefore be trained. (Raina et al., 2009) is one of the earlier explorers of using GPUs to train models and some more commonly heard ones are AlexNet (Krizhevsky et al., 2012) and (Cireşan et al., 2013).

One-shot learning

The problem with supervised learning regarding features, in this case concerning the tracking of Padel players, is that it is not feasible to attain training data. Training a conventional neural network would require multiple images per class. There is no way of knowing what player will be in the video. Imagine a similar case with automatic passport control at airports. This would require a dataset with everyone in the world with several images of them to train. To avoid this problem, the technology of one-shot learning can be used. With this, you train a deep learning algorithm to extract feature vectors. After training, you can send in new classes, in this case, player identities that haven't been seen before, through the network and compare the existing feature vectors in the database(gallery) to dynamically add new classes (feature vector types) to the database without needing to retrain the network. This makes it possible to classify new players entering the court without having prior training data on these individuals.

3.3 Deepsort and Kalman filter

Kalman filtering

Kalman filtering, commonly used in statistics and control theory, is an algorithm that utilizes a series of measurements observed under a certain timespan. Kalman filtering can produce unknown variables that usually is more accurate than variables based on a single measurement alone. It achieves this by estimating a joint probability distribution over the variables for each timeframe in the set timespan. The Kalman filter assumes that each variable in the time frame is both random and Gaussian distributed. Therefore, the variables have a mean value μ, which equates to the center of the gaussian distribution. The variables also have an uncertainty value: σ^2. Together these two values can be mapped between two variables, in this case, position and velocity, to form a Gaussian blob. Intuitively these two variables do correlate. These types of variable correlations are important because they can act as another source of information. As they are correlated one measurement can give information on what the other one could be. Therefore, the variables can be put into a covariance matrix $\sum i, j$.

$$\mu = \begin{bmatrix} position \\ velocity \end{bmatrix}$$

$$P_k = \begin{bmatrix} \Sigma_{pp} & \Sigma_{pv} \\ \Sigma_{vp} & \Sigma_{vv} \end{bmatrix}$$

Where μ is the best estimate and P_k is its covariance matrix. To further predict the state of these two variables, in this case, position and velocity, a kinematic formula is applied. A kinematic formula contains displacement, time interval, initial velocity, final velocity, and constant acceleration. With these values, it is then possible to create a prediction matrix for the next timestep.

There may be external forces affecting the values as well. If these external forces are known, they can be added to the equation as a vector to correct future predictions. With this, it is then possible to create a best new prediction which is made from the previous best prediction from an earlier timestep. Then the prediction is corrected from the known external force vector. Then a new uncertainty is predicted the same way by looking at the previous best uncertainty along with environmental uncertainty.

Deepsort

Deepsort is an extension and improvement of the original Simple Online and Realtime Tracking (SORT) algorithm (Bewley et al., 2016). Deepsort uses a conventional single hypothesis tracking methodology with recursive Kalman filtering along with frame-by-frame data association.

Furthermore, the track handling and Kalman filtering framework for Deepsort is for the most part identical to the original SORT algorithm (Wojke et al., 2018).

Deepsort considers 8-dimensional space features $(u, v, \gamma, h, \dot{x}, \dot{y}, \dot{\gamma}, \dot{h})$, Where (u, v) are the 2D center coordinates of a bounding box and γ is the aspect ratio. Aspect ratio is the ratio between the width and height of an image. Finally, h which is the height of the bounding box. These features along with their respective velocities in image coordinates (Wojke et al., 2018). To track a player Deepsort then utilizes the standard Kalman filter as mentioned previously with a constant velocity motion, and a linear observation model where it takes (u, v, γ, h) as observations of the object state (Wojke et al., 2018).

Deepsort can count the number of frames since the last successful track. Older tracks (meaning having a high number of frames since the last successful track) than a certain pre-set value are considered to have left the scene. These old tracks are then removed from the set of objects to be tracked. Similarly, in many re-id models when a new track gets detected, it checks if the new track can be associated with an existing track. If the new track is not able to associate with one of the currently tracked objects a new track will be initiated. However, these new tracks have a tentative track during their first three frames (Wojke et al., 2018). If a track cannot be successfully associated within these three tentative frames it will be deleted.

3.4 Transformers

Transformers have been gaining popularity ever since they were first introduced back in (Vaswani et al., 2017) utilizing only the attention mechanisms. The transformers showed that they were able to yield state-of-the-art performance with a fraction of the training time used. This was in the field of NLP (Natural Language Processing).

First Transformer

When (Vaswani et al., 2017) created the first transformer, it was decided to have an encoder-decoder structure since this was what the most competitive similar models used. The encoder maps an input sequence of symbol representations $(x_1, \dots, x_n)$ to a sequence of continuous representations $z = (z_1, \dots, z_n)$ which the decoder then uses to generate an output sequence $(y_1, \dots, y_n)$ of symbols one at a time. The model is at each step auto regressive so that it consumes the previously generated symbols as an additional input when generating the next one.

The model utilizes a structure using stacked self-attention and pointwise, fully connected layers for both the encoder and the decoder, shown in two separate halves in Figure 3.

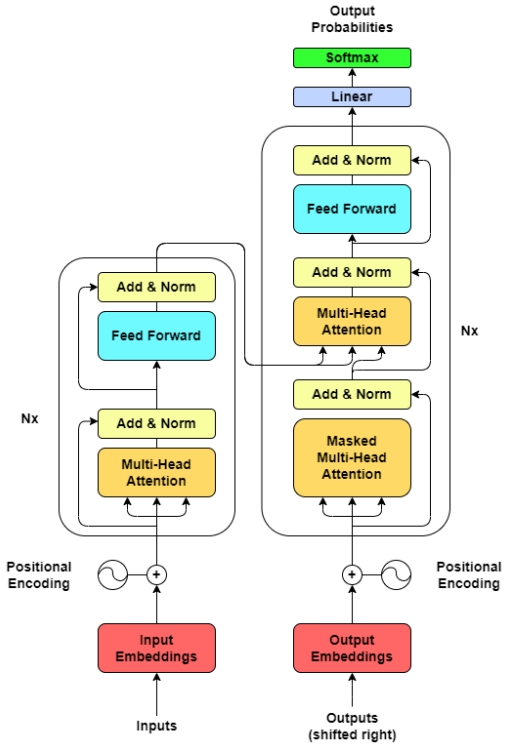

Figure 3 Original Transformer architecture illustrated from (Vaswani et al., 2017)

The encoder is composed of N = 6 identical blocks including 2 sub-layers each. The layers consist of a multi-head (self) attention (MSA) and a Feed-Forward layer, which are explained later. Before entering the encoder, the input word input goes through the input embedding and the positional encoder. The input embedding maps the input word to a place within the embedding space. This is where similar words are close to each other e.g., guitar and violin could be considered close. The output from the input embedding is a vector that then goes to the positional encoder. Since words have different meanings depending on where they are located within a sentence, the positional encoder will generate a vector that has information on the distances between the words within the sentence. (Ramachandran et al., 2019) use the following functions to calculate this:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right)$$

Equation 1, Sin positional encoder

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right)$$

Equation 2, Cos positional encoder

where the d_{model} is the output dimension of sub-layers within the model, they use 512. The pos marks the position of the word within the sentence and i is the dimension (or length) of the sentence. At this stage, the word has been translated to a word vector containing positional information which gives context to a sentence.

The multi-head attention layer is concerned with calculating the importance of words within a sentence. For each word, an attention vector is outputted capturing the contextual relationships between the words within a sentence. It outputs i amount of attention vectors. After that, the feed-forward layer which consists of a feed-forward network is applied to each attention vector so that it is more readable by other modules.

The add & normalization layer step consists of using layer normalization (Ba et al., 2016). The following function:

$$LayerNorm(x + Sublayer(x))$$

Equation 3 Add & normalization using layer normalization

is used to add the x (word vector) to the $Sublayer(x)$ which is the previous layer's output.

The decoder is also composed of N = 6 identical blocks but includes 3 sub-layers instead. It has two similar sub-layers to the decoder, but an additional masked MSA layer. The role of the masked MSA layer is similar to the MSA block. The difference is that while training the network to translate from one language to another, some of the target language labels are masked (hidden) from the model. It is so that the network cannot use the entire answer for the translation immediately. As the words within the sentence are iterated through, more of the sentence is visible for use.

After the masked MSA layer, the process is similar to the encoder, but with both the original sentence from the encoder and the new target sentence. After the feed-forward block, the data is input into the linear layer which is another feed-forward layer to expand the dimensions into the number of words in the target language. Finally, the softmax layer transforms the data into a probability distribution which then is readable by humans. The output word is the word that has the highest probability.

Visual Transformer

(Dosovitskiy et al., 2020) introduced the transformers into the field of Computer Vision and called it ViT (Vision Transformer). They took a standard transformer from the field of NLP and applied it directly to images with as few modifications as possible. They state that when trained on "mid-sized" (Dosovitskiy et al., 2020, pp. 1) datasets the transformer did not generalize well. Therefore, they pre-trained their model on other datasets and their dataset to finally obtain 88.55% accuracy on ImageNet and beat other state-of-the-art models in several image recognition benchmarks. Sometime later (Touvron et al., 2020) managed to reach 85.2% accuracy on the ImageNet while solely using the ImageNet dataset without pretraining. They built upon the ViT introduced in (Dosovitskiy et al., 2020).

When the ViT was introduced by (Dosovitskiy et al., 2020), previous computer vision convolutional architectures had remained dominant, but ever since the success of Transformers in NLP, several researchers have tried to combine them with CNN-like architectures (X. Wang et al., 2018), (Carion et al., 2020), and some replacing convolutions entirely (Ramachandran et al., 2019), (H. Wang et al., 2020). The replacement of convolutions did not scale too well on hardware because of specialized attention patterns.

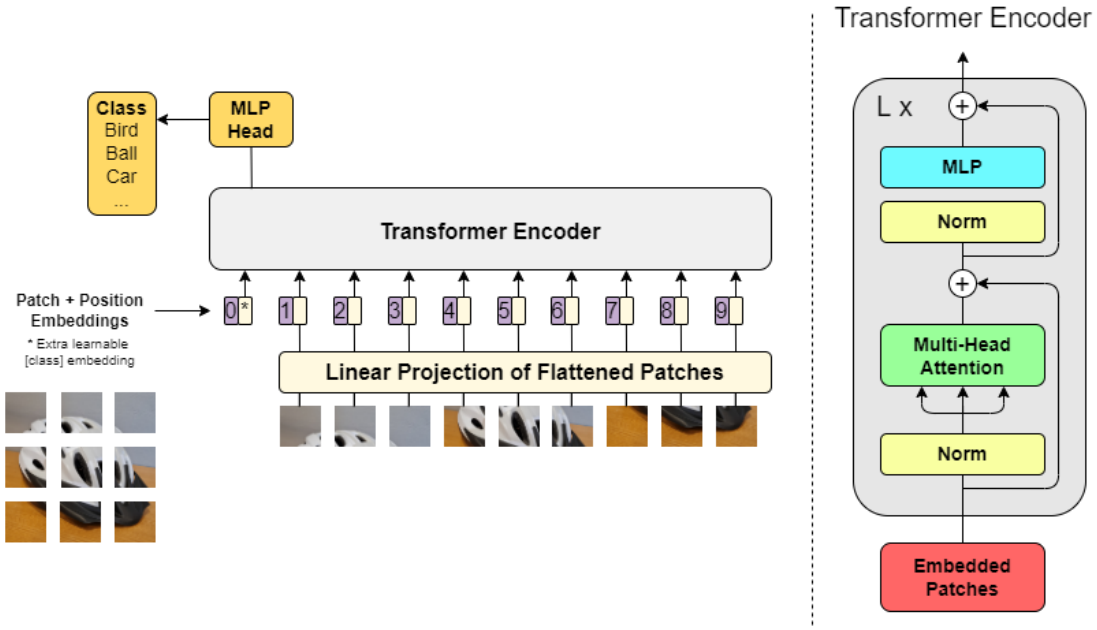

Figure 4 Visual representation of ViT (Vision Transformer) illustrated from (Dosovitskiy et al., 2020)

Figure 4 represents the ViT. Comparing it to the original Transformer in Figure 3, the difference can be seen in the steps before entering the Transformer encoder itself. The standard Transformer takes a 1D input as a sequence of tokens as seen in Figure 3. To be able to handle images, these would need to be reshaped somehow. (Dosovitskiy et al., 2020) decided to reshape an image from

$$x \in \mathbb{R}^{H \times W \times C}$$

Equation 4

into a sequence of flattened 2D patches

$$x_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$$

Equation 5

where (H, W) is the height and width of the original image, C is the number of color channels. (P, P) are the dimensions of each image patch and N is the resulting number of patches as can be seen in Equation 6

$$N = HW/P^2$$

Equation 6

The order of layers within the Transformer encoder block is slightly modified so that layer normalization (LN) is applied before the multi-head attention and multi-layer perceptron (MLP). The abbreviations clash a bit since in Figure 3 they use "Add & Norm" as a step within the blocks. In Figure 4 the "Add" part is visualized just like a (+) sign and the "Norm" is correlated to the LN. This choice by (Dosovitskiy et al., 2020) was inspired by (Q. Wang et al., 2019) and (Baevski & Auli, 2018).

