Programa de Pós-Graduação em

Computação Aplicada

Mestrado Acadêmico

Eduardo Souza dos Reis

PastLens: Granting Temporal Consistency to Multi-person Pose Estimation Through Longer Receptive Fields

São Leopoldo, 2019

Eduardo Souza dos Reis

PASTLENS: GRANTING TEMPORAL CONSISTENCY TO MULTI-PERSON POSE ESTIMATION THROUGH LONGER RECEPTIVE FIELDS

Dissertação apresentada como requisito parcial para a obtenção do título de Mestre pelo Programa de Pós-Graduação em Computação Aplicada da Universidade do Vale do Rio dos Sinos — UNISINOS

Advisor:
Prof. Dr. Rodrigo da Rosa Righi

São Leopoldo
2019

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)

Reis, Eduardo Souza dos

PastLens: Granting Temporal Consistency to Multi-person Pose Estimation Through Longer Receptive Fields / Eduardo Souza dos Reis — 2019.

70 f.: il.; 30 cm.

Dissertação (mestrado) — Universidade do Vale do Rio dos Sinos, Programa de Pós-Graduação em Computação Aplicada, São Leopoldo, 2019.

"Advisor: Prof. Dr. Rodrigo da Rosa Righi, Unidade Acadêmica de Pesquisa e Pós-Graduação".

1. Human Pose Estimation. 2. Multi-person Scenarios. 3. Receptive Fields. 4. Spatio-temporal Features. I. Título.

CDU 004.732

Bibliotecária responsável: Amanda Schuster — CRB 10/2517

Eduardo Souza dos Reis

PastLens: Granting Temporal Consistency to Multi-person Pose Estimation Through Longer Receptive Fields

Dissertação apresentada à Universidade do Vale do Rio dos Sinos – Unisinos, como requisito parcial para obtenção do título de Mestre em Computação Aplicada.

Aprovado em 28/02/2019

BANCA EXAMINADORA

Manuel Menezes de Oliveira Neto - UFRGS

Cristiano André da Costa - UNISINOS

Luiz Gonzaga da Silveira Junior - UNISINOS

Prof. Dr. Rodrigo da Rosa Righi (Orientador)

Visto e permitida a impressão
São Leopoldo, 15/04/2019

Prof. Dr. Rodrigo da Rosa Righi
Coordenador PPG em Computação Aplicada

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior Brasil (CAPES) - Código de Financiamento 001 / This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001

ABSTRACT

Accurately estimating poses of multiple individuals in unconstrained scenes would improve many vision-based applications. As a few examples: person re-identification, human-computer interaction, behavioral analysis and scene understanding. Through the advancements on convolutional networks' research, body part detectors are now accurate and can estimate spatial positioning on still images in real-time (30 FPS), for both single- and multi-person scenarios. In turn, multiple individuals interacting in videos impose additional challenges, such as person-to-person occlusion, truncated body parts, additional assignment steps and more sources for double counting. In the last few years, many advancements contributed towards this goal and partially solved some of these challenges. Nonetheless, dealing with long-term person-to-person occlusion is not possible in still images, due to the lack of discriminative features to detect the occluded individual. Most reviewed works solve this problem by collecting motion features that correlate body parts across multiple video frames, exploring temporal dependency. Usually, these approaches either rely only on adjacent frames to keep it close to real-time or process the whole video beforehand, imposing global consistency in an offline manner. Since most of the cited applications rely on near real-time processing in combination with complex human motions, which are not depicted in just a couple frames, we propose the PastLens model. Our main objective is to provide a cost-efficient alternative to the tradeoff between the number of correlated frames and the estimation time. The model impose spatio-temporal constraints to the convolutional network itself, instead of relying on arbitrary designed temporal features. We stretch the receptive field of the mid layers to also include the previous frame, forcing further layers to detect features that correlate poses across the two frames, without losing the per-frame configuration. Moreover, we do not constraint the representation of such features, allowing it to be learned throughout the training process, alongside the pose estimation. By pose estimation and tracking, we refer to the localization and tracking overtime of head, limbs and torso, followed by the assembling of these body parts into poses that correctly encode the scene. We will not evaluate our approach on benchmarks for facial keypoints or gesture recognition. PoseTrack is the dataset of choice for both training and validation steps, since it provides a publicly available benchmark for estimating and tracking poses, in addition to a leaderboard that enable direct comparison of our results with its state-of-the-art counterparts. Experimental results indicate that our model can reach competitive accuracy on multi-person videos, while containing less operations and being easier to attach to pretrained networks. Regarding scientific contributions, we provide a cost-efficient alternative to impose temporal consistency to the HPE pipeline, through receptive field increase only, letting the temporal features' representation to be learned from data. Hence, our results may lead towards novel ways of exploring temporal consistency for human pose estimation in videos.

Keywords: Human Pose Estimation. Multi-person Scenarios. Receptive Fields. Spatio-temporal Features.

RESUMO

Estimar com precisão poses de vários indivíduos em cenas sem restrições beneficiaria muitas aplicações de visão computacional. Alguns exemplos: reidentificação de pessoas, interação humano-computador, análise comportamental e compreensão de cenas. Através dos avanços na arquitetura das redes convolucionais, detectores de partes do corpo são precisos e podem estimar o posicionamento das partes em imagens estáticas em tempo real (30 FPS), tanto em cenários com uma única pessoa, quanto com várias pessoas interagindo. No entanto, vários indivíduos na mesma cena impõem desafios adicionais, como oclusão inter-pessoas, partes do corpo truncadas, etapas adicionais de atribuição e mais fontes para contagem dupla. Nos últimos anos, muitos avanços contribuíram para esse objetivo e resolveram parcialmente alguns desafios. Mas, lidar com a oclusão inter-pessoas de longo prazo não é possível em imagens estáticas, devido à falta de atributos discriminativos para detectar o indivíduo ocluído. A maioria dos trabalhos revisados resolve esse problema coletando atributos de movimento que relacionam as partes do corpo em vários quadros do vídeo, explorando a consistência temporal. Normalmente, esses trabalhos consideram apenas quadros adjacentes para correlacionar esses atributos em tempo real ou processam o vídeo inteiro de antemão, para depois impor consistência global de maneira *offline*. Como a maioria das aplicações citadas dependem do processamento próximo ao tempo real, em conjunto com a análise de movimentos humanos complexos, que não são detectáveis em poucos quadros, propomos o modelo PastLens. Nosso principal objetivo é prover uma alternativa com melhor custo benefício do que a atual escolha entre o número de quadros a serem correlacionados e o tempo de estimação. O modelo impõe restrições espaço-temporais à própria rede convolucional, ao invés de depender de atributos temporais definidos de forma arbitrária. Nós alongamos o campo receptivo das camadas intermediárias para também conter o quadro anterior, forçando as camadas subsequentes a detectarem atributos que correlacionam as poses nos dois frames, sem perder a configuração das poses quadro-a-quadro. Além disso, nós não restringimos a representação destes atributos, permitindo que a mesma seja aprendida durante o processo de treinamento, juntamente com a estimativa de poses. Por estimativa e rastreamento de poses, nos referimos à localização e rastreamento da cabeça, membros e torso, seguidos da combinação dessas partes em poses que descrevam corretamente a cena. Nós não avaliaremos nossa abordagem em *benchmarks* para *keypoints* faciais ou reconhecimento de gestos. PoseTrack é a base de dados escolhida para avaliar nosso modelo, visto que ela fornece uma referência pública para estimação e rastreamento de poses em videos, além de um quadro de classificação, permitindo a comparação direta de nossos resultados com o estado-da-arte. Os resultados dos experimentos indicam que nosso modelo atinge acurácia competitiva nos vídeos com múltiplas pessoas, mas contém menos operações e é mais fácil de adaptar para modelos pré-treinados. Em relação as contribuições científicas, nós disponibilizamos uma alternativa eficiente para impor consistência temporal à estimativa de poses humanas usando apenas o aumento dos campos receptivos, deixando que a representação de atributos temporais seja definida pelos dados. Assim, nossos resultados podem levar a novas formas de explorar a consistência temporal na estimativa de poses humanas em vídeos.

Palavras-chave: Estimativa de Poses Humanas. Cenários com Múltiplos Indivíduos. Campos Receptivos. Atributos espaço-temporais.

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

HPE	Human Pose Estimation
CNN	Convolutional "Neural" Networks
PAF	Part Affinity Fields
TFF	Temporal Flow Fields
NMS	Non-maximum Suppression
ILP	Integer Linear Programming
DSC	Depthwise Separable Convolution
CONV	Regular Convolution

CONTENTS

1 INTRODUCTION

Estimating human poses[1] is a fundamental task for many trending Computer Vision applications, namely: human-robot interaction, gaming, sports analysis, surveillance, scene understanding, virtual reality, driver-assisting, activity recognition and healthcare (HOLTE et al., 2012; YANG; RAMANAN, 2013; SHOTTON et al., 2013; SUN; KOHLI; SHOTTON, 2012; TOSHEV; SZEGEDY, 2014; DANTONE et al., 2014; SARAFIANOS et al., 2016; CHEN et al., 2017). Many important milestones have been achieved over the last decade, but hard challenges such as occlusion and unexpected poses, still remain (GIRSHICK et al., 2011; IQBAL; GALL, 2016). These challenges persist due to the nonlinear and highly dimensional nature of the human motion space, resulting in a plethora of different poses. Also, variety of both appearance and background clutter for real-world scenarios is very large. Abstracting these difficulties to enable generic models is not trivial and may lead to the loss of useful contextual information.

A well established way to overcome such challenges is the use of markers that reliably track human subjects' motion (HOLTE et al., 2012; SIGAL; BLACK, 2010). Yet, these marker-based solutions are expensive, intrusive and not applicable to most unconstrained scenarios, in which the subjects are not cooperative. Therefore, outside the gaming or visual effects industries, most applications need markerless (vision-based) solutions (LIU et al., 2015). With respect to markerless human pose estimation, the task can be defined as determining precise pixel location of body keypoints, from single or multi-image inputs, followed by the estimation of the human subjects' poses (NEWELL; YANG; DENG, 2016). It assumes that all the scene information is captured by external sensors, such as RGB or RGBD capable cameras. In addition, it is a cheaper, although less accurate, solution compared to its marker-based counterpart.

Even though local evidence from image patches is important for identifying body parts, a coherent final pose estimate requires kinematic knowledge, due to the degrees of freedom on human articulations and self or external occlusion. Thus, human pose estimation (HPE) heavily relies on contextual information (TOSHEV; SZEGEDY, 2014; NEWELL; YANG; DENG, 2016). The need for context led to a shift from local body part detectors, followed by spatial reasoning, to strong context-aware detectors, such as convolutional networks (IQBAL; GALL, 2016; INSAFUTDINOV et al., 2016; WEI et al., 2016; NEWELL; DENG, 2016). Through community efforts, nowadays there is enough annotated data available to train these networks for both single- and multi-person pose estimation tasks (ANDRILUKA et al., 2014; LIN et al., 2014; IQBAL; MILAN; GALL, 2016; GÜLER; NEVEROVA; KOKKINOS, 2018).

After witnessing great advancements in the field due to the application of convolutional networks as the standard body part detectors, researchers turned their attention towards unconstrained multi-person scenes. Mainly due to the greater range of real world applications that rely on such scenes. In cluttered environments, where multiple individuals may occlude each

[1] All images used to illustrate cluttered scenes in this work are licensed under a Creative Commons Attribution 4.0 International license.