Locally Aware-Transformer

At the time of writing, the top-performing person re-id transformer on PapersWithCode is the Locally Aware Transformer (LA-Transformer) (Sharma et al., 2021). They state that currently, the Vision Transformers are starting to replace pure CNNs for a variety of object recognition tasks. The name LA-Transformer comes from the fact that transformers yield a local token apart from the main global classification token. Techniques to use these local tokens are an active research area from which the name Locally Aware Transformer has emerged. This model outperforms "all other state-of-the-art published methods" at the time of writing (Sharma et al., 2021).

(Beal et al., 2020) stated, "*The remaining tokens in the sequence are used only as features for the final class token to attend to. However, these unused outputs correspond to the input patches, and in theory, could encode local information useful for performing object detection"*. They noticed that they have a strong connection to the original input patches and could therefore consider their use to enhance feature representation of the original image to more strongly couple vision transformers to fully connected (FC) classification techniques. The use of local tokens with FC classification techniques is the primary intuition behind the LA-Transformer (Sharma et al., 2021).

(He et al., 2021) was the first to employ Vision transformers in re-id and managed to receive competitive results compared to the current CNN models. (Sharma et al., 2021) extends on those results in several ways but mostly because they also aggregate the globally enhanced local tokens using a PCB-like (Part-based Convolutional Baseline) strategy that takes advantage of the spatial locality of these tokens. But the difference is that (He et al., 2021) make use of their local tokens through a shuffling step which does not take advantage of the 2D spatial locality information inherent in the ordering of the local tokens. "*LA-Transformer overcomes this limitation by using a PCB-like strategy to combine the globally enhanced local tokens while first preserving their ordering in correspondence with the image dimension."* (Sharma et al., 2021). The authors of the LA-Transformer (Sharma et al., 2021) have taken both global and local information of the picture, something that the previously mentioned in the chapter Literature findings when talking about either extracting globally or locally personal features. This has before been a choice of using either one, but the LA-Transformer makes use of both.

The architecture of the LA-Transformer can be seen in Figure 5. The figure consists of two main parts. Firstly, the backbone and then the locally aware network. Both are interconnected and trained as a single network.

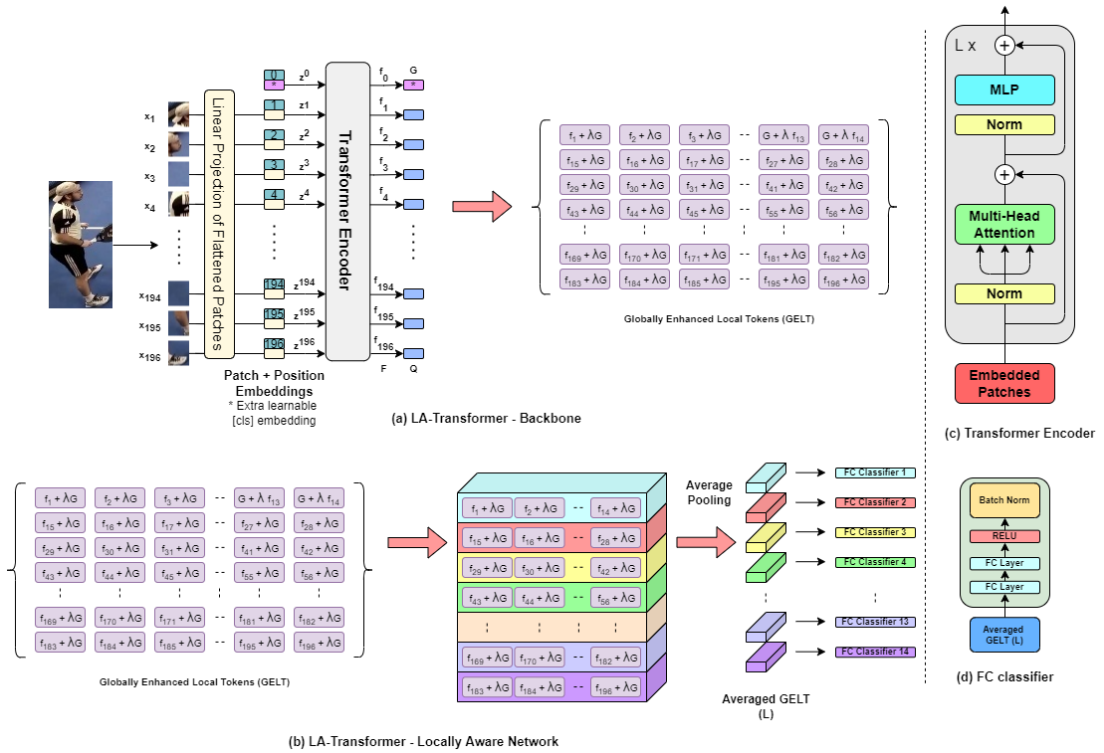

Figure 5 Architecture of LA-Transformer illustrated from (Sharma et al., 2021)

Part (a) shows the backbone of the architecture. The backbone is the ViT, the one proposed by (Dosovitskiy et al., 2020) and shown in Figure 4. The ViT generates tokens $F = f_0, f_1, \dots, f_N$. The token f_0, also known as the global token is referred to as G. The rest of the outputs $f_1 \dots f_{196}$ are referred to as local tokens or Q. Globally Enhanced Local Tokens (GELT) are obtained by the combination of G and Q using weighted averaging and are arranged into a 14 × 14 2D spatial grid. Part (b) shows the locally aware network. The GELTs go through average pooling row-wise, and the result can be seen as L. L is then fed to the locally aware classification module. In part (c) the architecture of the transformer encoder can be seen as well as in Figure 4 and in part (d) the architecture of the FC classifier is shown.

The backbone of the LA-Transformer consists of the ViT. The ViT requires a lot of training to become useful. The person re-id datasets do not have the amount of data to fully re-train such a network. The required number of images is between 14-300 million for an efficient training process. The Market-1501 dataset (L. Zheng et al., 2015) has ca 36000 images where ca 13000 are for training as seen in chapter Market-1501 dataset. Therefore, the pretrained ViT from (Dosovitskiy et al., 2020) is applied and slightly retrained in the finetuning part of chapter Training.

The ViT requires the images to be in 224x224 pixels as input, therefore the images are resized to this size as later explained in chapter Pre-processing. The authors of (Sharma et al., 2021) then continue by converting the image into N number of patches $x_p^i | i = 1, \dots, N$. The patches are then projected into D dimensions linearly using the patch embedding function: $(E(x_p^i) | i = 1, \dots, N)$ Equation 9, using a convolutional layer with a kernel size of 16x16. To not overlap the patches, a stride of 16 is used by the authors. The D dimensions number corresponds to number of channels and is set to 768 "which represents the size of the embedding" (Sharma et al., 2021, pp. 5). N is the total number of patches and can be calculated using Equation 7. If there is no padding, H, K are the height and width of the image, K_H, K_W are the height of the kernel and S is the kernel stride, which was 16.

$$N = \left(\frac{H - K_H}{S} + 1\right) \times \left(\frac{W - K_W}{S} + 1\right)$$

Equation 7 Function for number of patches

$$N = \left(\frac{224 - 16}{16} + 1\right) \times \left(\frac{224 - 16}{16} + 1\right) = 196$$

Equation 8 Equation 7 with values inserted

Thereafter, the class embedding x_{class} is attached as an additional variable to the patch embedding $(E(x_p^i)$ which then will keep the information of the entire image and serve as the global vector. (Sharma et al., 2021) The position embeddings P is added to the resulting vectors to conserve the positional information (Sharma et al., 2021). Finally before the transformer encoder, the final batch of vectors z_0 (Equation 9) is input into the transformer encoder (Figure 5) to create $N + 1$ amount feature vectors where N is the number of patches (196) plus class embedding.

$$z_0 = [x_{class}; E(x_p^1); E(x_p^2); \dots; E(x_p^N)] + P$$

Equation 9 Patch embedding function

Transformer Encoder

As for the Transformer encoders, (Sharma et al., 2021) use $B = 12$ blocks instead of the 6 in (Vaswani et al., 2017). Each one consists of an alternating MSA and MLP. The LN is applied before each MSA and MLP layer and a residual connection is applied after each respective block. The pass-through of all the B blocks can be seen in the following equations (Equation 10, Equation 11, Equation 12) where the final encoder output is F:

$$z_b' = z_{b-1} + MSA(LN(z_{b-1}))$$

Equation 10 MSA step with LN and residual connection

$$z_b = z'_b + MLP(LN(z'_b))$$

Equation 11 MLP step with LN and residual connection

$$F = LN(z_B)$$

Equation 12 Final output with LN

The difference compared to (Dosovitskiy et al., 2020) is that in the LA-Transformer all of the features z_B are used along with the class embedding.

The output from the encoder is $N + 1$ feature vectors where the global token $G = f_0$ and the local tokens $\mathrm{Q} = [f_1, f_2, f_3, \dots, f_N]$ where N is the number of patches, as seen in Figure 5. L is defined as the averaged GELT after using Equation 13.

$$L_i = \frac{1}{N_R} \sum_{j=i*N_R+1}^{(i+1)*N_R} \frac{(Q_j + \lambda G)}{(1 + \lambda)} \qquad i = 0 \dots N_C - 1$$

Equation 13 Local vector averaging

The N_R and the N_C are defined as the number of patches per row and column respectively. In the case of (Sharma et al., 2021) and using the ViT backbone which requires images in 224x224 and calculating N with Equation 8, the values of N_R and N_C equals to $\sqrt{N} = \sqrt{196} = 14$. Equation 13 takes a row of patches and creates one local vector per row.

N_C amounts to the number of FC classifiers. The FC classifier can be seen in Figure 5 and consists of two fully connected layers, Rectified Linear Unit (ReLU), and a batch normalization (Ioffe & Szegedy, 2015). The final output of the LA-Transformer is defined as y as seen in Equation 14, Equation 15, and Equation 16:

$$y_i = FC_i(L_i) \qquad i = 1 \dots N_C$$

Equation 14 Fully connected layer

$$score = \sum_{i=0}^{N_C} softmax(y_i)$$

Equation 15 Score calculation

$$prediction = argmax(score)$$

Equation 16 Final prediction

The outputs, y, passes through the softmax which are then summed together by Equation 15. The softmax with the highest score is the finals prediction and id of the found person Equation 16.

The ReLU is a widely used activation function that causes faster training and improves computational cost (Vargas et al., 2021), created by (Hinton, 2010). It is a simple max function that causes the data that comes through to be set to 0 if it previously was lower than 0 and always takes the highest value as output. This can be seen in Equation 17 where x is the input.

$$f_{ReLU}(x) = \max(0, x)$$

Equation 17 ReLU function

3.5 Ethics

Due to the nature of where this technology originates from, which is surveillance, ethics is of a big concern. With the increase in security and forensics with the help of improved access to media technology, surveillance cameras are increasing rapidly in numbers (Nambiar et al., 2020). According to the British Security Industry Association, there is one camera for every 11 people in the UK (Barrett, 2013). The average Londoner is caught on a surveillance camera on average 300 times a day (Pillai, 2012). To respect the integrity and privacy of players on Padel fields utilizing the AI-tracking and techniques described in this paper, we avoid saving images of the players' long term. It is also attempted to treat images in the algorithm as little as possible and instead only use them for feature extracting to generate feature vectors and thus going forward only using these feature vectors. This means that we attempt to minimize how long the images of players are stored on disk.