Figure 1: Multi-person pose estimation in a cluttered environment. Many of the common challenges in such scenes are depicted, such as ambiguous body part assignments, different parts sharing the same image region and partial part occlusion. Nonetheless, person-to-person occlusion remains as the main offender. For instance, most body parts of the occluded individual (highlighted with checkers) are partially or fully occluded, providing insufficient appearance features to be detected as valid candidates.

Source: Elaborated by the author.

other, estimating coherent poses is a hard task, since small errors in the estimates may lead to false pose configurations or miscount the number of persons (PAPANDREOU et al., 2017). The studied literature defines the goals of multi-person body pose estimation as threefold: (i) to assign every image pixel to a body part or to the background (LADICKY; TORR; ZISSERMAN, 2013); (ii) discover the number of human subjects in the image; and (iii) assign these detected body parts to pose configurations that correctly encode the scene. In the remainder of this work, pose estimation is used to refer to the body pose estimation variant.

We focus our research on the methods that deal with challenges particular to scenes containing multiple interacting individuals. In addition, we constrain the input to color images, or video sequences, depicting a single view of the scene. The main challenges for these scenarios are listed on Figure 1, proposing an operation room as a cluttered environment instance. Such challenges led the state-of-the-art towards the use of video temporal information and to the hard task of tracking poses overtime. Without any association between multiple frames, person-to-person occlusion, truncation, varying number of individuals and motion blur bring many false estimates. Therefore, temporal consistency is required to improve pose estimation both locally frame-by-frame and globally throughout the whole video. In order to provide this temporal consistency, we redefine the HPE task as highlighting spatio-temporal dependencies between body parts across adjacent video frames, instead of separately estimating poses frame-by-frame and later correlating them. The result is the PastLens model, that contributes to the HPE research

as a cost-efficient alternative to the state-of-the-art multi-person pose estimation pipeline, in addition to not relying on arbitrarily designed temporal features.

1.1 Motivation

Motivated by all the real world applications listed earlier, many researches showed interest for the area, resulting in remarkable advancements and an exponential growth in the capacity of body part detectors and trackers. Nonetheless, most methods still struggle with self-symmetries and there is no clear trend towards which parts of the pose estimation pipeline are crucial to this increase in the accuracy of the final estimates (XIU et al., 2018). For instance, the best methods on MPII pose estimation benchmark (ANDRILUKA et al., 2014) have many differences regarding how to approach the problem, but achieve similar accuracy (XIAO; WU; WEI, 2018). With respect to research opportunities, person-to-person occlusion is still a critical unsolved challenge for multi-person pose tracking in cluttered or crowded scenes (ANDRILUKA; ROTH; SCHIELE, 2008; XIAO; WU; WEI, 2018; DOERING; IQBAL; GALL, 2018).

There are two ways to ensure temporal consistency in videos: online or offline. Online approaches estimate poses in real-time (30 frames per second) by correlating the current frame to past data, aiming at reducing temporal inconsistencies between adjacent frames. In contrast, offline approaches enforce global consistency by estimating poses frame-by-frame for the entire video and enhancing these estimates through historical data afterwards. Although many works have been conducted after the release of the first big and publicly available dataset for multi-person pose tracking (ANDRILUKA et al., 2017), tracking poses overtime is still a largely unaddressed problem. Most state-of-the-art works rely on a two steps approach: an off-the-shelf pose estimator is used in combination with a simple heuristic for person association overtime (e.g. color features or dense temporal connectivity), conducted in an offline manner (INSAFUTDINOV et al., 2017; GIRDHAR et al., 2017). Just a handful of methods try to leverage features designed specifically for online tracking of poses (XIU et al., 2018; DOERING; IQBAL; GALL, 2018). However, these fast online methods usually consider two or three adjacent frames to guarantee temporal consistency, which is too short of a time interval for most, if not all, complete human motions. They are limited to this short time interval due to their architecture scaling linearly to the number of frames being correlated. For instance, a couple of frames add a whole new branch to the convolutional network, while a third one would lead to a similar increase in complexity.

Unfortunately, current methods propose highly refined convolutional networks to extract multi-frame motion cues and ensure temporal consistency. Regardless of these methods reaching state-of-the-art results on both image and video based benchmarks, there is no clear understanding of how or by how much each of these refined modules contribute towards the extraction of knowledge on the human kinematics; nor a clear path towards scaling these methods to enable longer sequences of frames to be processed online. We redeem a valid contribution to the

current literature any sort of temporally consistent estimator that do not depend on tailor-made modules nor limits the representation of temporal features to a fixed encoding. Even more remarkable would be to achieve competitive results by relying only on spatio-temporal image evidence, since it would allow a better scaling of current architectures towards longer frame sequences.

1.2 Research Question

The PastLens model try to answer the following research question: *How would a model that enforces spatio-temporal dependencies and learns how to encode temporal features from data perform on the task of estimating poses of multiple interacting persons in videos?*

As aforementioned, tracking body part motion across just a few frames is not enough to capture longer and more complex human motions. In addition, long-term occlusions are common in unconstrained videos and may last for several frames. Current motion features are too heavy complexity-wise to be calculated over many frames in real-time. Hence, it may be easier to ensure temporal consistency by estimating poses based solely on the spatio-temporal features, relying less on local (per-frame) image evidence. We frame HPE as estimating poses for the current frame given its correlation to the previous one, rather than its current spatial configuration. By doing that we can assert if the model learned to penalize unfeasible pose sequences and estimate sequences of poses or just approximated a frame-by-frame estimation.

1.3 Objectives

In this work, we propose to deal with the task of augmenting current pose estimation models by stretching their receptive field to also include previous frames, therefore, estimating spatial-temporal relations between frames, rather than just detecting image regions that most likely contain body parts. The main objective of this research project is: *train a model to estimate poses by evaluating the likelihood of pose sequences, relying solely on spatio-temporal features extracted by a feature merging module*.

Choosing the right way to measure similarity between a pair of poses or body parts across time is fundamental to the task of estimating persons' poses in videos, but it becomes even more impactful in multi-person scenarios. As an example, location based metrics as PCKh (ANDRILUKA et al., 2014) or IoU (LIN et al., 2014) are very accurate for single image estimation, but assume that spatial changes will be smoothly overtime, struggling when there is either rapid movement or abrupt camera changes, including scale variations due to the camera zoom (DOERING; IQBAL; GALL, 2018). Regarding appearance based metrics, such as optical flow, they are unable to handle some types of appearance changes such as occlusion or truncation. Exploring the output of current convolutional architectures for HPE, we propose PastLens: a multi-person human pose estimation model that augments single-frame estimation through

spatio-temporal features learned direct from data, imposing less constraints to their representation. Further, we keep it unaware of pretrained models' details, enabling it to be used on top of most state-of-the-art work.

To achieve this main objective, it was further reduced to a sequence of smaller secondary objectives:

i Review the basic concepts and techniques of human pose estimation and tracking;

ii Research the state-of-the-art on multi-person pose tracking and evaluate their reported results on publicly available datasets;

iii Choose the best features between the ones proposed by related works to ensure temporal consistency in an online manner;

iv Based on the chosen feature, iterate over possible adaptations to enable the model to better scale to longer time intervals;

v Develop and train a convolutional network that accurately assigns body parts to individuals, using our proposed spatio-temporal features;

vi Evaluate our model on a publicly available and broadly adopted benchmark.

To the best of our knowledge, there is no model proposed in the literature that try to implicitly encode temporal dependence of sequential pose configurations. Moreover, we keep the model simple and attachable to other bigger and more complex models, keeping it closer to an augmentation module rather than a completely new architecture, enabling further evaluation on other state-of-the-art convolutional networks. Evaluating the accuracy of these estimates in a reproducible way is another project decision that severely modify the final results. We adopted a public available and largely used benchmark called MS-COCO (LIN et al., 2014) to evaluate single-frame estimates and the PoseTrack benchmark (ANDRILUKA et al., 2017) for training on video sequences.

1.4 Text Structure

The remainder of this study is organized as follows: Chapter 2 introduces the reader to the basic concepts behind multi-person human pose estimation in videos, exploring the omnipresent convolutional networks as current state-of-the-art body part detectors. Chapter 3 discusses the reviewed literature and presents a comparative analysis of the works closer related to our own model. Chapter 4 details the PastLens architecture and justifies project decisions. Chapter 5 describes the adopted evaluation methodology, proposes test scenarios, discusses experimental results and evaluates the effectiveness of our model in a publicly available benchmark. Finally, Chapter 6 brings the final remarks, limitations and open issues for future work.

2 BACKGROUND

HPE is the task of estimating the 3D configuration (pose) of a human body given a sensor input or the projection of this pose in a 2D image (SHOTTON et al., 2013; BOGO et al., 2016). Human bodies can be modeled as collections of rigid parts connected by flexible joints. This model is represented as a set $P = \{p_{head}, p_{neck}, ..., p_{left-ankle}\}$, that encodes a discrete collection of body parts. For example, regarding body estimation, the set typically contains head, neck, shoulders, elbows, hands, torso, hips, legs, knees and ankles; being further augmented with a set of per-part configurations, including position, appearance and scale, but without facial or gestural features, such as the mouth's or finger's motion. The task of estimating these per-part configurations can be formulated as a structured optimization problem, which approximates a set of parameters to minimize a cost function measuring the similarity between the image I and the space of possible poses resulting from these configurations (RAMAKRISHNA et al., 2014). In other words, the objective is to approximate a set of image patches $B = \{I_{p1}, I_{p2}, ..., I_{pn}\} \subset I$ that correctly encode the positions of the human body parts in I. By connecting this set of parts in a graph structure (kinematic chain) it is possible to impose kinematic constraints, such as: the neck being directly connected to the head; length ratios between parts; or different degrees of freedom for the joints (CHEN; WEI; FERRYMAN, 2013). These constraints confine the search space to a smaller set of valid poses, reducing complexity.

Body pose estimation can be further divided into two categories: full or upper body pose estimation. Full body methods estimate the set of joints defined earlier. Upper body variants discard lower limbs and hips since most images available in the media only picture the upper half of the human body and it also encodes the most informative joints (EICHNER et al., 2012). On both scenarios, the problem of estimating poses, with an array of calibrated cameras, constrained environments, and willing individuals is mostly solved (HOLTE et al., 2012; SIGAL; BLACK, 2010). Yet, there are approaches that rely on a small set of cameras as sensors, without markers attached to the human subject nor his cooperation. These are the targets of this study due to the broader range of applications they enable. Furthermore, we focus our attention on methods that explicitly deal with multi-person scenarios, i.e. scenes with multiple individuals interacting with each other and their surroundings.

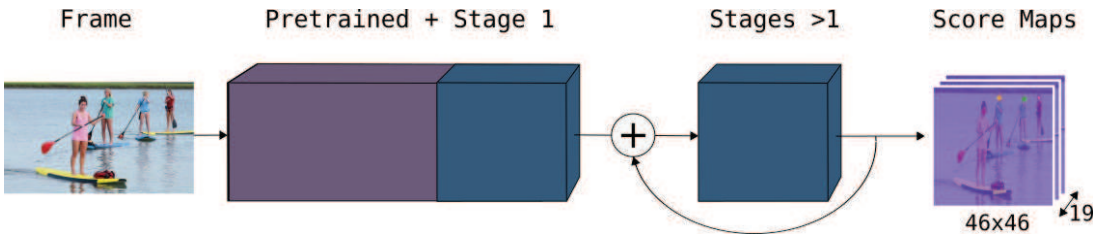
2.1 Convolutional Networks as Body Part-Detectors

HPE methods prior to the rise of deep learning and the convolutional networks (CNNs), relied on the extraction of hand-crafted low and mid-level features, including HOG blocks, edges, contours, and color histograms (BOURDEV; MALIK, 2009; DANTONE et al., 2014; SHOTTON et al., 2013; TAYLOR et al., 2012; YANG; RAMANAN, 2013). The current literature trend is to apply deep learning discriminative methods, which outperforms most part-based models (TOMPSON et al., 2014; BELAGIANNIS et al., 2016a; CHEN et al., 2017). Among

these approaches, most of them use convolutional networks to output a heatmap, which encodes a per-pixel likelihood for joint locations in the image (TOSHEV; SZEGEDY, 2014; TOMPSON et al., 2014; RAMAKRISHNA et al., 2014; CAO et al., 2016). This map is called confidence, belief or score map and usually represents a single joint probability distribution over the whole image. For the remainder of this document we will refer to these as score maps. Figure 2 provides an example architecture for this pipeline, it is worth noticing the Gaussian peaks identifying the correct head positions on the output score map. Discriminative methods received more attention because it is harder to map body parts to precise pixel locations (regression) than mapping them to Gaussian peaks in score maps (YANG et al., 2017). With respect to body pose estimation, this shift from regressing to vectors of part locations to estimating these peaks, led to the main advancements during recent years (TOMPSON et al., 2014; MEHTA et al., 2017).

CNNs are collections of layers that extract an hierarchy of features from images, through a set of filters optimized with supervised learning. In summary, a standard network is composed of convolutional layers that apply a convolution operation over every input image channel, followed by a non-linear transformation, usually ReLU (KRIZHEVSKY; SUTSKEVER; HINTON, 2012). As a short example, we can describe a scenario where the first layer receives a color image $3 \times 340 \times 220$ and outputs a collection of 64 feature maps $64 \times 336 \times 216$. We suppose that the layer has 64 filters with size 5×5. The scenario also assumes a stride, the number of pixels that the filter is dislocated at each step of the convolution, of 1 and stops when the filter reaches a pixel outside the image boundaries. Convolutional layers may also be interleaved with pooling (downsampling) layers, in order to reduce memory consumption. Increasing the stride discards more pixels, resulting in an alternative downsampling process used by fully convolutional networks (no pooling layers).

Figure 2: A deep learning pipeline broadly adopted as body part detector in current literature: state-of-the-art network (purple) for image classification, such as ResNets (HE et al., 2016) or Mobilenets (HOWARD et al., 2017), are trained in a prediction-refinement manner (WEI et al., 2016). The first stage (blue) receives an image as input while deeper stages receive the image evidence and previous stages‘ outputs. After a sequence of refinement steps, the final output is a set of per-joint score maps. We assume 19 body parts, 46×46 score maps.



Source: Elaborated by the author.

The size of its filters defines a layer's receptive field: the bigger it is, the more context information it provides, enabling the resulting feature maps to be more context-aware. More precisely, each position (activation) in a feature map will be a high value if the feature it encodes is detected in the receptive field it was calculated from, and a low or zero value otherwise,

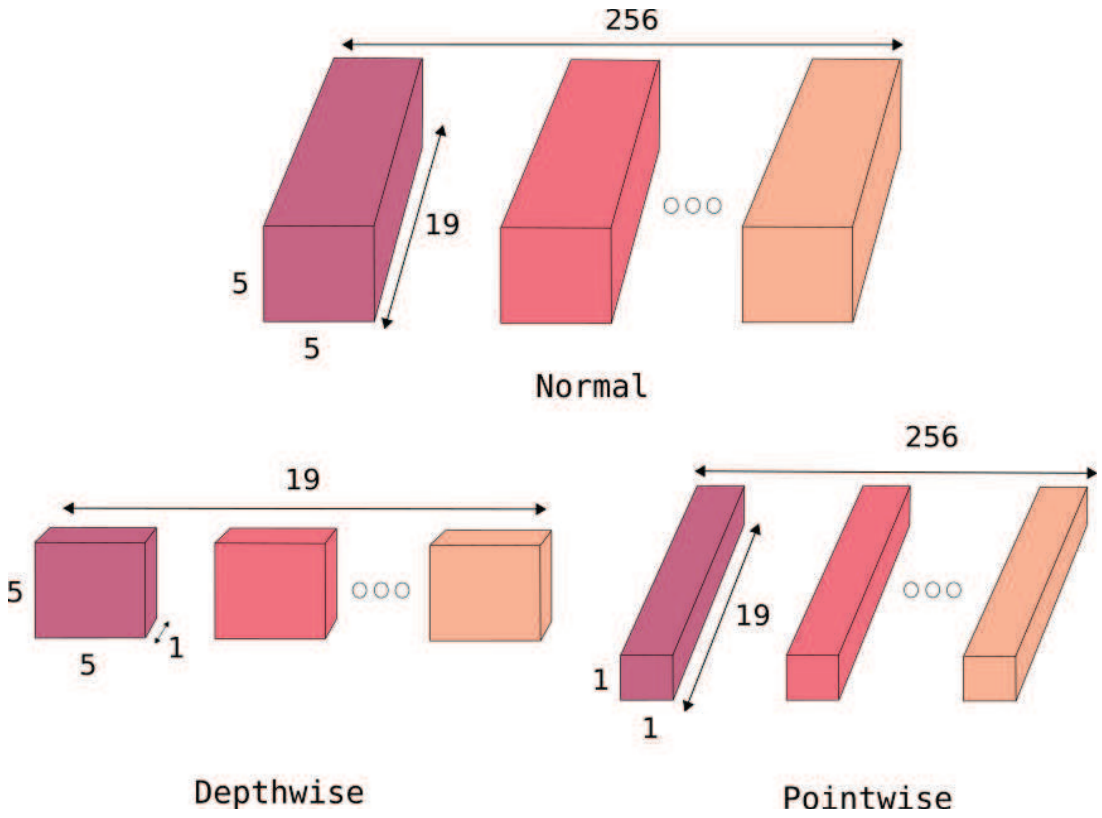
assuming a ReLU non-linearity function. Therefore, by increasing this receptive field size, the calculated features can encode a bigger and more informative image region. Still, based on the aforementioned scenario, in practice, for color images there would be $3 \times 5 \times 5$ filters per feature map on the first layer l_1, one for each color channel. Further layers would have $C \times 5 \times 5$ filters per feature map, where C is the number of maps from the previous layer. Regarding the training process, an advantage of the convolutional networks is being fully differentiable, allowing the use of backpropagation to adjust the filter weights (WEI et al., 2016). The final network output is a set of high-level features maps encoding semantic knowledge.

In the human pose estimation context, convolutional networks can capture context information due to not being reliant on local part detection (TOSHEV; SZEGEDY, 2014). Additionally, they do not need a feature engineering step, because filter weights are adjusted to find the most discriminatory feature set directly from data. This property makes the problem of estimating coherent poses considerably easier since well-known feature descriptors, including HOG and color histograms, may fail to encode different viewpoints or limb deformation. In order to effectively explore these properties, the network architecture requires bigger receptive fields, which can be achieved in the following ways: pooling at the expense of precision; increasing the filter size while also increasing the parameter space; or by increasing the number of convolutional layers, resulting in a deeper architecture (WEI et al., 2016). The current trend is to use fully convolutional networks, apply downsampling through bigger strides and increase the receptive fields with a deeper architecture (NEWELL; YANG; DENG, 2016; CHEN et al., 2017). However, stacking too many layers on CNNs may result in vanishing gradients (HE et al., 2016). A general solution to this problem is proposed by HE et al. (2016) and tailored for structured problem, such as pose estimation, by Wei et al. (2016). Their strategy is similar to skip-connections (MAO; SHEN; YANG, 2016), enforcing intermediate local supervisions, i.e. every stage has its own loss layer. It is an effective solution for pose estimation, therefore we use it in our own body part detector.

2.1.1 Depthwise Separable Convolutions

Regarding the CNN adopted as a base for our model, its main advantage is the use of a factorized version of convolutional layers, reducing the number of computations and parameters (HOWARD et al., 2017). This particular approach is called Depthwise Separable Convolution, widely used by current literature and based on the works of Jaderberg, Vedaldi e Zisserman (2014); Lebedev et al. (2014). Depthwise refers to the fact that it does not decompose a convolution operation spatially (filterwise), instead, it divides the whole process into two steps: the *depthwise* convolution and the *pointwise* convolution. During the *depthwise* step the input's depth, e.g. number of channels of an image is kept and a different filter is used for each of these channels. Later, the next step is to apply a *pointwise* convolution using 1×1 filters, convolving in a point-by-point manner. It uses as many filters as the output's number of channels and has

Figure 3: Comparison of a normal convolutional layer and the two steps that compose Depthwise Separable Convolutions. This example assumes an input of dimensions 46×46 with 19 channels and 5×5 filters for the current layer. Further, the desired output must have 256 channels and stride 1, resulting in a $42 \times 42 \times 256$ output. Using a regular convolution the layer needs $256 \times 5 \times 5 \times 19 \times 42 \times 42 = 214.502.400$ operations. In contrast, the *depthwise* step needs only $19 \times 5 \times 5 \times 42 \times 42 = 837.900$ and the *pointwise* $256 \times 1 \times 1 \times 19 \times 42 \times 42 = 8.580.096$, adding up to $9.417.996$ operations.



Source: Adapted from Howard et al. (2017).

a depth equal to the inputs's. An example showing the significant decrease in the number of operations per-layer is portrayed in Figure 3.

With respect to accuracy, the experiments reported by Howard et al. (2017) shows that these factored convolutions reach similar overall accuracy to the regular ones in tasks such as image classification, captioning or HPE, while keeping the number of operations an order of magnitude lower. Nonetheless, reducing the number of parameters have the downside of being prohibitive to shallow networks or the ones with small number of feature maps per-layer, since these characteristics would decrease the already limited number of variable weights available to the network during training.

2.2 Multi-Person Pose Estimation and Tracking

Defining human pose estimation as the task of finding the correct configuration for body parts of a single, pre-localized, person was a driver for progress. Current research efforts try to

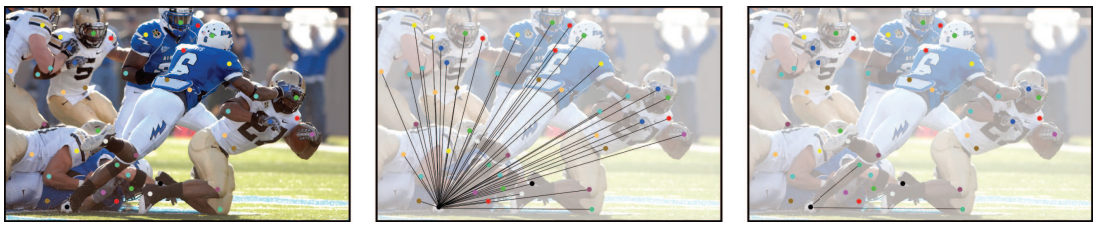
solve the multi-person variant: estimating poses of an unknown number of interacting individuals. This view is closer related to most real-world applications, since crowded environments may be unavoidable. Mainly for computational reasons, most research prior to 2013 assume that the individuals are correctly pre-localized (YANG; RAMANAN, 2013). Keeping this assumption, most approaches can be adapted to multi-person images. However, this section focus on methods tailored specifically for estimating poses of multiple individuals in still images or videos, since they greatly outperform naive approaches in practice (CAO et al., 2016). Transitioning from single to multi-person is hard due to many inherent challenges: unknown positioning, different scales, person-to-person occlusion, partial visibility, contact (pose merging), truncation (escaping the image boundaries) and a much larger state space. There is also the double counting problem: some of the body parts do not have many distinguishing features (e.g. elbows) or may be hard to assert positioning (left vs right hands), thus, the network may assign the same image region to both right and left score maps or to multiple individuals (VARADARAJAN; DATTA; TICKOO, 2017). Regarding the unknown number of individuals, it makes the segmentation step harder and increases the computational complexity as it gets bigger. While many advancements happened in recent years, occlusion and background complexity remain as open issues (CHEN et al., 2017).