Surveillance

In many countries, video surveillance is considered either a primary tool to enforce security or simply a crime deterrent tool. If an incident occurs, law enforcement authorities can review the available video footage, and identify a set of subjects of interest, by matching the captured images/video to their associated identities. (Yaghoubi et al., 2021)

From a vision and surveillance point of view, person re-id is a highly relevant, contemporary topic with great significance in research and application development. Systems have to re-identify persons in large camera networks (Nambiar et al., 2020). Within person re-id, surveillance is the most applied area and has been thoroughly researched. The rationale many developers or researchers use for justifying the relevance and importance of work conducted in this area is increased security and catching/preventing e.g. robberies by creating "intelligent technology". (D. Wu et al., 2019) discusses a robbery and homicide case that happened in China, in 2012. The police force mobilized more than 500 officers to analyze thousands of videos in an attempt to track the robber's last 24 hours. If there instead had been intelligent technology automating the surveillance video

analysis. The analysis could have been expedited significantly or been conducted in real time.

Ethics is an important consideration in surveillance technology research and engineering. This also applies to person re-identification. In 2018, a lawsuit was filed against Walmart in the United States of America. The lawsuit was claiming that their video recording technology violated the Song-Beverly Act, which prohibits businesses from recording customers and extracting personal identification information, such as eye color, hair color, and facial features, as a condition of accepting a credit card payment (Bender & Dori, 2019). However, overseas in Europe, the General Data Protection Regulation (GDPR) protects an individual's private information, where large fines are issued when not complying with the legislation (Koptelov, 2021). In 2019, the Swedish Data Protection Authority (DPA) fined a municipality for using facial recognition technology for students to track attendance in school (European Data Protection Board, 2019). The ethics, privacy issues, and dilemmas regarding the technology of person re-id are complicated but will be heavily regarded throughout this thesis.

3.6 Evaluation metrics

Relevant evaluation metrics are important. Different fields within AI have different standard, or conventional, evaluation metrics. Conventionally within the field of person re-id two metrics are commonly used: Cumulative Matching Characteristics (CMC) and Mean Average Precision (mAP) (Sharma et al., 2021). CMC utilizes a per-step top-k function which gives a score of 1 or 0 depending on if the returned gallery samples contain the query identity. For example, if top-k is used, where $k = 5$, the model returns a set of 5 images from the gallery, and if the expected query identity is found within this set of 5 images, the evaluation metric returns a score of 1. Furthermore, the 5 images returned are also ranked from most likely to least likely in the set. This is then done n times and the average is calculated and returned as the CMC metric.

$$Acc_k = \begin{cases} 1 \ \ if\ top-k\ ranked\ gallery\ samples\ contain\ the\ query\ identity \\ 0 \ otherwise \end{cases}$$

Equation 18, CMC

The Mean Average Precision metric is, in the person re-id field, calculated from a set of queries. It's a popular metric to measure how the model is doing on the task it was trained on. To calculate the mAP two other values need to be calculated first: precision and recall. Precision measures how accurate the prediction is. This means the percentage of all correct predictions. Recall on the other hand measures how well the model can find all the positives. These two values can then be presented as a curve. The average precision is the mean of precision at a specific threshold in the curve while the mAP can be measured as the area under the curve of these two values.

The mAP over a set of queries is defined as:

$$mAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q}$$

Equation 19, mAP on a set of queries

Where Q is the total amount of queries and $AP(q)$ is the average precision of a given query, q.

The reason these two metrics were used other than being frequently used in the field is because with the help of these two metrics it is possible to evaluate the model's performance both quantitively and qualitatively. Picking up on the top-k ranking explained earlier in this chapter, imagine a CMC metric where the rank 5 accuracy is 70%. That means that a correct prediction of the correct identity will occur in the top 5, 70% of the time. This means that the better the algorithm, the higher the top-k CMC percentage. However, CMC only summarizes the accuracy of these top-k ranked sets, it ignores other factors such as similarity scores and strong versus weak scoring hits (Grother et al., 2019). Therefore, mAP is utilized to also get an understanding of the precision of the model as it cannot be gathered from the CMC metric.

3.7 Transfer learning

Ideally, there would exist enough labelled data so that a new domain could be fully learned just based on that data. Unfortunately, this is not always the case and because of this transfer learning exists. "Transfer learning aims at improving the performance of target learners on target domains by transferring the knowledge contained in different but related source domains." (Zhuang et al., 2021, pp. 43).

In Figure 6 some intuitive examples are presented about knowledge transfer. The first example could be seen as teaching a person to play Western chess after already having learned to play Chinese chess. The second example could be about learning to ride a motorbike after already having experience riding a regular bike.

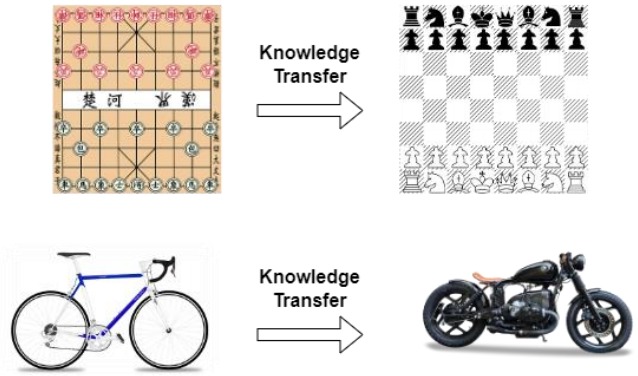

Figure 6 Intuitive examples of transfer learning. Illustrated from (Zhuang et al., 2021). The utmost example is a knowledge transfer from Chinese chess to Western chess. The bottom example is a knowledge transfer from bicycles to motorbikes.

As humans learn new things through years of growing up and aging, we tend to take with our previous experiences. For example, Magnus Carlsen the current world champion in chess (Western chess), would have an easier time trying a new board game such as Chinese chess for the first time compared to someone who never has played any other chess variant. The same thing goes for the second example in Figure 6. Someone who has never sat on a bicycle before will have a harder time learning to balance on a motorbike compared to someone who has.

This does not only apply to humans. If a program has been taught to recognize bicycles, it will require fewer new images to learn a new similar subject as motorbikes as compared to training to recognize airplanes.

Additionally, if a certain program knows how to recognize motorbikes, transfer learning on the same program but with new motorbike images, can be called domain adaptation. This is usually done when wanting to increase the performance of the original program in a slightly new domain. E.g., the first model was trained to only recognize motorbikes from the side, this new adaptation could add new data so that the new program learns to recognize motorbikes from a different viewing angle as well.

Transfer learning into a new domain and only knowing the labels of the target domain is known as inductive transfer learning (Zhuang et al., 2021). (Sharma et al., 2021) use this method when they are transfer learning their new model on top of the old model which has been trained on ImageNet (Fei-Fei., 2009). ImageNet dataset contains 3.2 million images over 5247 categories. This can give a model high generalizability since it is trained on many different subjects. (Sharma et al., 2021) use something called fine-tuning to train their model very carefully. This usually includes tweaking many parameters to perfectly train the desired model. Learning rate, learning decay, and layers are some of the important parameters to have in mind.

Transfer learning involves freezing the layers of the model. It disables their ability to be changed and therefore not be overwritten. This is very important if you want to keep the features of the original network and not overwrite them. Learning rate is a variable that controls how much the network is adjusted after each learning epoch. This value spans from 1 to 0. A higher number will affect the weights more. Adjusting the learning rate to the correct level is hard. Too high of a value will cause the model to miss its optimal accuracy, too low of a value will make the training take a long time and potentially miss the optimal accuracy. Learning decay is a variable that is multiplied by the learning rate. The value is between 1 and 0. The purpose of the decay is to gradually decrease the learning rate. As mentioned, a learning rate that is too high can cause sub-optimal training, a learning decay can help training in the later stages. The authors of (Sharma et al., 2021) unfreeze the top two frozen layers of the model every two epochs. Additionally, every 2 epochs they apply the learning decay. With this technique, the layers within the model are gradually unfrozen as training goes on and the learning rate is decreased to not unlearn any important features of the previous model.

3.8 Data augmentation

Data augmentation is concerned with augmenting data to create more similar data. This is to increase the amount of data available. Models such as Transformers require large amounts of data to fully train and the amount of data is not always available. Transfer learning has been created to alleviate this, but it is not always enough if a domain change is of need.

When it comes to image data, the images can be manipulated so that they look alike but slightly different. Images can be rotated, flipped, resized, the color changed, etc. The purpose of increasing dataset size is to create less "perfect" data. E.g., the object of interest might not be fully in frame, or the color is slightly darker. Training on this data will prepare it for situations where the model might stumble on this kind of data. This will make it more generalizable. Additionally, data can be normalized. This is useful in cases when the data has high sporadic behavior. If not normalized, models can start to assume that certain extreme values have a higher impact on the output, which is not always true.

4 Method and implementation

In this chapter, the work process is presented with the theoretical backing of the information presented in the Theoretical background chapter. The process of data gathering, usage of model, development of a new method, and dataset construction is explained along with some newly discovered techniques. The developed person re-id system is inspired by the workflow structure presented in the chapter Literature findings recommended by (Ye et al., 2022).

A mixture of both open and closed-world settings is used for multiple reasons. The images later described are regular images, meaning no sketches or text descriptions, which places this in a close-world setting. The bounding boxes are generated by an object detector later explained. For the quantitative benchmark, the bounding boxes are processed beforehand whereas for the qualitative test and real-world setting, the bounding boxes need to be taken care of in real-time placing this as both closed and open-world setting.

A dataset is created to increase the performance in the new domain, more on this in chapter Dataset. The dataset contains enough labeled images to make a complete dataset. Therefore, supervised learning is the chosen method placing this in the closed-world setting in the benchmark scenario. In the qualitative test and real-world setting, the model needs to handle these bounding boxes, placing them in the open-world setting. More on this in coming chapters Data collection, Dataset, and LA-Transformer.

4.1 Camera Setup

The Padel court is generally an easy environment when it comes to detecting players since there is minimal occlusion. Occlusion is where an object or a person might block the view of another person. This generally happens more when e.g., looking a people walking on a crowded street and people crossing each other. In the current setup, the background is relatively consistent, compared to other environments where similar re-id techniques are operating in. One of these environments is the benchmarking dataset Market-1501. The dataset consists of frames, or partial frames, of people taken from an everyday setting in a supermarket at Tsinghua University (L. Zheng et al., 2015). While on a Padel court there is very little occlusion and only 4 people are on the court. Padel courts are typically the same color blue.

Figure 7, Market-1501 dataset versus Padel dataset

The optimal camera setup which is used in WPT should be 15,5 meters behind the court and elevated at 7,6 meters. The advantage of this is that it creates a good overview of the entire court and ensures that the players are always visible. The drawback here is that the metal structure that holds the Padel court together, including the walls, can cause some occlusion of the players. This is mostly hidden due to the smart placement of the camera that aligns some thicker parts of the walls to the bottom part of the net if seen from the perfect angle. But it can still cause some occlusion.

Unfortunately, not all courts have the possibility to create such a camera setup. Most courts and indoor courts are tightly placed next to each other. The solution then is to place the camera behind the back glass as far up as possible. The current solution utilizes a wide-angle camera lens to be able to fully capture the court when the camera is mounted closer. The short end of a Padel court typically consists of 5 glass panes. The camera lens is mounted in the middle of the middlemost glass pane. It is approximately 30cm away from the glass pane at a height of about 3m.

This way most of the court is visible. The biggest drawback with this setup is that the closest back corners are not fully visible. This will not likely cause one to miss any big parts of the game but remains a problem. The model will lose track of the players in this area. However, this camera setup will be used as its most common for recreational courts to use due to the limited space.

4.2 Data collection

Person re-identification requires images of the players in some way. The most used method is to have an object detector that extracts the images of the players. The role of the object detector is to locate and detect specific targets. This can then be set to only look for players, which then will provide coordinates within the image in the form of a bounding box around the player. Some examples of what this can look like can be seen in Figure 8. In the images the id of the person can be seen, the class which equals to person, and the confidence that it is the specific class. Additionally, the bounding box can be seen in which the class is predicted to be located.