Multi-person human pose estimation has been tackled in two ways so far: person detector followed by single-person pose estimators (top-down) or joint detection followed by a part-assignment step (bottom-up). Top-down approaches can also be seen as the generation of object proposals followed by the highlighting of keypoints. Hence, the pipeline is very similar to segmentation tasks, assuming a scenario where the object class is constrained to people (HE et al., 2017; PAPANDREOU et al., 2017). Although more common, top-down approaches solely rely on the human detection algorithm precision and may assign the same body part instance to multiple overlapping individuals (PISHCHULIN et al., 2016). In addition, the run-time is proportional to the number of people in the image, since more people equal more executions of the single-person pose estimator. As a consequence, bottom-up approaches, which assemble the body parts into coherent poses after a global part-detection stage, may lead to faster results (CAO et al., 2016). Nonetheless, in crowded scenes assigning the located body parts to coherent poses, as well as associating these poses to the correct individuals, can be a quite difficult task and may result in ambiguous poses because people are too close together (FANG et al., 2017; PAPANDREOU et al., 2017). Figure 4 depicts the differences between the two strategies.

Our approach is closer related to the bottom-up pipeline, thus we will pay more attention to it. Approaching multi-person pose estimation in a bottom-up fashion usually rely on an initial body parts detection step, which generate score maps or unary potentials (pictorial-structures) for every visible part disregarding the number of individuals in the image (PAPANDREOU et al., 2018). Following the pipeline, relations, usually pairwise, between these part estimates are encoded in a graph structure. This framework represents an instance of the minimum cost

Figure 4: The top row illustrate the standard top-down approach for tackling multi-person pose estimation: each person in the image is tracked, then a single-person pose estimation method detect body parts and estimate each pose configuration individually. Most approaches also need an additional step to remove outliers because of imperfect segmentation. Regarding bottom-up approaches (bottom row), a part-detector identifies every body part in the image, a dense graph connecting all body parts is established and a multicut subgraph partitioning problem is defined to reduce this set of edges to a subset that correctly encode the poses. Given the computational complexity of reducing this densely connected graph, it is necessary to apply contextual knowledge like in the right-most picture. Only the edges of a single part (right ankle) are presented to avoid visual clutter.

(a) Top-down



(b) Bottom-up

Source: Elaborated by the author.

subgraph multicut and is usually solved through Interger-Linear Programming (ILP). How many edges or its value policy may vary, but the objective is to reduce this densely connected graph of possible pose configurations to a subset containing only kinematically feasible and correctly estimated poses for every individual (PAPANDREOU et al., 2017; IQBAL; MILAN; GALL, 2016; NEWELL; DENG, 2016; VARADARAJAN; DATTA; TICKOO, 2017).

This pipeline raise two challenges: how to measure confidence in these pairwise relations and how to relax this densely connected graph to a smaller sparse representation. Detecting body parts in a multi-person scenario outputs score maps with more than one peak, since there are multiple instances of the same part. By densely connecting parts across the score maps, most limb configurations are actually false candidates. To discern between true and false pose candidates it is necessary to measure confidence. The simplest way would be to calculate midpoints between the two connected parts, but approaches that do not encode orientation fall short on crowded scenes due to multiple part estimates being in a feasible distance from the first part location (CAO et al., 2016). Assuming an accurate confidence measurement, non-maximum suppression (NMS) is the standard method to filter these pose candidates NEUBECK; VAN GOOL (2006). Relaxation is also necessary due to the matching of an unknown amount of body parts into an unknown amount of individual poses, configuring a K-dimensional matching problem,

known to be NP-hard (INSAFUTDINOV et al., 2016). Strategies to reduce this graph density are broadly studied in the pose estimation literature, but most have similar run-time/accuracy trade-off (VARADARAJAN; DATTA; TICKOO, 2017).

Tracking poses throughout video frames is a fairly new problem, mainly the multi-person variation. Most works, prior to the rise of convolutional networks as state-of-the-art body part detection, frames this problem as an error minimization (ANDRILUKA; ROTH; SCHIELE, 2008; CHERIAN et al., 2014; TOKOLA; CHOI; SAVARESE, 2013). It was generally solved using the pictorial structures framework (FELZENSZWALB; HUTTENLOCHER, 2005) augmented with a third potential, besides the unary (per-part) and pairwise (pairs of body parts), to ensure temporal consistency. These approaches relied on hand-crafted features, which must be defined before the training step. Hence, these models are weak at generalizing to unexpected poses or ambiguous limb configurations due to person-to-person occlusion. Moreover, mapping individual states for each articulation independently is only feasible for a small set of object classes (e.g. walking pedestrians). It is necessary to reason over the pairwise relationship between joints and the global arrangement of individuals in the scene. Similar to many vision tasks, the progress on human pose tracking was significantly accelerated by deep learning. State-of-the-art approaches rely either on optical flow (XIAO; WU; WEI, 2018) or temporal edges (IQBAL; MILAN; GALL, 2016; INSAFUTDINOV et al., 2017). Nonetheless, to achieve real-time performance other researches adopt simpler, yet discriminative enough, temporal features tailored specifically for human motion tracking (XIU et al., 2018).

3 RELATED WORK

The multi-person pose estimation research witnessed many breakthroughs during last decade, but most authors are still doing ablation studies to understand which techniques impact the final estimates the most. It is not clear which set of features is more relevant to body part detectors, nor the loss function that better regress to this ideal set. In turn, temporal consistency is a safe bet to deal with long-term occlusion, since it is hard to estimate poses given only a small set of partially visible joints without any historical data. Thus, we reviewed works that aim at estimating and tracking poses of multiple individuals in unconstrained videos. By unconstrained, we refer to works that do not impose constraints on the video subject, camera angle or number of persons. Additionally, we do not review single-person pose estimation in videos because there is no person-to-person occlusion and the space of feasible pose sequences, which is the main focus of our model, is way smaller. The exception being the work of Song et al. (2017), that explore temporal consistency in single-person scenes through a set of features closely related to other studied method.

3.1 Research Methodology

This study aims at augmenting HPE models by enabling them to better scale towards receiving longer sequence of frames as input. Yet, our interest is the discovery of subjects body poses in images, not the interpretation or semantic meaning of such poses. Therefore, gesture and activity recognition, both fields closely related to HPE, are not reviewed. Marker-based HPE is not included either, since it is well-solved problem and its intrusiveness makes it less appealing (HOLTE et al., 2012; SIGAL; BLACK, 2010). In order to allow better quality assessment and reproducibility, the principles described in the next subsections guided this review. The corpus of articles and the scope of this study are reduced through three filters: selection criteria, quality metrics and the keywords that compose our search phrases.

3.1.1 Search Strategy

A set of keywords, listed on Table 1, is defined in order to explore the human pose estimation literature as widely as possible, within the scope of this study. The keywords were divided into two categories by their role in the search phrase: context or objective. Context mimics this study's scope and refers to human body pose estimation and methods. Objective is a set of words commonly used to refer to the identification and highlighting of human body joints. Furthermore, it also shows the permutations used within the set of keywords to limit the search space. Resulting searching phrases are used as filters on the following database libraries: IEEE

Xplore Library[1], ACM Digital Library[2], ScienceDirect[3], Springer Link[4] and Google Scholar [5]. These contain the five conferences identified as the main sources of breakthroughs and with the higher impact factor among the explored ones: ACM SIGGRAPH; European Conference on Computer Vision (ECCV); IEEE Computer Vision and Pattern Recognition (CVPR); IEEE International Conference on Computer Vision (ICCV); and IEEE International Conference on Image Processing (ICIP). Conferences are filtered by relaxing the criteria in the following ways: since most articles in these conferences target practitioners, methods applied to specific domains are considered, and preference is given to newer or well known (over 100 citations) studies. Methods ranked on PoseTrack that publicly reveal their methods are fully reviewed as well.

Table 1: Set of keywords used to compose search phrases.

Goal	**Keywords**
Context	Human, Body, Multi-Person, Part-based, Exemplar-Based
Objective	Estimation, Recognition, Recovery, Model, Regression
Permutations	*Context* + "Pose" + *Objective* *Context* + "Pose" + "Tracking" *Context* + "Estimation" "Articulated" + *Context* + *Objective*

Source: Elaborated by the author.

3.1.2 Corpus Selection

Towards reviewing only studies that are inline with the objectives of this research, two steps are taken: first, *filtering* by the selection criteria; second, a *quality assessment* step based on our quality metrics. Next, both steps are detailed and justified. For the *filtering* step, the selection criteria is defined by seven constraints:

(i) *Publication date*: in order to reflect the most recent findings, the scope is limited to the last decade of HPE research (2008-2018);

(ii) *Journal or conference*: taking advantage of the better reviewing process, only work published in journals and conferences are considered;

(iii) *Markerless method*: only methods that estimate human pose directly from the pixel space without relying on intrusive markers are considered;

[1]https://ieeexplore.ieee.org
[2]https://dl.acm.org
[3]https://www.sciencedirect.com
[4]https://link.springer.com
[5]https://scholar.google.com

(iv) *Addresses multi-person tracking scenarios*: based on title and abstract review, some articles that misleadingly use the keywords, address a method tailor-made for specific domains or simply apply an off-the-shelf body part detector on top of a standard tracking method, without any adaptation for multi-person scenarios, are discarded;

(v) *Non duplicate*: besides the direct duplicates found during the process of article selection, we also discard the work that differs too little from its approach original idea or state-of-the-art counterpart;

(vi) *Color images*: initial estimation must be done from video frames captured by RGB capable devices;

(vii) *English*: only English written articles are considered.

Articles that pass through the *filtering* step go to the *quality assessment* step, which comprises full text review and paper removal according to four quality metrics: contains reproducible research results; position itself through related work analysis and state-of-the-art comparison; clearly describes their experiments context, hypotheses and assumptions; apply the proposed method to public available video benchmark. The final corpus of articles is considered related work and grouped by approach similarity in the following sections. During the analysis of research opportunities or gaps, we considered only works that achieved a remarkable result in the PoseTrack benchmark, to enable a fair comparison to the state-of-the-art.

3.2 Literature Review

Hand-crafting discriminative features for estimating and tracking poses is a hard task (INSAFUTDINOV et al., 2017). Hence, most works that tried to solve multi-person pose tracking before the rise of convolutional networks as the standard body part detectors, achieved underwhelming results. Yet, some of these older approaches had cleaver ideas to deal with multi-person scenarios. Tokola, Choi e Savarese (2013) reason over temporal path hypothesis for each body part, instead of connecting estimations made on a single frame across time. Thus, they can reward trajectories that are consistent throughout the whole video. Their experiments show that, given correctly estimated path hypothesis, the proposed method approximate the ground truth poses very well with about 100 path candidates. Parts that have a higher frequency of movement, such as the wrists, perceived greater improvements. Even though for single-person scenarios this strategy is easily applicable by just choosing the path with the highest confidence, on multi-person scenarios these paths must be evaluated jointly, in a person-wise manner, considerably increasing the computational complexity. Following a different idea, Cherian et al. (2014) try to leverage temporal consistency to improve the estimation of hard to detect body parts (e.g. wrists, elbows) in each frame. They cast the problem as minimizing the set of edges in a graph that densely connects all detected body parts spatially and temporally between

adjacent frames. This spatio-temporal graph formulation achieved moderate success when combined with CNNs (IQBAL; MILAN; GALL, 2016; INSAFUTDINOV et al., 2017). Although in its original form the pipeline was constrained to simple human action sequences, small tweaks to this pipeline originated most of the current state-of-the-art methods.

The next sections are organized as follows: Section 3.3 describes the approaches proposed by the benchmark's creators as reference results; Section 3.4 brings the current best placed methods on the leaderboards that track poses after processing the whole video; Section 3.5 shows the current state-of-the-art trackers for real-time processing; and Section 3.6 briefly explains the work of Cao et al. (2016), a state-of-the-art real-time multi-person pose estimator, that is applied by our PastLens model during the body part detection and poses assignment steps.

3.3 Baseline proposals

PoseTrack: As a first successful work that targets multi-person pose tracking in videos there is the work of Iqbal, Milan e Gall (2016), which also provides the PoseTrack dataset. The idea is to augment the regular bottom-up pipeline with temporal consistency. After the body part detection step, they build a dense spatio-temporal graph. Two types of edges are proposed for this graph: spatial edges fully connecting all parts in a frame and temporal edges, connecting body parts of the same type across two or more sequential frames. Costs for these temporal edges are the probabilities of a part belonging to the same person across two frames. These probabilities are calculated through a logistic regression method that receive as input dense correspondences extracted with the DeepMatching tool (REVAUD et al., 2015). By optimizing their constrained (ILP) formulation, they jointly associate body parts to individuals in a single frame and the same individuals across frames. Their formulation tries to target occlusion, truncation and temporal assignment simultaneously. Gurobi (OPTIMIZATION, 2014) is the chosen solver, combined with an additional step to remove any trajectory shorter than 7 frames. This strategy is based on the assumption that trajectories that are too short have a higher chance of being false candidates, since humans tend to move slower and stay in the scene for longer. A later work improves their model and better performs on the proposed dataset (ANDRILUKA et al., 2017).

ArtTrack: By Insafutdinov et al. (2017) this model is similar to PoseTrack and also adopt a bottom-up pipeline, but it offloads a large share of the reasoning about the body part assignment to a CNN. The main difference is in the body part detector itself, that is conditioned to the human subjects locations, assigning parts simultaneously to the detection step. More precisely, they associate the head to the individuals‘ locations, since it is the easier to detect joint, and use this pre-localization as input to the body part detector. Another improvement to the detector is the intermediate supervision layers. Each body part is estimated at different layers of the CNN, following a kinematic proximity order starting from the head. This kinematic order implicitly allow the detection process to be guided by easy to detect parts, such as the head and neck. The authors state that this approach improved the detection of parts that are

harder to detect (e.g. ankles, knees, wrists). The model is suitable for scenes with an unknown number of people and applies repulsive edges in addition to the temporal ones. Costs for these repulsive edges are inversely proportional to the spatial distance. These repulsive edges have the function to remove connections between body parts that carry little to no information about other parts' behaviour overtime. For instance, the position of left ankle with respect to the right wrist. Reducing the graph to a sparser representation also reduced the runtime considerably. For the temporal edges they use a combination of three features: Euclidian distance between part candidates, SIFT (LOWE, 2004) features for orientation and features extracted with by DeepMatching. Although these features are complementary to each other, they are calculated for adjacent frames only. Moreover, their ablation study shows that, individually, these features only represent small improvements on the MPII benchmark.

3.4 Offline Methods

DetectAndTrack: Girdhar et al. (2017) use a 3D variation of the convolutional network architecture proposed by He et al. (2017), with time as the third dimension, to define temporal relationships across adjacent frames. Body part detection is done frame-by-frame and temporal consistency is calculated through a sliding window (3 frames at a time). They propose a top-down strategy, in which every detected person bounding box in the current frame is densely connected to every other bounding box in the previous frame. Afterwards, this temporal graph is reduced to a bipartite matching formulation, having the bounding box overlap (IoU) between the ground truth bounding box and the estimated one as edge costs. Hungarian algorithm is applied like in Cao et al. (2016) to solve this matching problem. Further, the information captured in small time intervals is mixed into a light-weighted long-term tracker. Their formulation is nearly two times faster than the ILP formulations described earlier and grows linearly with the number of individuals, although it is not fast enough for real-time tracking, because it must process entire sequences of frames by sliding their fix sized time window.

FlowTrack: Improving on their previous work on DetectAndTrack, Xiao, Wu e Wei (2018) change their similarity measure from spatial distance to optical flow. They store a pool of motion features calculated between a body part and all the part candidates of the same type on adjacent frames. To track poses, they combine these local pairwise features into pose tracks and compare these tracks to the features calculated for the current frame. This comparison between past tracks and the current frame is done through greedy matching. After identifying every match and classifying the new poses through NMS, they add them to the pool. The pool is organized as a double-ended queue, which removes the oldest frame's poses when a new one is added. In addition, they also adjust bounding box detections on each frame by the temporal displacement, estimated with optical flow. The authors state that flow-based pose similarity is better than bounding box similarity in the presence of fast movements. Hence, they apply Object Keypoint Similarity (LIN et al., 2014) as the costs for their spatial-temporal graph edges,

instead of IoU. Currently this approach is the state-of-the-art tracker regarding accuracy, but the matching procedure is still offline.

3.5 Online Methods

PoseFlow: In Xiu et al. (2018), the authors propose the first online pose tracker. It is based on a top-down strategy on top of their own temporal features called *pose flows*: pairwise connections between the motion of joints in two adjacent frames. The work of Fang et al. (2017) is used as a per-frame pose estimator, but they apply a new NMS method to choose the correct pose candidates. It is tailored specifically for their temporal features, avoiding greedy matching. They also recur to a sliding time window, but do not keep past data for future comparison, nor vary the window's size. Therefore, pose flows are always calculated in sets of 3 sequential frames. In scenarios where the individuals are highly occluded, confidence on the single frame estimates is very low. However, since they have the pose flow calculated for adjacent frames, they can replace this low confidence estimates with high confidence ones from prior frames.

JointFlow: The model proposed by Doering, Iqbal e Gall (2018) was close related to our own model. It applies the convolutional network proposed by Cao et al. (2016) in a Siamese architecture (two images in parallel) to estimate poses for two consecutive frames. Score maps from both frames are used as input to another CNN, responsible for predicting Temporal Flow Fields (TFF). TFFs represent the movement of body parts with a set of unit vectors, one per pixel, which encode the direction and intensity of motion between two frames. These vector fields are used as similarity measure for connecting similar body parts across frames. As a major advantage, this measurement is not task agnostic, it is specifically tailored for pose tracking in videos. Following the common bottom-up pipeline, first every body part in the two processed frames is densely connected in a graph, then this graph is reduce to a set of more likely temporal connections through greedy matching. Accumulated TFFs of each pose are connected between frames as well, in order to enable per-pose similarity measurements. Each pose detected in the current frame is assigned to a pose track detected in a previous frame or to a new one, if there is no matching with high enough confidence. The authors claim that greedy matching is enough to achieve close to state-of-the-art results, while keeping real-time performance.

3.6 Pose Estimation Methods

Part Affinity Fields: the first multi-person pose estimator to reach real-time performance on 2D image inputs was proposed in Cao et al. (2016). They jointly predict a set of score maps for each part and a set of 2D vector fields, the part affinity fields (PAFs). Each PAF encodes the likelihood of connection between two body part candidates. Their convolutional network use iterative refinement with intermediate supervision, as described in Wei et al. (2016), for both branches (score maps and PAFs). Since the score maps may contain false part estimates,

it needs NMS to filter the less likely part locations for each map. However, there are multiple individuals in the image, so each score map will have multiple candidates for the part it encodes, resulting in many possible configurations for the limbs. Moreover, there will be many false candidates among these configurations. Each limb candidate is scored by the alignment, calculated through a line integral, of the predicted PAF with the ground truth line connecting the two parts that form the limb. They relaxed the otherwise NP-hard problem of matching this graph of parts candidates into the ground truth collection of limbs through contextual knowledge: since the expected kinematic chain is known, firstly they approach the problem as bipartite graph matching, i.e., the initial edges are all the possible connections between a pair of parts (e.g. neck and head) and the goal is to find a subset of edges where no two edges share a node; second, they find full body poses by matching a minimal number of edges to build a human skeleton. As an example, they only evaluate matches of right wrists with right elbows. The relaxed version shows results close to ILP solutions (PISHCHULIN et al., 2016; INSAFUTDINOV et al., 2016), but in a fraction of the time. The authors speculate that, since the association criteria is calculated pairwise over a big receptive field, it implicitly encode global context. Remarkably, using PAFs the authors achieved similar results to the same part detector applied with ground truth pairwise connections, implying that they measure close to perfect inter-part connections.

3.7 Research Opportunities and State-of-the-art Comparison

The current trend is to approach the pose tracking problem as a sub-graph partition problem: a dense graph connecting every part candidates between two consecutive frames is reduced to a set of most likely tracks by either ILP or the Hungarian algorithm (ANDRILUKA et al., 2017). Once the tracks are estimated, NMS is used to further reduce this set of pose candidates by the confidence of their temporal edges. This approach resemble the body part detectors based on the DeepCut method (PISHCHULIN et al., 2016). Remarkably, the work of Xiao, Wu e Wei (2018) shows that a simple Hungarian matching algorithm on top of good body part estimates is superior to the proposed PoseTrack's baselines. They also showed on their previous DetectAndTrack model that a simple metric achieves good results, because the quality of the pose estimates is more critical to the final tracking accuracy than the similarity measurement itself. However, for real-time processing these measurements must use simple, yet discriminative, features that can be calculated rapidly. Further, to deal with challenges such as long-term occlusions, these features must be light enough to be matched against historical data in real-time.

In order to further justify the PastLens model as an actual contribution to the state-of-the-art on multi-person pose estimation, we compare the related works with our model on 5 attributes: single- or multi-person methods, online or offline methods, time interval used for assuring temporal consistency across frames, runtime, and arbitrary designed temporal feature representation. The first two attributes penalize works which disregard multi-person scenarios or act in an offline manner, since both constraints are prohibitive for most applications. The third attribute,

analyzes the works by how many frames they use to impose temporal consistency. Runtime is a recurring concern for vision methods and became our forth attribute for comparison. Finally, constraining the detection of cross-correlation between frames to a pre-defined representation is limiting to both the scaling of the model and the potential accuracy ceiling. Table 2 clearly shows a research opportunity: using current methods the user has to either choose near real-time methods or the ones that deal with longer time intervals, on top of being constrained to a single temporal feature representation. On the other hand, long-term occlusion, motion blur and complex human motions are all part of real world scenarios and require more than a couple frames to be dealt with. Thus, being able to leverage this temporal information across longer time intervals, while keeping the processing time close to real, would be an impactful contribution to the literature.