Figure 8, Object detection of the same players in two different scenarios. The first two images from left to right show two different players who have been detected. The right-most image shows the two previous players are in the same image. The bounding box represents where the person is believed to be. They are labeled with an ID, object class, and YOLOv5's confidence that they are this specific class represented by the decimal number.

This method is used in two phases for this study. First off, it will be used in the data collection phase to collect images for the dataset. Thereafter, in the final testing phase, this method will serve to supply the re-id model with images of the players.

In the data collection phase, the object detection model, specifically "You Only Look Once" version 5 (Yolov5) together with Deepsort tracks the players. This gives a base tracker that is somewhat able to track the player during certain scenarios. The YOLO model can be applied to Padel scenarios where there is no id-swapping and clear visibility of all the players to prevent a player from getting a new id. The Yolov5 model together with Deep Sort can relatively easily extract images of the player and store them in separate folders depending on the id. One run on a short video should give 4 folders, named 1-4 and each folder should contain only images of 1 specific person. However, videos heavy with players occluding other players, entering blind spots, and leaving courts can cause players to be mislabeled. Therefore, some manual correcting will be required. Despite this, the method can collect data way faster than with fully manual labeling. Manual labeling consists of drawing a bounding box of an object e.g., a person, and then giving them an id or a class. This usually also generates some additional data on the coordinates of the person within that image. This information is redundant when training a person re-id model to learn specific identities.

The videos are run at 30 fps and there should be up to 4 people playing. Under the assumption that all persons are detected all the time, the model should be generating 30 x 4 = 120 images per second. If this is run on a video that is 1 minute, 120 * 60 = 7200 images collected and labeled.

Unfortunately, all the data is not "clean". The model is very good at tracking the person even though the person is not always fully visible. This causes many scenarios where the image saved is not of good quality which could worsen the result when applied to models. Some examples of this can be seen in Figure 9. The figure contains three images of player legs that could be very hard to predict.

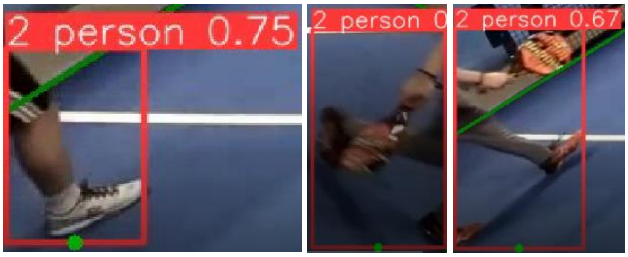

Figure 9, Object detection, examples of data that is bad. The first image from left to right contains a player's leg. The middle images also contain the player's legs bug with a racquet that blocks some of the legs. The third image contains a player's leg and some part of the forearm. These images contain less information about the player than a full-body image would.

This type of data is removed so that the model is not trained, validated, or tested on this. This is to reduce the amount of data that has less information contained. E.g., in Figure 9 the images could be very hard to correctly identify creating the possibility of training a worse model if these are used in the training process.

4.3 Datasets

Heeding the warnings of previous works where some algorithms can have trouble on entirely new domains it was decided that a new dataset was to be made specifically tailored for Padel. To gather sufficient data the process explained in the Data collection chapter was conducted on several videos to yield as much data as possible. It is important to have a lot of inter-class variation when making the dataset since the plan with the model was to prepare it for one-shot learning. Therefore, the datasets should contain different angles of the same player. This will increase the robustness of always identifying the correct player.

Padel dataset

The train folder contains 192130 images in 50 classes and the validation folder contains 82327 images in 50 classes making it a 70/30 split of the total 274457 images. After the split, the classes in the train folder contain between 1000-10000 images per class and in the validation folder, each class has ca 400-4400 images. Additionally, some testing images were taken that are of completely different players compared to the train and validation images. These images were first gathered, preprocessed, and then randomly sorted. When using the method in chapter Data collection to collect the images most of them look identical due to the speed at which the images are saved. Therefore, it was decided to randomly pick 99 images for each class of the ca 3000 images collected. This way every player should have as much variation as possible. The 99 images are then split 80/20 to gallery and query respectively over the 38 different captured classes. The testing set contains in total of 3762 images. Each image is resized to a 64x128x3 (height x width x color). The height and width are represented in pixels and the color is represented by the number of colors. 3 means that RGB (red, green, blue) colors are included. Each color is represented by a color depth of 8-bit. This translates to 256 shades per color

channel where any of the RGB colors can have a value between 0 and 255. Usually, the 0 is black or no color and the 255 is full brightness in the specific color. The disk size of the train/validation dataset is 1,04GB and the test set is 20,6MB. The choice of resolution (64x128) is inspired by the Market-1501 dataset (L. Zheng et al., 2015) where they also capture pictures of people in relatively the same proportion.

Market-1501 dataset

The Market-1501 dataset consists of 32668 labeled images spanning 1501 different identities officially (L. Zheng et al., 2015). The version of the dataset of which (Sharma et al., 2021) use contains 32036 images. The data is distributed in 4 different folders. A training folder containing 12151 images in 748 different classes, a validation folder containing 743 images in 743 classes, a gallery folder containing 19544 images in 764 classes, and a query folder containing 3326 images in 741 classes. The data distribution within the classes can vary from ca 3-55 images in train and gallery. In the query, the images are distributed so that every class has ca 3-6 images and in validation, every class has one image.

For clarity's sake, Figure 10 has been created to visualize the usage of the different datasets and the resulting models.

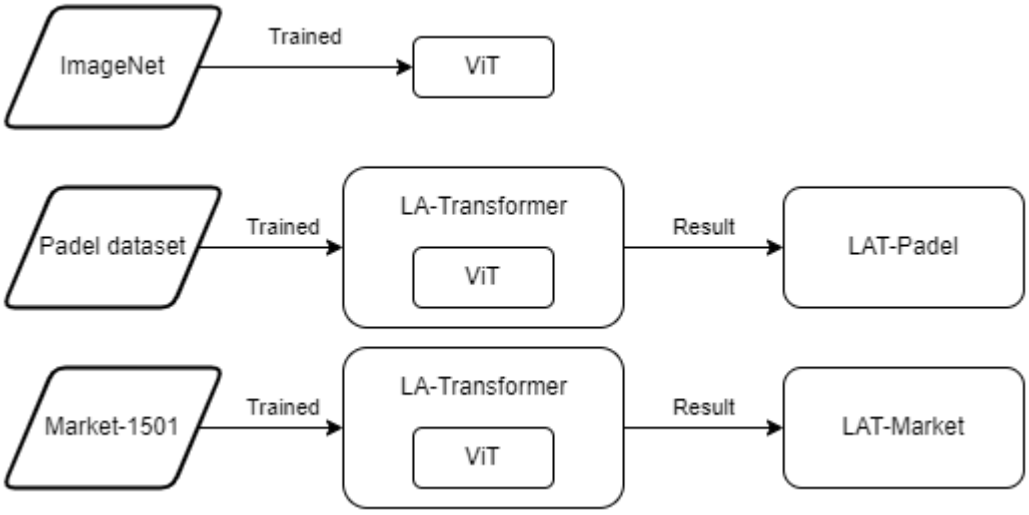

Figure 10 Visualization of dataset usage. LAT stands for LA-Transformer and the name stands for what dataset it has been trained on.

4.4 Pre-processing

When an image is taken, some adjustments should be made to help the model as much as possible before performing predictions. As mentioned in the Padel dataset chapter, the images are in the resolution of 64x128x3 (which stand for (height x width x color)). So, the optimal resolution for the images before the model predicts should be the same. During the training, both the training and validation folders are normalized individually and during inference, both the gallery and query folders are normalized individually. All the folders are normalized with a known mean of

[0.485, 0.456, 0.406] and standard deviation of [0.229, 0.224, 0.225]. They represent each value within the picture as stated above (height x width x color). These have been previously calculated on the ImageNet dataset which the ViT backbone has been trained on. The benefit of this is that the images from the datasets will now look more like the pictures the ViT has already trained on before. This will then increase the performance and reduce training time. The normalization brings all the values from potentially being sporadic to a known max and min value. This is a requirement for all pre-trained Torch models (Torch Contributors, n.d.). Furthermore, the images need to be resized to a resolution of 224x224 as the backbone ViT architecture takes images of this size. The query has an additional step which is a random horizontal flip that can randomly flip the input image. This is all based on the (Sharma et al., 2021) creator of the LA-Transformer.

4.5 LA-Transformer

The LA-Transformer is used with minimal adjustments. The same ViT backbone is used and further transfer-learned on the Padel dataset mentioned in chapter Datasets. More about the training process in chapter Training. As input during inference, the LA-Transformer receives an image from the so-called "query". The query consists of an image of a person who needs to be identified, visualized in Figure 11. The LA-Transformer then turns this query image into a feature vector. This query vector then gets compared to the vectors in the gallery of already known persons and determines which one it believes it to be. It then returns the class id along with the confidence value of the person. Each identity consists of a, or several, feature vectors.

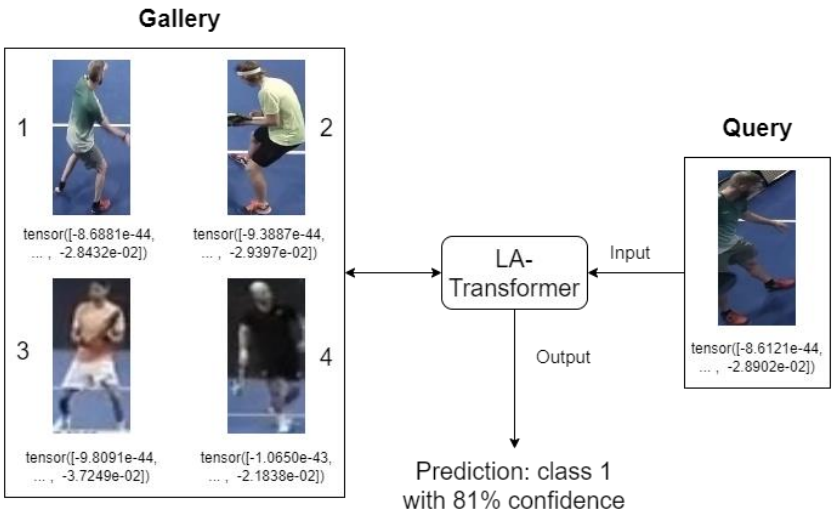

Figure 11, Visual representation of inputs and output through the LA-Transformer as a player in the form of images and feature vectors in tensor variables

These images are all gathered when processing the video. At the beginning of a video when all ids are known with the help of Deepsort, a burst of images is taken to establish a good knowledge base of each player in the gallery. Images can then be added as the processing goes on. To add more images to the gallery, two different

variables control the rate at how often this happens. Firstly, a counter is set on each player that increases every iteration. If and only if the counter reaches 120 or above, which equals four seconds an image will be saved. Secondly, if the model predicts a specific player with 90% confidence, an image of that person will be added to the class as further knowledge. This way the knowledge bases for the classes are allowed to increase and be updated without adding potentially wrong images.

However, there is a max size for the gallery that is set by a variable. When an image is added but the gallery is at the max size the oldest image will be removed and the newest image will be added. We call this technique Short Term Gallery Memory (STGM), and it is inspired by cache eviction policies commonly used in web development. The benefit of this technique is reduced inference time because only fresh images are stored and not every image. In this case, an STGM size of 10 images was used. This also ties back to the ethical concerns presented in the Ethics chapter of not storing images of players long-term.

The methods of adding new images during run time as mentioned previously in this chapter along with the STGM method are crucial to have a reasonable inference time. Adding images of every tracked player in every frame would quickly amount to a lot of images needing to be feature extracted. This would not only take a significant amount of time but also compound every second. Since the LA-Transformer is originally built to only execute on benchmarking datasets, it has no way of working with a flow of images throughout a video. The model assumes that the images in the gallery and query exist because it is in a close-world setting.