Table 2: Comparison between the proposed model and related pose tracking methods.

Work	Multi-person	Online	Time interval	Runtime	Temporal Features
Tokola, Choi e Savarese (2013)	No	No	All frames	Non real-time	Path candidates
Cherian et al. (2014)	No	No	Adjacent frames	Non real-time	Temporal edges
PoseTrack (IQBAL; MILAN; GALL, 2016)	Yes	No	Adjacent frames	Non real-time	Temporal edges
ArtTrack (INSAFUTDINOV et al., 2017)	Yes	No	Adjacent frames	Non real-time	Localized Temporal edges
Girdhar et al. (2017)	Yes	No	All frames	Non real-time	Bounding boxes
FlowTrack (XIAO; WU; WEI, 2018)	Yes	No	All frames	Non real-time	Optical flow
PoseFlow (XIU et al., 2018)	Yes	Yes	Adjacent frames	Real-time	Pose Flows
JointFlow (DOERING; IQBAL; GALL, 2018)	Yes	Yes	Adjacent frames	Real-time	TFFs
PastLens	Yes	Yes	Adjacent frames	Near real-time	Learned from data

Source: Elaborated by the author.

Regarding real-time processing for tracking, the state-of-the art are the *Temporal Flow Fields* proposed by the JointFlow model. Their work was based on other lightweight features for associating body part to full poses in still images, the Part Affinity Fields. Both approaches have some shortcomings. PAFs are very good at assigning body parts to poses in still images, but have as input image evidence of a single frame, therefore, if there is severe occlusion or any misalignment, it can lead to flickering or even full individuals not being recognized. The TFFs map dependency of joints between pairs of frames only, because it would have to correlate the whole collection of score maps from past frames, deeming processing longer time intervals in real-time an unfeasible challenge. Moreover, it overloads the network with information that has little to no value kinematically (e.g left-foot optical flow to estimate head positions). An alternative approach is to encode composed features that simultaneously represent the spatial configuration of the scene and temporal dependencies among frames. This is the path followed

by the PastLens model, avoiding a restrained temporal feature representation, followed by a shallow correlation of such features with spatial evidence encoded by the CNN's weights. Another core advantage of our model is that it can be attached to any existing architecture that outputs a score map, following the well spread Toshev e Szegedy (2014) approach.

4 THE PASTLENS MODEL

Pursuing a cost-efficient solution to the problem of adding temporal consistency to the estimation and tracking of poses in unconstrained videos, we developed the PastLens model. Our goal is to stretch the CNN's receptive field to also include previous frames, driving it to the detection of "time-wise" features (multi-frame) on top of the spatial ones (single-frame). By estimating and tracking poses, we refer to the localization and tracking overtime of head, limbs and torso, followed by the assembling of these body parts into poses that correctly encode the scene. An important constraint observed during the review process is that most applications rely on real-time processing to offer a visual feedback to the user. Hence, we also take in consideration the processing time of our whole pipeline, from inputting the video frame to the final estimation step. With respect to the main challenges faced on multi-person scenarios, being able to associate body joints throughout longer time intervals would allow the pose estimators to better recover from wrong estimates due to long-term occlusion or motion blur. We try to allow longer intervals by decreasing the impact of adding more frames to the CNN's input. The idea is to encapsulate the temporal features in the pose estimator weights alongside the spatial ones, instead of designing a separate model. Further, PastLens do not constraint these features to any formal representation, letting the network learn the best way to encode temporal consistency alongside the training process. This Chapter is organized as follows: Section 4.1 lists and justifies our project decisions; Section 4.2 explain in detail the PastLens architecture, highlighting its scientific contributions; Section 4.3 details our feature merging strategy and Section 4.4 follows up about the iterative refinement step.

4.1 Design Decisions

The PastLens model tries to achieve both near real-time pose estimation and implicitly deal with temporal dependencies on unconstrained and markerless multi-person video sequences. Solving both problems at once is a complex task and would require big research teams and a high-end infrastructure. Therefore, we impose a few constraints, aiming at keeping the scope of this project manageable. We covered only main body pose estimation, meaning that approaches for estimating foot, hand or face details are not investigated. With respect to the available cameras, we assume a single monocular RGB capable device without depth information. Not relying on multiple views to deal with occlusion goes against many state-of-the-art solutions, but enables a larger spectrum of applications to benefit from our model. Multi-view approaches also require the individuals to be associated across all views, which may lead to even more false pose candidates per-frame (BELAGIANNIS et al., 2016b). In contrast, we do not make assumptions about the classes of action or interactions between the human subjects.

Being able to evaluate the accuracy of our outputs in a reproducible way is another desired outcome. We use a public available and widely adopted benchmark called PoseTrack to evaluate

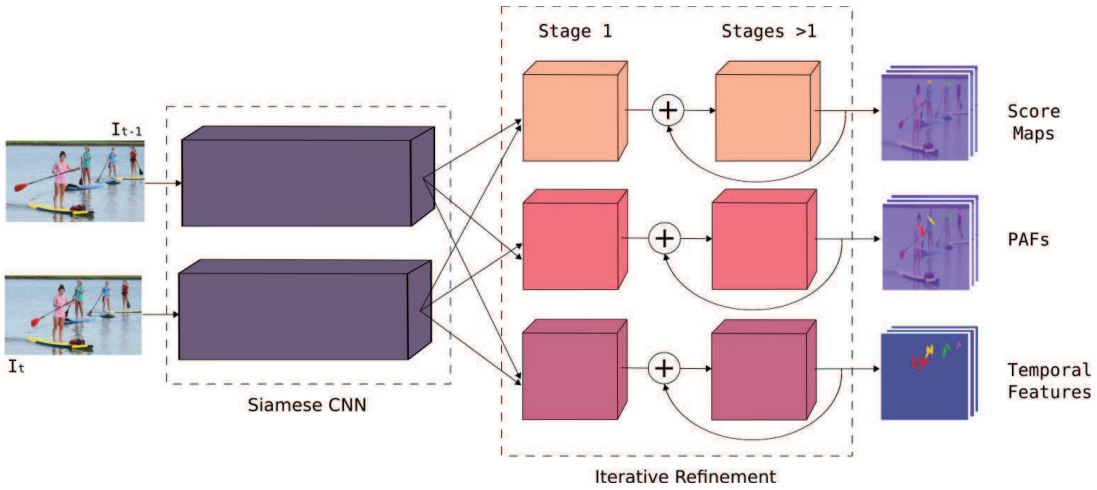
our model. It provides a collection of annotated video sequences and a leaderboard to compare our results with the state-of-the-art approaches. Each individual in one of these sequences has spatial annotations for 17 joints: headtop, headbottom (neck), nose, ears, shoulders, elbows, wrists, hips, knees and ankles. We constrain our model to the same set of joints in order to enable one to one comparisons, while assuming direct correspondence between headbottom and neck. Moreover, this benchmark may give a different track ID when the same individual re-enters the scene after leaving the camera field-of-view. Relaxing this constraint would require person re-identification, a challenging problem that is broadly studied in the literature (ZHENG; YANG; HAUPTMANN, 2016). Solving this limitation is out of the scope of this research project. Another decision based on reproducibility is the chosen architecture. Since one of the objectives is to keep our model agnostic to the chosen CNN, we adopted the Mobilenet v1 (HOWARD et al., 2017) due to its wide adoption by current literature and ease-of-use.

Varying the time interval between frames during training is mandatory to deal with complex human motions, but it may reduce confidence on fast movements that occur in a smaller set of frames. Nonetheless, our objective is to remove the need for explicit temporal features extraction, reducing the impact on the model runtime as more frames are concatenated to the input. Also, in the dataset used during training, neither the time interval between frames nor the frame rate were linear. For example, a fast movement may be captured by two annotated frames spaced by 10 in between frames, while a slower movement is annotated every third frame. Thus, we rely solely on the variation present in the data provided by our chosen benchmark to cover this gap between fast and slow movements. Another relevant project decision regards the way we are going to tackle multi-person pose estimation as a whole. Our model is developed using the bottom-up pipeline, mostly because all the related works that achieved real-time processing also adopted the same pipeline. It would be enlightening to execute tests on a top-down variation of our model, since they outperform bottom-up methods for single-frame pose estimation with respect to accuracy, mainly due to double counting (XIU et al., 2018). Yet, top-down approaches heavily rely on the person detection step, which would have too big of an impact on the pose estimation results, turning the already hard task of evaluating how our model encode temporal consistency into an unfeasible chore.

4.2 Architecture

A video sequence is a set of n sequential frames $V = \{I_1, I_2, ..., I_n\}$, containing a varying number of individuals, or none, per frame. In addition, each person depicted in one of these frames is represented by a set of body parts $P^i = \{p^i_{headtop}, p^i_{headbottom}, ..., p^i_{left-ankle}\}$ containing the 17 types of parts annotated in the PoseTrack dataset, where $p^i_{headtop} = (x_{headtop}, y_{headtop})$ are the 2D coordinates estimated for the head of the i^{th} person in frame I_n, which contains person P^i. At every frame I_n our goal is to correctly estimate the pose of every person in the depicted scene and keep it consistent with past frames. In the following sections, we refer to the

Figure 5: An overview of the architecture behind the state-of-the-art bottom-up multi-person pose estimation models (CAO et al., 2016; DOERING; IQBAL; GALL, 2018). As a first step, two sequential frames I_t and I_{t-1} serve as input to Siamese CNN. Each network composing this CNN calculates spatial features for one of the frames in the input pair separately. Later, their outputs are concatenated and forwarded to the other three branches, which compose the iterative refinement step. All three apply the same process, where each stage receives the base outputs concatenated with the last stage's, but differentiate the loss layer and thus the expected output. Note that one of such branches is tailor-made to map temporal features.