An interesting phenomenon appears when applying this technique which can be equated to the sliding window effect. The images in the STGM represent a recent timeframe of images ensuring that the LA-Transformer has a fresh memory of identities, consisting of the gallery folder. Since the images are recent it helps with the accuracy of the LA-Transformer. This allows the LA-Transformer to tackle problems that previously without the STGM were very challenging such as side switches between teams. Images are taken just before the players enter or leave the court or when they approach the camera. This allows for an easy transition if a player decides to leave the court and then reenter the playing field. An approximate visualization of how the gallery can look during player can be seen in Figure 12.

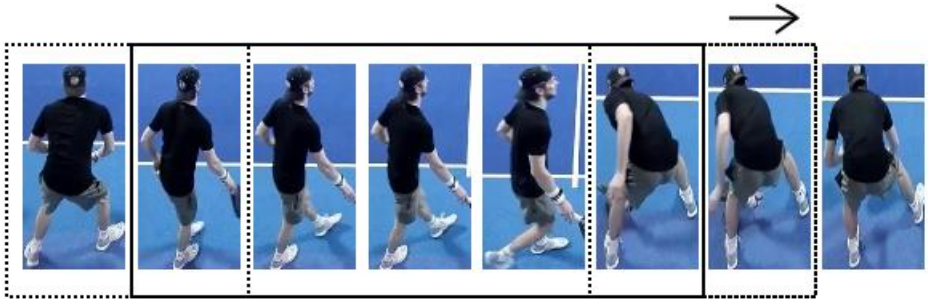

Figure 12, STGM sliding window visualization

4.6 One-shot learning

One thing that the LA-Transformer with the STGM implementation takes advantage of is One-Shot learning. In this case, it is when the model must learn an identity/player from a few images. After the short burst of images is done, as mentioned in chapter LA-Transformer. The model must learn every player's features uniquely so that new images that come in can be correctly identified to be one of those players.

4.7 Training

The original LA-Transformer is trained and evaluated on the Market-1501 dataset. In this study, the LA-Transformer is transfer learned on the same ViT backbone as the authors (Sharma et al., 2021) did. The same LA-Transformer is used with some modifications to allow this new data to work but the same training process is used. In the training process, the train and validation folders are used from the Padel dataset in chapter Datasets.

The batch size has been reduced to allow the model to complete the training through all 30 epochs. The authors (Sharma et al., 2021) mention that they use an "Nvidia RTX2080 TI with 11 GB VRAM, and 64 GB RAM". In this study, the hardware is not the same. The specification of the computer hardware used in this thesis is:

- NVIDIA GeForce GTX 1080 clocked at 1949MHz with 8192MB of accessible V-ram clocked at 5005MHz
- Mechanical Seagate hard drive at 7200rpm
- Intel i5-6800K
- 16GB of DDR4 ram clocked at 4200MHz. Corsair

During early testing and experimentation, the environment would crash due to a video memory limitation. Therefore, the batch was reduced to 16 instead of the original 32.

In training, the same block-wise finetuning technique is used. This is what allows (Sharma et al., 2021) to get high accuracy. At the start, the entire network is frozen. Every 2 epochs, they unfreeze one more layer of the network and multiply the learning rate by a learning rate decay value. Their learning rate starts at 0,0003 and the learning rate decay is 0,8. This same process of multiplying the learning rate by the decay and unfreezing the layers in the network goes on until the network is completely trained at 30 epochs. This helps in "mitigating the risk of catastrophic forgetting of the pre-trained weights" (Sharma et al., 2021, pp. 6) and keeping what the ViT has learned before.

Every time they have trained their model for one epoch, they evaluate it to see if the model has improved. Here the validation data is used and if it has improved, the weights are saved. The same saving method is used but modified a bit. Instead of overwriting the same weights, the weights a renamed uniquely so that all of them are saved in case of the need to go back. Every single epoch, the entire model is

also saved. This is so that in case the model starts to overfit, or anything goes wrong, the alternative exists to go back to a previous version. The advantage of saving the entire model is that you get all the needed values if training needs to resume at another time. This is especially important in this scenario since the learning rate is not static and neither are the frozen players. CrossEntropyLoss is the chosen loss function based on (Sharma et al., 2021).

While training the new version of the La-Transformer on the new dataset, the mentioned data distributions from chapter Datasets are used.

Important libraries and versions for the sake of reproducibility are as follows:

- Torch 1.10.2+cu113
- Torchaudio 0.10.2+cu113
- Torchvision 0.11.3+cu113
- Python 3.7.11
- Timm 0.5.4
- Numpy 1.21.5
- Tqdm 4.63.0
- Faiss-cpu 1.7.2
- Ipykernel 6.9.1
- Ipython 7.31.1
- Ipywidgets 7.6.5
- Pillow 9.0.1
- Pip 21.2.4
- Shapely 1.7.1

4.8 Testing

As described in the Evaluation metrics chapter the two primary and most conventional metrics used to evaluate the findings are CMC and mAP. The CMC rankings 1, 5, and 10 will be calculated along with an mAP. These will be presented in the upcoming Findings and analysis chapter.

Benchmarking

To perform a benchmarking of the models, many steps need to be done in the correct order. It begins with loading the data in the two required folders, a gallery, and a query. Then each folder is feature extracted separately, and this is what takes time. After the features have been extracted. The next large step is to concatenate the averaged GELTs. The technicalities of these steps exist in the chapter Locally Aware-Transformer. Once these steps have been done along with some minor adjustments, the similarities need to be calculated. (Sharma et al., 2021) use a library provided by (Johnson et al., 2021) called FAISS. With this, the result will be provided, which query images belong to which class with the most similarity.

5 Findings and analysis

In this chapter results and benchmarks between the original transfer learned weights (provided by the authors of the Locally Aware Transformer) (Sharma et al., 2021) and the transfer learned weights produced within this thesis will be presented. They will from now on be represented by either LAT-Market or LAT-Padel depending on which dataset they have been trained on. Both models will be analyzed to gain further knowledge about their strengths and weaknesses. The primary metrics that the models will be evaluated on are presented in the chapter Evaluation metrics.

To thoroughly evaluate the models both a quantitative and a qualitative evaluation will be done to fully explore the capabilities of both models. Firstly, a benchmark will be done on both datasets with both models. Secondly, a thorough evaluation will be done on 12 different videos with both models.

5.1 Benchmarks on datasets

Both models are tested on datasets in the same environment. Firstly, both models are evaluated by inference on the Market-1501 dataset. This is the same dataset as the authors (Sharma et al., 2021) trained their model on, the LAT-Market. Secondly, both models are tested on the Padel dataset created and mentioned in chapter Datasets. This is the same dataset that this study's model LAT-Padel is trained on. In Table 1 the testing results for the Market-1501 dataset are shown and in Table 2 the testing results for the self-made dataset are shown.

Table 1 Benchmarking results on the Market-1501 test dataset. Higher value is better

Model	**mAP**	**Rank1**	**Rank5**	**Rank10**
LAT-Padel	0,7288	0,7562	0,8679	0,8999
LAT-Market	0,9275	0,9827	0,9976	0,9982

Table 2 Benchmarking results on the created Padel test dataset. Higher value is better

Model	**mAP**	**Rank 1**	**Rank 5**	**Rank 10**
LAT-Padel	0,9439	0,9529	0,9889	0,9945
LAT-Market	0,9355	0,9446	0,9861	0,9931

The results from Table 1 shows how both the LAT-Padel produced in this thesis and the LAT-Market perform on the Market-1501 test dataset. It can be noted that performance increases significantly for both models as the top-k rises. Furthermore, a clear observation is that the LAT-Market performs ca 10-20% better than the LAT-Padel in general on the Market-1501 dataset. The mAP shows the LAT-Market performing 20% better than the LAT-Padel. Looking at the Ranks1-10 the score varies between ca 10 to 13%. On the Market-1501 test dataset, the winner is LAT-Market.

The results from Table 2 show how both LAT-Padel and the LAT-Market perform on the Padel dataset. It can be noted that performance increases significantly for both models as the top-k rises. Furthermore, the first clear observation is that the performance is similar between the two models. Even though it is close, the LAT-Padel outperforms the LAT-Market in every metric. In general, the difference between the scores is less than 1%. On the Padel test dataset, the winner is LAT-Padel.

5.2 Qualitative analysis

During testing, it was noted that the LAT-Padel had several occasions where it outperformed the LAT-Market when applying them to videos concerning Padel. This is hard to see in clear numbers, therefore the choice was made to also conduct a qualitative analysis of both models. 12 scenarios were extracted from Padel videos where it was suspected that both models could have a hard time identifying and tracking the correct players.

These videos contain multiple scenarios where the models will have to re-id a person. Out of the 12 scenarios, five of them specifically test reentering of the players into the court after leaving it. Three scenarios focus on reentering players as well as occlusion of players. Two scenarios focus on side switches of the teams where the two players from the far away side of the court come and play closer to the camera and nice verse with the other team. Two scenarios focus on a potential id switch where players move in front of each other. Depending on what the possible problem is with every video, the focus is slightly adjusted to what to focus on, but in general, the main goal with the videos is to look at the ids of the players and to see if they are consistent and correct.

Both authors participated in grading the clips on their performance. A grade scale was set up to mitigate the possibility that they would grade the clips inconsistently. One author solely graded the LAT-Padel model and the other the LAT-Market model. During grading, the authors could not see what the other author had graded on specific clips. This was done to also counteract any bias towards the model produced and trained in this thesis when it comes to grading. One thing that this does not prevent is the possibility of creating a bias per author. One author might be more generous when it comes to grading the scenarios than the other. This is something that unfortunately cannot be mitigated. As can be noted in Table 3 it was decided to not punish the model in the analysis for predicting someone as "UKN" or unknown identity. This is because it is better to be aware that the model is simply not confident enough about some identities rather than it predicting the wrong identity.

Table 3 Grading scale

Grade	Description
10	No ID jittering, Flawless result
5	Slight ID jittering (~1 second duration), or "UKN" prediction
1	Wrong prediction, excessive ID jittering

The clips were graded according to the Table 3 Grading scale. Numbers in between these are simply a mix of them. ID jittering is a phenomenon where the models get confused over a couple of frames causing them to switch the predicted ID, usually between two similarly looking identities, over a couple of frames.

Before getting into the results of the analysis there is one thing to highlight, a common problem with both models that were noted during the analysis. In the clips, there seem to be specific conditions that throw off both models, whether it is an inherent problem with the LA-Transformer or the ViT backbone is not clear. Footage of normal gameplay with immobile players either before the serve or when the ball is on the far side of the court both models appear to have problems with the tracking. The tracking/identity-tracking becomes unstable for a few frames causing ids to rapidly switch.

An example of a clip is of a person leaving the court and re-entering, a common occurrence in Padel. Both models appear to handle the re-entry well. The LAT-Padel appears to maintain a smoother track with minimal id flickering. It also handles post-occlusion well where one player is occluding the other player. A general difference between both models is that the LAT-Padel appears to be more confident. This is advantageous for this model since it allows for the model to store more frequent fresh images in the STGM allowing for more consistent tracking. This is also supported by the observation from the LAT-Market as it from time to time appears to be struggling with id flickering on tracked players

A trend that was noticed for both male and female players is that the majority appear to be dressed very similarly, with white shoes, black skirt/track pants, and a black t-shirt. This posed a problem for the authors when reviewing processed videos as they were in some cases not able to distinguish between players. On a positive note, the LAT-Padel was able to. This brings confidence that the LAT-Padel can feature extract and find useful features for identification purposes that humans might struggle with.

The main drawback of this system, in which the LA-Transformer exists, is that the model is fundamentally dependent on the YOLOv5 model to detect persons. If a person cannot be detected, no inference from the re-id model will be done which was noted in a brief part of one of the clips.

The result for all the scenarios is seen in Table 4. The table consists of all the scenarios listed row-wise. For each scenario, each model has been evaluated on that scenario and given a score by one of the authors. The final score of the models can be seen at the bottom and the result is 52 to 60 where the LAT-Padel wins over the LAT-Market. Purely looking at the final results the LAT-Padel seems to win a lot, however, this changes when looking at Figure 13.