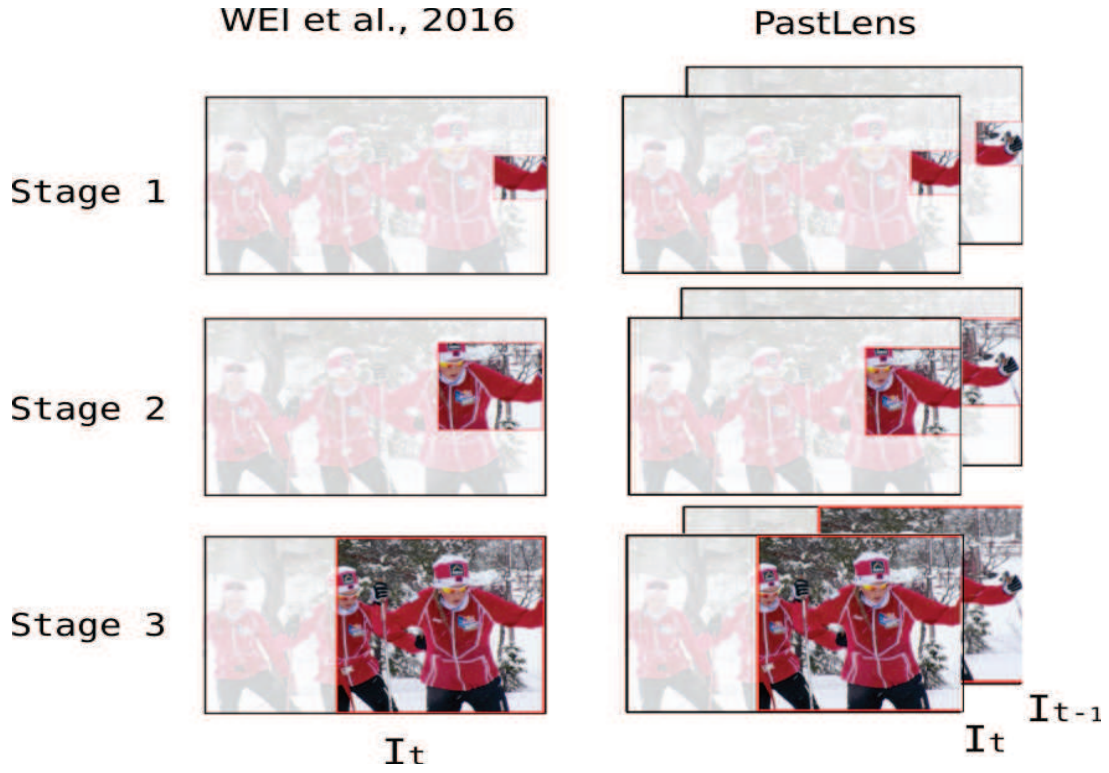


Source: Elaborated by the author.

current video frame in any given time instant t as I_t. As portrayed on Figure 5, most state-of-the-art solutions are composed of a similar pipeline: a base Siamese network (BERTINETTO et al., 2016) followed by two branches, a detector of spatial features, which outputs a score map, and another detector for assignment features, PAFs for instance. Notice that there is also a third branch, which detects temporal features to correlate frames overtime. Our research objective can be understood as removing the need for this last branch.

The base Siamese network duplicates the initial layers in order to separately extract features from each frame. Later, it concatenates the set of features from each frame as an input to the refinement stages. After the Siamese network, the second step is based on the CNN architecture used by Wei et al. (2016), which was later widely adopted by the HPE literature. It improves the work of Ramakrishna et al. (2014) by replacing random forests with convolutional networks, enabling the model to learn both image and context features directly from data. They define their version of *Convolutional Pose Machines*, as sequences of smaller CNNs (stages), where each network creates a set of score maps to estimate the image patches that most likely contains body parts. Score maps from current stage are used as input to the next stage, in an iterative refinement process. At every further stage the receptive field gets bigger, and more contextual information is available. Through this additional contextual information, estimation accuracy also rises. Training is done from scratch and augmented with intermediate (stage-wise) loss functions, to avoid vanishing gradients. In practice, we intend to use the improved version of Cao et al. (2016), since it also outputs the aforementioned PAFs. *Part Affinity Fields* are

Figure 6: Comparison of the effective receptive field on each stage (centered at right wrist) of our architecture vs Wei et al. (2016) explored by related work. At each consecutive stage the effective receptive field is larger, due to the downsampling applied by previous layers. Notice these changes are mimicked on the previous frame I_{t-1}. Previous frame I_{t-1} is flipped to better depict the symmetric increase on both frames.



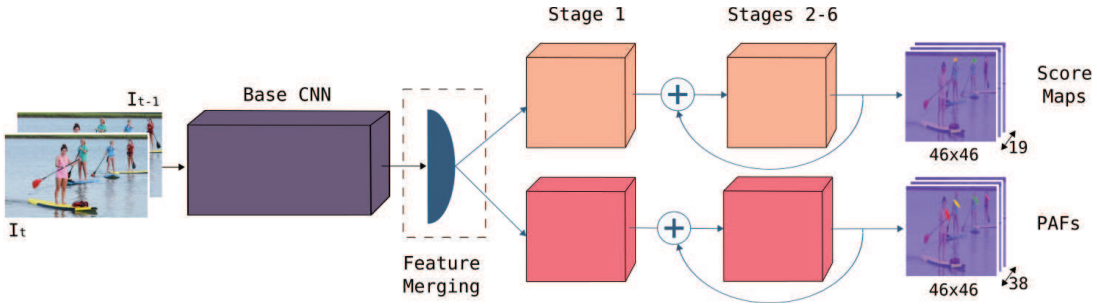
Source: Elaborated by the author.

composed of vector fields that encode part-to-part associations and make the model robust to multi-person scenarios. Their results are remarkable, not only due to achieving state-of-the-art, but also by enabling 2D multi-person pose estimation in real-time.

Each stage has access to a slightly larger receptive field due to the inputs getting condensed into smaller feature maps. In the HPE context, larger receptive fields allow the model to capture long-range spatial dependencies between body parts that are not directly connected in the proposed kinematic chain. In order to enable temporal consistency, a binding between the two frames must be provided during the training process. Instead of empirically crafting a feature capable of correlating two consecutive frames, the PastLens model increases the CNN's receptive field after each stage symmetrically on both frames, as depicted by Figure 6. This time-wise stretch combined is the result of a feature merging module, which merges temporal and spatial features into composed spatio-temporal features and is the core of our model. By merging spatial features from frames I_t and I_{t-1}, we force the network to highlight features regarding the disparity between such frames, i.e. it cannot detect features for each of the frames separately, it must prioritize features that correlates both frames in order to minimize the loss.

Providing a general overview, Figure 7 portrays the PastLens architecture. Unlike the standard pipeline described earlier, we do not rely on any explicit feature encoding, besides the

Figure 7: An overview of our proposed PastLens model. The base CNN and the refinement process follow the same premises of other works in the literature. The main difference is the absence of an additional CNN to add temporal consistency to the output, thus reducing the impact of receiving longer time intervals as input.

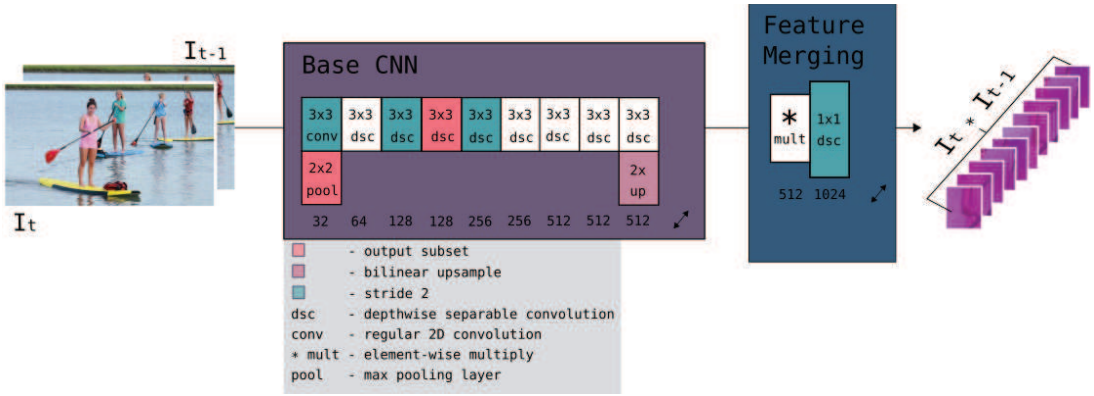


Source: Elaborated by the author.

spatial ones (score maps and PAFs). Moreover, there is neither a branch nor pre-defined constraints regarding the temporal features. The sets of spatial and temporal features are combined by multiplying feature maps element-wise across all channels, after the non-linearity step. We must use element-wise multiplication because we care about the spatial configuration, which would be lost if we picked the dot product. Experimental results on the action recognition task shows that multiplicative combinations perform better than additive on spatio-temporal contexts (PARK et al., 2016). Although this procedure effectively double the training operations (assuming two frames as input), it keep the exact same number of operations during test time, contrasting with Siamese networks that must keep the enlarged structure on both scenarios.

This way, the set of filters after merging contains only $N \times 3 \times 3$ weights, instead of $2 \times N \times 3 \times 3$ as of a regular CNN's layer. Since the base network will extract spatial features from the input, including both frames resulted in spatio-temporal features that highlight interdependence, as shown by our experimental scenarios described on Chapter 5.1. Spending part of the available weights to extract correlation features may harm the set of spatial-only features that our model is capable of extracting. This tradeoff between improving temporal consistency and losing relevant spatial features is further discussed in Chapter 5. We chose to impose this correlation between consecutive frames in the output of our base CNN, differently from other models that extract spatial features in two separate base networks and only impose any sort of temporal consistency on top of downscaled score maps. Hence, we force the model to condense more features on the same set of filters, but allow our refinement networks to detect body parts and assign these parts to complete poses based on a more constrained set of inputs. By constrained we refer to the way that adding a previous frame limits the set of feasible pose configurations for the current frame. Also, since we rely on spatial features calculated before any task specific branch, the model as a whole is not affected by pose estimates in past frames, only by inaccurate feature detection, which is way less impactful due to the broader collection of feature maps in these mid layers (final estimates have 17 maps, while the layers go up to 512)

Figure 8: PastLens feature merging module and base CNN architecture. The feature merging step receives both frames I_t and I_{t-1} as input and pre-proccess (base CNN) their initial spatial features. Later, an element-wise multiplication between the two resulting sets of spatial features merges them into a spatio-temporal set. These mid tier outputs are concatenated into a final set $I_t * I_{t-1}$ that will be iterativly refined.



Source: Elaborated by the author.

4.3 Feature Merging and Base CNN

HPE research had many breakthroughs regarding pose estimation in still images, yet, the current literature still suffers with flicker and temporal inconsistency in videos. The lower accuracy seen in video applications compared to images is most likely due to the difficulty in designing CNNs capable of learning temporal and spatial features simultaneously. Averaging outputs from two frames in separate may lead to small improvements in accuracy, but the temporal network would be prone to confuse two similar motions and the spatial network two similar body parts. This behaviour was already observed on the action recognition context and would be engraved by the similarity between poses and human subjects. Therefore, combining this information earlier to force the model to consider both temporal and spatial feature spaces may lead to a better outcome.

Park et al. (2016) combined both sorts of features through multiplication and slightly improved accuracy. Their work concerned action recognition and had a predefined set of possible human actions to be mapped. We focus our attention on unconstrained videos and consequently increase the scope. In turn, the space of possible pose configurations is lower than the full set of possible human actions and similarity between poses is higher than between actions. Combining features by multiplication will produce higher activations where the highlighted feature is present on both frames, thus, enforcing a spatio-temporal dependency. For instance, our model must use the displacement between an elbow in frame I_t and some wrist occupying the same image region in I_{t-1}, instead of only relying on visual cues from the current frame. This feature merging module is a CNN itself, detailed in Figure 7.

This module is trivial to implement in the proposed architecture and adds $C \times S_{out\_size} \times T_{out\_size}$ operations, where C is the number of channels from the two output layers that will be merged, and S_{out_size} is the dimension of the feature maps from the base CNN's output with respect to the current frame. Consequently, T_{out_size} regards the base CNN's output , but receiving the previous frame I_{t-1} as input. Following Park et al. (2016) we also added an additional $1 \times 1 \times 1024$ layer, which works as a linear increase in the number of channels, without reducing the input dimensions. During training by backpropagation, calculating partial derivatives and error gradients is also straightforward.