Table 4 Grading of the scenarios

Scenario names	**LAT-Market**	**LAT-Padel**
1_person_out	5	7
1_person_reenter	8	8
2_reenters	9	8
reenter	1	5
corner_blindspot	3	6
play_occlusion_leaving_and_entering_court	1	2
occlusion_and_reenter (1)	9	7
occlusion_and_reenter	7	5
long_sideswitch	1	1
sideswitch_occlusion	1	1
id_switch_and_reenter	6	7
id_switch_test	1	3
Total points	52	60

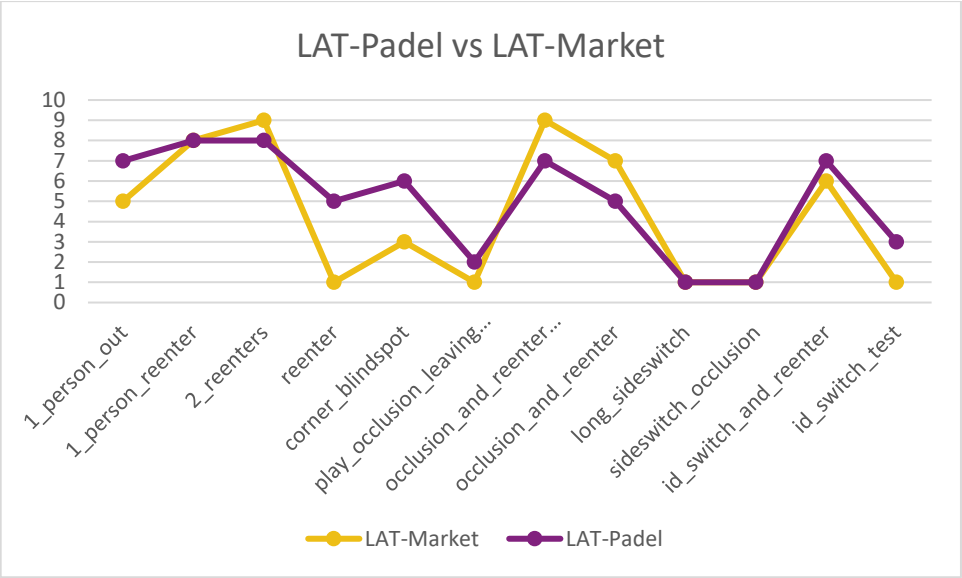

Figure 13 Results from the qualitative test. LAT-Padel versus LAT-Market

As can be seen in Figure 13 both models have been presented along with their grades on each clip that they were graded on. It can be noted that they outperform each other in different areas. As was mentioned previously in this chapter there exists a possibility of inconsistent grading between the two authors. Therefore, grades, where the difference is only 1, exist within the range of error. This means that grades, where the difference is more than 1, are of interest. Looking at the results it can be deduced that LAT-Padel outperformed LAT-Market on 4 clips and two of them quite significantly. LAT-Market outperformed LAT-Padel on 2 clips.

Attention Visualization

As machine learning has been spiraling toward bigger and deeper AI models, the trade-off has been predictive performance versus explainability. As the accuracy and/or performance have increased, the possibility to understand why the models do certain decisions had been lost. Figure 14 illustrates this, as the models used in this thesis are deep learning models, the hardest ones to understand. Even though this is the case, many attempts have been made to try to extract certain elements from the models to try to further understand some of the decisions. The ViT has been the subject of analysis for some time and many different researchers have tried to extract the attention layers from the model. Therefore, the decision was made to further analyze the backbone of the LA-Transformer, ViT, and investigate what captures the attention of the model. This was done using a modified version of the repository provided by (Gildenblat, 2020). By looking at heatmaps extracted from the attention layer in the backbone it is possible to speculate about the performance of the model, for example, if it needs more training. It can also be used to speculate about what the model is looking at the most or deem important in a picture when re-identifying an identity. In this chapter some generated heatmaps will be shown, analyzed, and speculated around to gain knowledge discussed previously.

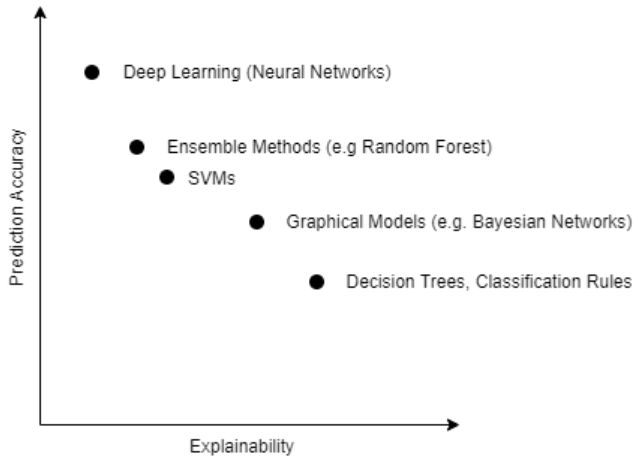

Figure 14 Predictive accuracy versus Explainability illustrated from (Dam et al., 2018)

Figure 15 Attention layers focusing on the back of the head and pants

In this scenario, Figure 15, the attention layers focus on the head of the players, as well as part of the pants. Here it seems to have been decided that the dark head is a good distinctive feature along with the dark pants.

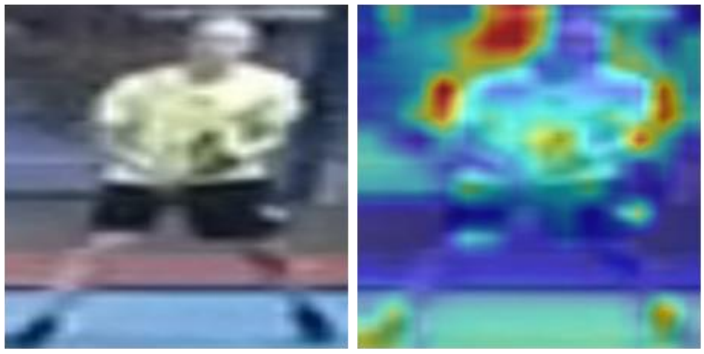

Figure 16 Attentions layers focusing on the background and forming a silhouette around the person

Figure 16 on the other hand shows less clear focus. The attention is focused on the background around the player. It forms a silhouette around the player showing that the player itself is not the main interest. The right shoulder, from the reader's point

of view, seems to have some interest along with the shoes. However, it can also be seen from another point of view that the model is creating an outline of the identities against the background.

Figure 17 Attention layers focusing on the top right pixels that stand out in red color

As can be seen in Figure 17, something else captures the attention of the model in the top right corner. Further investigating the picture reveals that the top rightmost corner of the picture reveals a small part of the red floor outside the court. Colorwise of this image this part stands out from the rest of the picture and it is believed what captures the attention of the model in this case.

Figure 18 Attention layers focusing on the back of the head

Moving on, in Figure 18 where the image doesn't contain any small areas of interest as in previously described images it can be seen that the model directs its attention to the player, primarily the upper body region and mostly the head. However small areas in the heatmap also reveal that a small amount of attention is given to both hands or lower arms as well as the feet.

Figure 19 Attention layers focusing on the upper back part of the player

Similarly as seen in Figure 18, Figure 19 shows the same characteristics. However, it is also possible to see that the attention also bleeds over into the net in the middle of the court in the background. A possible explanation could be that if the model is able to see a net in the background it can therefore know which side of the court the player is because the furthest away players (from the camera perspective) will not have a net behind them. In other words, a net in the background could be, according to the model, an important feature for this some identities.

Figure 20 Attention players focusing on the head and the background of the player

Once again in Figure 20, the head of the player seems to be of high importance. Additionally, the attention seems to lay a lot on the top background of the images. It could be that the model is looking for a net, or the letters on the net, as it is of high distinctive value, adding to the probability of it being the same player if the net, or letters on the net, exists.

As has been speculated around in this chapter the net appears to be an important feature other than the actual player. This could be the result of many different factors. With the nature of the LA-Transformer model and the usage of GELT, the token for each patch that the image gets split into is enhanced with global information which could be a reason why this behavior occurs. It could also be as speculated early that the net gives a good indication of the position of the player and effectively filters away 50% of the identities or simply an error in training, the reason remains inconclusive.

6 Discussion and conclusions

In this chapter, everything will be tied up. The method will be discussed including why some of the choices of this thesis were done. Some strengths and weaknesses along with some things that went well and less well. Additionally, the purpose and research questions will be reiterated to see if they have been answered.

A discussion of the findings will be held where the authors will state more subjective thoughts on the previous results. Here the purpose and the research questions will be listed and gone through individually to see what the results say about them.

And then the thesis will be finally concluded with some additional thoughts about some future work the authors think could be followed upon this thesis.

6.1 Discussion of method

Experimentation is something that is commonly used within the field of computer science and especially when it comes to testing out models within the field of Artificial Intelligence. Today there are a lot of different benchmarking datasets that are out there, and some of them have become standard to use when evaluating models. For example, in the field of object detection, the MS COCO (Lin et al., 2014) and ImageNet (Fei-Fei., 2009) are very well known. Benchmarking on these datasets is a relatively easy way of seeing how you are performing against other models within the same field. The same applies to the field of person re-id, the Market-1501 (L. Zheng et al., 2015) dataset is one of the most popular datasets within the field. Benchmarking on this immediately gives one an understanding of how well the model performs in the specific field on the same benchmark, dataset, and type of problem.

The purpose of this thesis is to see if it was possible to keep a consistent id of players throughout a Padel game. Firstly, a literature review was conducted to get an understanding of the topic and the field along with relevant models, benchmarking datasets, and evaluation metrics. Secondly, existing models were tested to see how they work. Through this, a lot of the technicalities were learned. During the early exploration, the first research questions were set to see how well the existing re-id models could perform in a new domain. After that, the second research question was created to further explore the possibility of improving the models in this new domain (Padel).

The way this thesis was conducted seemed the most reasonable. One of the goals was to see how well existing person re-id methods performed in Padel scenarios. Most literature in this field evaluates their models using evaluation metrics CMC and mAP, which were explained in the chapter Evaluation metrics. Therefore, it makes the most sense to use those same standards and metrics in the evaluation.

One additional contribution is the introduction of the developed STGM method presented in the chapter Locally Aware-Transformer. All the tests conducted in this thesis are possible because of this method. As mentioned before in chapter LA-Transformer, without the STGM the inference time of videos would be unreasonably high. This is because every saved image needs to be feature extracted in every frame for each player. These saved images quickly stack up to form a gallery size too big to have a reasonable inference time on an entire video. The STGM keeps the gallery size to a custom size where one can choose for more saved images in hopes of having better tracking with the trade of being inference time and vice versa. Due to how the LA-transformer was optimized to inference on data where all available images and identities were pre-known and not videos with dynamic identities this solution had to be implemented. This is because the LA-Transformer was built to work in a closed-world setting when it comes to data collection and bounding box generation.

One potential weakness is that this was only tried on one other existing model. This was mostly due to poor documentation and replicability from other authors. The authors would have happily tried more models if the possibility existed. Additionally, a more in-depth qualitative analysis would have been done if the idea was discovered earlier in the development process.

The authors think they have relatively achieved some of their goals. Looking at the two research questions, which will be answered in the chapter Discussion of findings, the results look promising from the chapter Findings and analysis. Although, the purpose of this thesis, was to keep track of player ids throughout a game seems quite hard today. However, comparing the solution produced within this thesis against the previous method of using YOLOv5 along with DeepSort, a significant improvement in re-identifying identities can be observed.

Training

As could be noted in chapter Findings and analysis, the LAT-Padel did not perform as well on the Market-1501 test dataset as the LAT-Market did on the Padel dataset. The Market-1501 dataset contains considerably hard data in terms of images. For example, the Market-1501 dataset has many images of partial persons/identities. When constructing the Padel dataset for this thesis a conscious decision was made to exclude images like these as can be seen in Figure 9. By doing this it is believed that the LAT-Padel may have become less robust towards partial images of persons, something that is very abundant in the Market-1501 dataset.

6.2 Discussion of findings

The purpose of the study is:

To investigate the applicability of existing state-of-the-art person re-id methods in the game of Padel, and how the methods can be improved to yield highly consistent and accurate performance in tracking all players through complete games of Padel.

and looking at the results from chapter Findings and analysis, this seems possible but not quite today. The benchmarking data from both Table 1 and Table 2 show promising result where the LAT-Market trained on the Market-1501 dataset seems to be very generalizable, and the LAT-Padel seem better on Padel data in terms of presented evaluation metrics. Although this looks good, real-world qualitative testing is a completely different case.

Looking at the first research question:

1. How well does an existing person re-id model perform on Padel matches when it comes to keeping a consistent and accurate id on all players.

The results from how the LAT-Market dataset did on the Padel dataset show surprisingly good results, achieving high scores in both mAP and CMC (Table 2). This means that the model that they have created is assumably good at generalizing to other domains such as Padel despite previous literature warning for bad performance in different domains. To further tie back into the first research question the results are mixed. The LAT-Padel can maintain a consistent track and accurate id on all detectable players. However, as mentioned in chapter Findings and analysis, id jitters can be observed but overall performance is consistent.

2. How can this model be improved upon to perform better in the new domain; Padel courts?

To improve the model, the LA-Transformer was trained on Padel data to improve results and to help it adapt to a new domain as mentioned in the chapter Training. Looking at results from both Table 1 and Table 2, it is shown that while the LAT-Padel took a substantial hit when it comes to being generalizable toward other datasets like the Market-1501, it managed to become more confident in its new domain. As discussed in chapter Discussion of findings, the LAT-Padel performs better on the Padel dataset in every metric including the mAP and CMC, and the LAT-Market performs better on the Market-1501 test dataset in every metric including the mAP and CMC. The results for the LAT-Padel are not quite as substantial in the benchmarking, but when it comes to the qualitative analysis, the LAT-Padel shows better results. The results gathered to evaluate the model only show results in terms of measuring metrics. Differences between LAT-Padel and LAT-Market could potentially have differences that can be hard to extrapolate by only looking at the measuring metrics, it was, therefore, appropriate to do a qualitative analysis to explore more tangible differences between the two.

The qualitative analysis was therefore conducted to further explore the strength and weaknesses between the two models on real-world unseen data. It was noted that generally, the LAT-Padel performed better or on par with the Market-1501 weights. However, it is quite remarkable that the LAT-Market has so good performance on the Padel dataset. It was not expected to perform that great because of what previous literature stated.

6.3 Conclusions

This thesis introduces the reader to the field of person re-identification. A thorough literature review has been conducted by searching some of the best-reviewed journals and conferences. Along with that, a purpose was set up including two research questions. A Transformer based model was chosen as the solution to the problems. This was then re-trained to work better in Padel scenarios on the dataset produced in this thesis. The results, first and foremost, suggest that the technology of person re-id significantly improves the results against a solution that relies on YOLOv5 and Deepsort for player tracking. Furthermore, the results suggest the novel LAT-Padel model outperforms the LAT-Market model concerning person re-identification in Padel games. However, when it comes to the Market-1501 dataset LAT-Market outperforms LAT-Padel by 20% mAP and 10-13% CMC. To further analyze LAT-Padel and LAT-Market a qualitative analysis was done. The results suggest that LAT-Padel which was specifically trained to identify Padel players performed better than LAT-Market. However, it was also noted that both models were affected by unknown and unseen influences which need to be further investigated in the future.

This thesis shows that it is possible to use person re-identification to enhance player tracking in Padel through both a quantitative and qualitative evaluation.

Future work

Throughout the work on this thesis, many ideas for future work have come up. Things that could have been done better or in a different way. We also elaborate on ideas of adapting existing methods in surveillance to computer vision in sports.

As discussed in the Discussion of method chapter it appears that the weights produced in this thesis are not as robust as models trained on other datasets. Therefore, it would be interesting to include images where the person is not always fully visible such as the ones depicted in Figure 9 in the training process to see if they have any effect on the robustness of the model. Alternatively, conduct transfer learning on the already trained Market-1501 weights to potentially benefit from the robustness of these weights. Furthermore, due to hardware constraints with running out of VRAM during training the batch size was decreased from 32 to 16. Batch size is an important hyperparameter that has a big role when it comes to training a network. A batch size that is too high can cause the network to have problems with achieving convergence (Kandel & Castelli, 2020). Too small of batch size on the other hand can cause the network to go back and forth without ever achieving an acceptable performance (Kandel & Castelli, 2020). However, a small batch size can

converge faster than larger batch size. Despite this, a larger batch size can approach or reach global minima which a small batch size cannot reach (Kandel & Castelli, 2020). Future work could therefore experiment more with larger or equal batch sizes to 32.

In this thesis, the method of utilizing STGM is presented. This is to minimize the inference time on each frame for each player. Future work could therefore explore the possibility of saving the feature vectors of players in a variable in RAM. In this way features for each identity can be appended to this variable faster because only one image needs to be feature extracted instead of each image stored in the gallery. This would speed up the inference process.

Additionally, a third dataset introduced into the thesis would have been interesting to see. For example, CHUK01 and CHUK02 (W. Li et al., 2014) are popular re-id datasets. They contain 971 identities over 3884 images and 1816 identities over 7264 images respectively. Testing the model produced in this thesis along with the model that was trained on the Market-1501 dataset on the CHUK datasets would have been interesting to see how they perform in a domain which none of them have been trained for.

Another interesting evaluation would have been to run the different models on an evaluation video. Metrics such as Multi Object Tracking Accuracy (Milan et al., 2016) would have been an interesting addition to the quantitative results. Since the authors believe that the produced model in this thesis did a better job qualitatively, maybe they also would have had a greater gap in accuracy compared to the other models trained on the Market-1501 dataset when it comes to keeping a consistent id on the players.

Finally, the creation of the qualitative analysis was realized too late in the time plan of this thesis. Therefore, the quality and validity suffered as a cause of this. In the future, a more systematic approach for this shall be taken to make sure any sort of bias is completely removed e.g.,

- Show a clip to the reviewer and not tell which model has produced that clip
- Get a third-party reviewer that does not have any bias
- Invite more reviewers and collect more results and then average the results to get a more confident opinion about the models
- Utilize more videos and scenarios to get a further understanding of the capabilities of the models

7 References

Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2018). Understanding of a convolutional neural network. *Proceedings of 2017 International Conference on Engineering and Technology, ICET 2017*, *2018-Janua*, 1–6. https://doi.org/10.1109/ICEngTechnol.2017.8308186

Almazan, J., Gajic, B., Murray, N., & Larlus, D. (2018). *Re-ID done right: towards good practices for person re-identification*. http://arxiv.org/abs/1801.05339

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). *Layer Normalization*. http://arxiv.org/abs/1607.06450

Baevski, A., & Auli, M. (2018). Adaptive Input Representations for Neural Language Modeling. *7th International Conference on Learning Representations, ICLR 2019*, 1–13. http://arxiv.org/abs/1809.10853

Bak, S., Zaidenberg, S., Boulay, B., & Bremond, F. (2014). Improving person re-identification by viewpoint cues. *2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 175–180. https://doi.org/10.1109/AVSS.2014.6918664

Barrett, D. (2013). *One surveillance camera for every 11 people in Britain, says CCTV survey*. The Telegraph. https://www.telegraph.co.uk/technology/10172298/One-surveillance-camera-for-every-11-people-in-Britain-says-CCTV-survey.html

Beal, J., Kim, E., Tzeng, E., Park, D. H., Zhai, A., & Kislyuk, D. (2020). *Toward Transformer-Based Object Detection*. http://arxiv.org/abs/2012.09958

Bedagkar-Gala, A., & Shah, S. K. (2014). A survey of approaches and trends in person re-identification. *Image and Vision Computing*, *32*(4), 270–286. https://doi.org/10.1016/j.imavis.2014.02.001

Bender, D., & Dori, Y. (2019). *Lawsuit Alleges That Self-Checkout Videos Violate the Song-Beverly Act*. EMERGING TECHNOLOGIES, LITIGATION. https://www.insideprivacy.com/united-states/litigation/lawsuit-alleges-that-self-checkout-videos-violate-the-song-beverly-act/

Bewley, A., Ge, Z., Ott, L., Ramos, F., & Upcroft, B. (2016). Simple online and realtime tracking. *Proceedings - International Conference on Image Processing, ICIP*, *2016-Augus*, 3464–3468. https://doi.org/10.1109/ICIP.2016.7533003

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-End Object Detection with Transformers. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12346 LNCS*, 213–229. https://doi.org/10.1007/978-3-030-58452-8_13

Cho, Y.-J., & Yoon, K.-J. (2016). Improving Person Re-identification via Pose-Aware Multi-shot Matching. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, *2016-Decem*, 1354–1362. https://doi.org/10.1109/CVPR.2016.151

Cireşan, D. C., Meier, U., Masci, J., & Gambardella, L. M. (2013). Flexible, High Performance Convolutional Neural Networks for Image Classification. *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence Flexible*, 1237–1242. https://www.aaai.org/ocs/index.php/IJCAI/IJCAI11/paper/viewFile/3098/3425

Dam, H. K., Tran, T., & Ghose, A. (2018). *Explainable Software Analytics*. http://arxiv.org/abs/1802.00603

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. http://arxiv.org/abs/2010.11929

ElitePadel. (2022). *Statistik padel i mars 2022_ antalet padelbanor, padelspelare och padelhallar*.

European Data Protection Board. (2019). *Facial recognition in school renders Sweden's first GDPR fine*. https://edpb.europa.eu/news/national-news/2019/facial-recognition-school-renders-swedens-first-gdpr-fine_sv

Farenzena, M., Bazzani, L., Perina, A., Murino, V., & Cristani, M. (2010). Person re-identification by symmetry-driven accumulation of local features. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2360–2367. https://doi.org/10.1109/CVPR.2010.5539926

Fei-Fei., J. D. W. D. R. S. L.-J. L. K. L. L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. *CVPR09*. http://www.image-net.org/papers/imagenet_cvpr09.bib

Gheissari, N., Sebastian, T. B., & Hartley, R. (2006). Person Reidentification Using Spatiotemporal Appearance. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06)*, *2*, 1528–1535. https://doi.org/10.1109/CVPR.2006.223

Gildenblat, J. (2020). *Explainability for Vision Transformers (in PyTorch)*. https://github.com/jacobgil/vit-explain

Google Scholar. (n.d.). *Top publications*. Computer Vision & Pattern Recognition. Retrieved February 8, 2022, from https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng_computervisionpatternrecognition

Gray, D., & Tao, H. (2008). Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 5302 LNCS* (Issue PART 1, pp. 262–275). https://doi.org/10.1007/978-3-540-88682-2_21

Grother, P., Ngan, M., & Hanaoka, K. (2019). Face Recognition Vendor Test (FRVT) Part 3 : Demographic Effects. *Nistir 8280*, *December*, https://doi.org/10.6028/NIST.IR.8280.

He, S., Luo, H., Wang, P., Wang, F., Li, H., & Jiang, W. (2021). *TransReID: Transformer-based Object Re-Identification*. https://doi.org/10.1109/iccv48922.2021.01474

Hinton, G. E. (2010). Rectified Linear Units Improve Restricted Boltzmann Machines. *Proceedings of the 27th International Conference on Machine Learn- Ing*, *3*, 807–814.

Hirzer, M., Roth, P. M., Köstinger, M., & Bischof, H. (2012). Relaxed Pairwise Learned Metric for Person Re-identification. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 7577 LNCS* (Issue PART 6, pp. 780–793). https://doi.org/10.1007/978-3-642-33783-3_56

Hou, R., Ma, B., Chang, H., Gu, X., Shan, S., & Chen, X. (2019). VRSTC: Occlusion-Free Video Person Re-Identification. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, *2019-June*, 7176–7185. https://doi.org/10.1109/CVPR.2019.00735

Huang, H., Li, D., Zhang, Z., Chen, X., & Huang, K. (2018). Adversarially Occluded Samples for Person Re-identification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5098–5107. https://doi.org/10.1109/CVPR.2018.00535

Huang, Y., Zha, Z., Fu, X., & Zhang, W. (2019). Illumination-Invariant Person Re-Identification. *Proceedings of the 27th ACM International Conference on Multimedia*, 365–373. https://doi.org/10.1145/3343031.3350994

IBM. (n.d.). What is Computer Vision? | IBM. In *What is computer vision?* https://www.ibm.com/topics/computer-vision

Ioffe, S., & Szegedy, C. (2015). *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. http://arxiv.org/abs/1502.03167

Johnson, J., Douze, M., & Jegou, H. (2021). Billion-Scale Similarity Search with GPUs. *IEEE Transactions on Big Data*, *7*(3), 535–547.

https://doi.org/10.1109/TBDATA.2019.2921572

Journal Rankings. (n.d.). *Scimago Journal & Country Rank*, *Computer V*. https://www.scimagojr.com/journalrank.php?category=1707

Kandel, I., & Castelli, M. (2020). The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset. *ICT Express*, *6*(4), 312–315. https://doi.org/10.1016/j.icte.2020.04.010

Karanam, S., Li, Y., & Radke, R. J. (2015). Person Re-Identification with Discriminatively Trained Viewpoint Invariant Dictionaries. *2015 IEEE International Conference on Computer Vision (ICCV)*, *2015 Inter*, 4516–4524. https://doi.org/10.1109/ICCV.2015.513

Koptelov, A. (2021). *Facial Recognition and GDPR: How to Stay Compliant?* The European Business Review. https://www.europeanbusinessreview.com/facial-recognition-and-gdpr-how-to-stay-compliant/

Kostinger, M., Hirzer, M., Wohlhart, P., Roth, P. M., & Bischof, H. (2012). Large scale metric learning from equivalence constraints. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, *Ldml*, 2288–2295. https://doi.org/10.1109/CVPR.2012.6247939

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, *25*. https://arxiv.org/pdf/1003.0358.pdf

Li, S., Xiao, T., Li, H., Yang, W., & Wang, X. (2017). Identity-Aware Textual-Visual Matching with Latent Co-attention. *2017 IEEE International Conference on Computer Vision (ICCV)*, *2017-Octob*, 1908–1917. https://doi.org/10.1109/ICCV.2017.209

Li, W., Zhao, R., Xiao, T., & Wang, X. (2014). DeepReID: Deep Filter Pairing Neural Network for Person Re-identification. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 152–159. https://doi.org/10.1109/CVPR.2014.27

Li, X., Zheng, W., Wang, X., Xiang, T., & Gong, S. (2015). Multi-Scale Learning for Low-Resolution Person Re-Identification. *2015 IEEE International Conference on Computer Vision (ICCV)*, 3765–3773. https://doi.org/10.1109/ICCV.2015.429

Li, Z., Lv, J., Chen, Y., & Yuan, J. (2021). Person re-identification with part prediction alignment. *Computer Vision and Image Understanding*, *205*(November 2020), 103172. https://doi.org/10.1016/j.cviu.2021.103172

Liao, S., Hu, Y., Xiangyu Zhu, & Li, S. Z. (2015). Person re-identification by Local Maximal Occurrence representation and metric learning. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, *07-12-June*(June), 2197–2206. https://doi.org/10.1109/CVPR.2015.7298832

Liao, S., & Li, S. Z. (2015). Efficient PSD Constrained Asymmetric Metric Learning for Person Re-Identification. *2015 IEEE International Conference on Computer Vision (ICCV)*, *2015 Inter*, 3685–3693. https://doi.org/10.1109/ICCV.2015.420

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 8693 LNCS* (Issue PART 5, pp. 740–755). https://doi.org/10.1007/978-3-319-10602-1_48

Marr, B. (2019). 7 Amazing Examples Of Computer And Machine Vision In Practice. In *Forbes*. https://www.forbes.com/sites/bernardmarr/2019/04/08/7-amazing-examples-of-computer-and-machine-vision-in-practice/?sh=51788dc01018

Martinel, N., Foresti, G. L., & Micheloni, C. (2019). Aggregating Deep Pyramidal Representations for Person Re-Identification. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, *2019-June*, 1544–1554. https://doi.org/10.1109/CVPRW.2019.00196

Matsukawa, T., Okabe, T., Suzuki, E., & Sato, Y. (2016). Hierarchical Gaussian Descriptor for Person Re-identification. *2016 IEEE Conference on Computer Vision and Pattern*

Recognition (CVPR), *2016-Decem*, 1363–1372. https://doi.org/10.1109/CVPR.2016.152

Milan, A., Leal-Taixe, L., Reid, I., Roth, S., & Schindler, K. (2016). *MOT16: A Benchmark for Multi-Object Tracking*. 1–12. http://arxiv.org/abs/1603.00831

Nambiar, A., Bernardino, A., & Nascimento, J. C. (2020). Gait-based Person Re-identification. *ACM Computing Surveys*, *52*(2), 1–34. https://doi.org/10.1145/3243043

Padgham, L., Lee, Y., Sadiq, S., Winikoff, M., Fekete, A., MacDonell, S., & Reid, I. (n.d.). *CORE Rankings Portal*. Retrieved February 8, 2022, from https://www.core.edu.au/conference-portal

Pillai, G. (2012). *Caught on Camera: You are Filmed on CCTV 300 Times a Day in London*. International Business Times. https://www.ibtimes.co.uk/britain-cctv-camera-surveillance-watch-london-big-312382

Plantinga, A. (1961). Things and Persons. *The Review of Metaphysics*, *14*(3), 493–519.

Raina, R., Madhavan, A., & Ng, A. Y. (2009). Large-scale deep unsupervised learning using graphics processors. *ACM International Conference Proceeding Series*, *382*. https://doi.org/10.1145/1553374.1553486

Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., & Shlens, J. (2019). Stand-Alone Self-Attention in Vision Models. *Advances in Neural Information Processing Systems*, *32*(NeurIPS), 1–13. http://arxiv.org/abs/1906.05909

Sarfraz, M. S., Schumann, A., Eberle, A., & Stiefelhagen, R. (2018). A Pose-Sensitive Embedding for Person Re-identification with Expanded Cross Neighborhood Re-ranking. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 420–429. https://doi.org/10.1109/CVPR.2018.00051

Sharma, C., Kapil, S. R., & Chapman, D. (2021). *Person Re-Identification with a Locally Aware Transformer*. http://arxiv.org/abs/2106.03720

Song, C., Shan, C., Huang, Y., & Wang, L. (2021). Mask-guided contrastive attention and two-stream metric co-learning for person Re-identification. *Neurocomputing*, *465*, 561–573. https://doi.org/10.1016/j.neucom.2021.09.038

Stekler, H. O., Sendor, D., & Verlander, R. (2010). Issues in sports forecasting. *International Journal of Forecasting*, *26*(3), 606–621. https://doi.org/10.1016/j.ijforecast.2010.01.003

Torch Contributors. (n.d.). *MODELS AND PRE-TRAINED WEIGHTS*. Retrieved April 12, 2022, from https://pytorch.org/vision/stable/models.html

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2020). *Training data-efficient image transformers & distillation through attention*. 1–22. http://arxiv.org/abs/2012.12877

Vargas, V. M., Guijo-Rubio, D., Gutiérrez, P. A., & Hervás-Martínez, C. (2021). *ReLU-Based Activations: Analysis and Experimental Study for Deep Learning* (E. Alba, G. Luque, F. Chicano, C. Cotta, D. Camacho, M. Ojeda-Aciego, S. Montes, A. Troncoso, J. Riquelme, & R. Gil-Merino (Eds.); Vol. 12882, pp. 33–43). Springer International Publishing. https://doi.org/10.1007/978-3-030-85713-4_4

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *2017-Decem*(Nips), 5999–6009.

Wang, H., Zhu, Y., Green, B., Adam, H., Yuille, A., & Chen, L.-C. (2020). Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 12349 LNCS* (pp. 108–126). https://doi.org/10.1007/978-3-030-58548-8_7

Wang, Q., Li, B., Xiao, T., Zhu, J., Li, C., Wong, D. F., & Chao, L. S. (2019). Learning Deep Transformer Models for Machine Translation. *Journal of New Media*, *2*(4), 137–148.

https://doi.org/10.32604/jnm.2020.014278

Wang, T., Gong, S., Zhu, X., & Wang, S. (2014). Person re-identification by video ranking. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8692 LNCS*(PART 4), 688–703. https://doi.org/10.1007/978-3-319-10593-2_45

Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local Neural Networks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 7794–7803. https://doi.org/10.1109/CVPR.2018.00813

Wang, Y., Wang, L., You, Y., Zou, X., Chen, V., Li, S., Huang, G., Hariharan, B., & Weinberger, K. Q. (2018). Resource Aware Person Re-identification Across Multiple Resolutions. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8042–8051. https://doi.org/10.1109/CVPR.2018.00839

Wei, L., Zhang, S., Yao, H., Gao, W., & Tian, Q. (2017). *GLAD : Global-Local-Alignment Descriptor for Pedestrian Retrieval*.

Wojke, N., Bewley, A., & Paulus, D. (2018). Simple online and realtime tracking with a deep association metric. *Proceedings - International Conference on Image Processing, ICIP*, *2017-Septe*, 3645–3649. https://doi.org/10.1109/ICIP.2017.8296962

Wu, A., Zheng, W.-S., Yu, H.-X., Gong, S., & Lai, J. (2017). RGB-Infrared Cross-Modality Person Re-identification. *2017 IEEE International Conference on Computer Vision (ICCV)*, *2017-Octob*, 5390–5399. https://doi.org/10.1109/ICCV.2017.575

Wu, D., Zheng, S. J., Zhang, X. P., Yuan, C. A., Cheng, F., Zhao, Y., Lin, Y. J., Zhao, Z. Q., Jiang, Y. L., & Huang, D. S. (2019). Deep learning-based methods for person re-identification: A comprehensive review. *Neurocomputing*, *337*, 354–371. https://doi.org/10.1016/j.neucom.2019.01.079

Xiong, F., Gou, M., Camps, O., & Sznaier, M. (2014). Person Re-Identification Using Kernel-Based Metric Learning Methods. In *Eccv* (pp. 1–16). https://doi.org/10.1007/978-3-319-10584-0_1

Yaghoubi, E., Kumar, A., & Proença, H. (2021). SSS-PR: A short survey of surveys in person re-identification. *Pattern Recognition Letters*, *143*, 50–57. https://doi.org/10.1016/j.patrec.2020.12.017

Yang, Y., Yang, J., Yan, J., Liao, S., Yi, D., & Li, S. Z. (2014). Salient Color Names for Person Re-identification. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 8689 LNCS* (Issue PART 1, pp. 536–551). https://doi.org/10.1007/978-3-319-10590-1_35

Ye, M., Liang, C., Wang, Z., Leng, Q., Chen, J., & Liu, J. (2015). Specific Person Retrieval via Incomplete Text Description. *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, 547–550. https://doi.org/10.1145/2671188.2749347

Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., & Hoi, S. C. H. (2022). Deep Learning for Person Re-Identification: A Survey and Outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(6), 2872–2893. https://doi.org/10.1109/TPAMI.2021.3054775

Yu, H.-X., Wu, A., & Zheng, W.-S. (2020). Unsupervised Person Re-Identification by Deep Asymmetric Metric Embedding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *42*(4), 956–973. https://doi.org/10.1109/TPAMI.2018.2886878

Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X., & Tang, X. (2017). Spindle Net: Person Re-identification with Human Body Region Guided Feature Decomposition and Fusion. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, *2017-Janua*, 907–915. https://doi.org/10.1109/CVPR.2017.103

Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., & Tian, Q. (2016). MARS: A Video Benchmark for Large-Scale Person Re-Identification. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 9910, pp. 868–884).

Springer International Publishing. https://doi.org/10.1007/978-3-319-46466-4_52

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., & Tian, Q. (2015). Scalable person re-identification: A benchmark. *Proceedings of the IEEE International Conference on Computer Vision*, *2015 Inter*, 1116–1124. https://doi.org/10.1109/ICCV.2015.133

Zheng, W.-S., Gong, S., & Xiang, T. (2011). Person re-identification by probabilistic relative distance comparison. *CVPR 2011*, 649–656. https://doi.org/10.1109/CVPR.2011.5995598

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2021). A Comprehensive Survey on Transfer Learning. *Proceedings of the IEEE*, *109*(1), 43–76. https://doi.org/10.1109/JPROC.2020.3004555

Zou, Y., Yang, X., Yu, Z., Kumar, B. V. K. V., & Kautz, J. (2020). Joint Disentangling and Adaptation for Cross-Domain Person Re-Identification. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 12347 LNCS*. Springer International Publishing. https://doi.org/10.1007/978-3-030-58536-5_6