We merged the features at the mid layers of our model, forcing it to encode fine spatio-temporal features on every further layer. Another benefit is that the merge step works as an additional regularization measure, since it makes most layers unable to reach the initial spatial features alone, which could be too specific to the set of images used during training, consequently avoiding overfitting (PARK et al., 2016). As aforementioned, the feature merging module happens after the layers composing our base CNN. It goes against our claim that we want the CNN to rely on spatio-temporal features from beginning to end, yet, initial layers on current convolutional architectures detect features that are too common to impose correlations towards multiple frames, such as edges and color gradients. Further, it imposes no restriction to the choice of base CNN, enabling the PastLens model to be indirectly improved by future research breakthroughs.

Regarding the base CNN, we implemented the Mobilenet portrayed on Figure 8. It is composed by nine layers applying depthwise separable convolutions with 3×3 filters, followed by regular ReLU non-linearity functions. The first layer applies a regular convolution with stride 2 in order to downsample the input, while also downsampling by $2\times$ max pooling (SIMONYAN; ZISSERMAN, 2014). Outputs from the stride increase follow the regular flow of the network, while the ones from pooling are kept to be used as part of the base CNN output. The fourth layer also saves a copy of its output as part of the final set. There is a bilinear upsample process on the last layer due to the output dimensions differing from the expected input dimensions on the refinement stages. Concatenating the output of different layers allow us to cover a wider range of feature map sizes, enabling further generalization.

Due to the base CNN representing most operations of our architecture, training it from scratch would take too long and become a bottleneck in itself, thus, we adopted a model pre-trained [1] on the MS-COCO dataset. This pre-training step targeted image classification, not HPE, but most initial layers end up learning simple features as color gradients or edges (SIMONYAN; ZISSERMAN, 2014). In the original work of Howard et al. (2017) the CNN has two more layers with 1024 channels each, but they would further reduce the size of the refinement step input. We empirically chose to discard them. Instead of adopting the original architecture proposed, increasing the number of channels would allow the network to learn more features and avoid having to discard features that are less frequent in the training dataset due to the limit

[1] Available at *https://github.com/ildoonet/tf-pose-estimation* on 01/2019

of filters per-layer. Nonetheless, modifying the CNN even further would be detrimental to one of the PastLens objectives: leveraging pretrained and publicly available models.

4.4 Refinement Stages

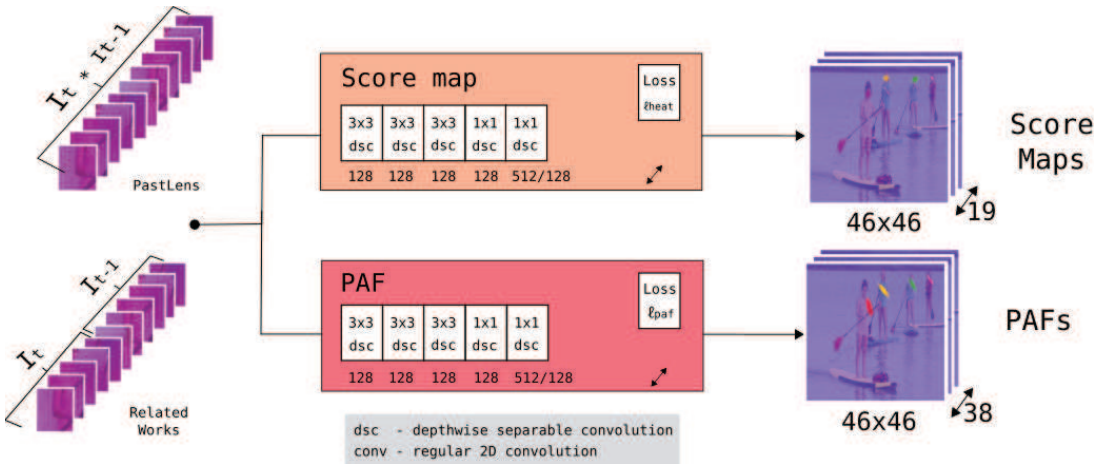
With respect to the iterative refinement, we chose a bottom-up architecture that separate the pose estimation task into two steps: the body part detection and the body part assignment. During experiments we noticed that enforcing temporal consistency on top of the PAFs, used for assigning each detected body part to a coherent pose, would require a 3D encoding. Although it is a valid option, works such as Girdhar et al. (2017) and Doering, Iqbal e Gall (2018) already explore this research opportunity with state-of-the-art results, for both online and offline pose estimation and tracking. Further, all reviewed works that relied on these temporal vector fields encoding, had to design new feature representations and complex Siamese CNNs.

As observed on Chapter 2, although each studied approach relies on considerably different encoding and CNN architectures, their improvements over baselines works, such as Iqbal, Milan e Gall (2016) or Insafutdinov et al. (2017), are similar. Through ablation studies, they have shown that their models improve overall accuracy, but we believe that the temporal consistency itself is responsible for a considerable part of such improvements. In other words, there is more space for advancing on the encoding of temporal features, than in adapting the CNN structure to cope with the notion of a sequential process.

Figure 9 differ our model from other reviewed works regarding the iterative refinement process. We keep the composition similar to our base CNN, applying 3×3 depthwise convolutions in a series of three layers, followed by two 1×1 layers to adjust the number of channels in the output as needed. Dimensions for both score maps and PAFs are the result of the chosen sequence of layers and module compositions and represent no particular design choice. Nonetheless, the number of channels on both outputs are dependent on the chosen dataset for training and the way ground truth PAFs are calculated. Combining both sets of joints from PoseTrack and MS-COCO, used by the pre-trained model, we end up with a set of 19 joints (original 17 and both eyes). Therefore the PAFs' set must account for 19 affinity fields on each direction, vertical or horizontal, composing a set of 38 maps.

Every stage has its own loss layer, ℓ_{paf} or ℓ_{heat} depending on the branch, to provide intermediate supervision. Thus, the stages are penalized separately by the backpropagation algorithm and optimize their filters accordingly. There are two main reasons to keep this behaviour: each stage correlate features based on a different receptive field and to avoid vanishing gradients. Regarding the receptive fields, in practical terms, it means that each of the stages will learn to highlight body part dependencies of different spatial ranges. For example, the first stage may have higher activations when there is small scale poses in the scene or correlations between body parts that are directly connected, such as wrists and elbows, while the later stages will concern long-range correlations (head to knees), up close poses or person-to-person occlusion.

Figure 9: Iterative refinement module design choices. Related works use as input the concatenation of two sets of features, one composed of spatial features highlighted on frame I_t and another for frame I_{t-1}. In contrast, on the PastLens architecture, each consecutive stage receives previous stage's output and the spatio-temporal features set $I_t * I_{t-1}$. Notice that the last layer of each branch has 512 channels on stage 1 and 128 on the other ones. This is due to the output of the first layer not being encoded as PAFs or score maps.



Source: Elaborated by the author.

Spatial estimates (score maps) and the base CNN, are adjusted by a per-stage squared $L2$ loss $\ell_{heat,\sigma}$:

$$\ell_{heat,\sigma} = \sum_{q \in I_t} RoI(q) \cdot \|S'_t(q) - S_{t,\sigma}(q)\|_2^2 \,, \tag{4.1}$$

where $S'_t(q)$ and $S_{t,\sigma}(q)$ are the ground truth and the predicted score map values during stage σ at pixel q in frame I_t, respectively. $RoI(q)$ is a binary mask calculated over frame I_t, with $RoI(q) = 0$ for every pixel q that is not part of an annotated body part region. Ground truth score maps are provided by the PoseTrack benchmarks through annotated keypoints (image coordinate pairs). Final score maps are calculated by applying a max operator over the concatenation of every stage's score maps. Max is better than average to aggregate score maps, since it avoids scenarios in which wrong estimates dilutes strong Gaussian peaks (CAO et al., 2016). We penalize the body part assignment estimates (PAFs), as well as the base CNN, through per-stage squared $L2$ loss $\ell_{paf,\sigma}$:

$$\ell_{paf,\sigma} = \sum_{q \in I_t} RoI(q) \cdot \|\psi'_t(q) - \psi_{t,\sigma}(q)\|_2^2 \,, \tag{4.2}$$

where $\psi'_t(q)$ and $\psi_{t,\sigma}(q)$ are the ground truth and the predicted PAF values during stage σ at pixel q in frame I_t, respectively. Ground truth's estimates are calculated following the same protocol described in Cao et al. (2016). In summary, for a pixel q that belong to any of the annotated joints p_t^i, its ground truth PAF is a unit vector that points from p_t^i to the correspondent p_t^j joint instance. By correspondent we refer to the parent joint in the kinematic tree. Clarify-

ing, the $p_{left-shoulder}$ is a parent node of the $p_{left-elbow}$, which is in turn a parent node for the $p_{left-wrist}$. At test time, PAF confidence is measured by the alignment of the predicted PAF and the limb that would be formed by associating parts p_t^i and p_t^j. By greedy matching, we reduce the dense graph of possible pose configurations to the most likely ones, having as edge values these confidence measures. To select between multiple candidates for the same pair of body parts, we use NMS to select the one with highest confidence for each of the body parts detected in the current frame. There is also the need for a lower confidence threshold to avoid false body part candidates, arbitrary set to 0.15 after a trial and error process.

There are six stages in our refinement set, also following the results presented by Cao et al. (2016). Adding any stages after the third one have diminishing returns regarding spatial accuracy of the body part estimates, but may be needed to encode coarse grained correlations between two adjacent frames. This hypothesis come from the way our receptive field expands linearly on both frames: only the latest stages would be able to cope with long-range spatial dependencies across two frames. The main tweak during training is that both loss functions, across all stages, are implicitly conditioned to relying on spatio-temporal features that satisfy pose configurations through both frames. Since the base CNN is connected to both branches by the feature merging module, their separate error gradients will jointly impact it. This way we avoid the case of our model end up approximating a single-frame pose estimator.

5 EXPERIMENTAL RESULTS

This chapter define our proposed experiments and validate the implementation of our PastLens model. It is composed of two distinct sections. First, Section 5.1 explains the experimental scenarios we developed to measure how impactful each project decisions was and summarize the evaluation metrics. In addition, it also share implementation details, training parameters and characteristics of the chosen benchmark, while addressing a few of its limitations. Later, in Section 5.2 we analyze the obtained results and compare it to the state-of-the-art methods.

5.1 Evaluation Methodology

Changing the context to the empirical evaluation of our model, the process of measuring runtime of CNNs is very sensitive to the computational infrastructure. Therefore, the comparisons are made between the base architecture and the final model, after adding the PastLens modules, discarding direct comparison to the numbers reported by related works. We deemed retraining their models to work inside the PastLens' infrastructure and elaborate an additional feature encoding for their output, not worth the effort, since even coding discrepancies would invalidate these experiments. Most well-konwn datasets are limited to single frame annotations (LIN et al., 2014; ANDRILUKA et al., 2014). In contrast, video based datasets usually annotate a single centered individual in each frame and constrain the videos to specific domains, such as walking pedestrians. To fill this gap, the PoseTrack dataset was proposed by Andriluka et al. (2017). We recur to this dataset to both train and evaluate our model. The following sections describe our computational node and test scenarios; further detail the available dataset; explain the adopted PCKh evaluation metric, on top of other studied alternatives.

5.1.1 Infrastructure and Test Scenarios

The testing node will be composed of 16 GB of DDR3 1866 MHz memory, an NVIDIA[1] GTX 960 GPU, which is a CUDA[1] 9.0 and cuDNN[1] 7.5 capable device. The CPU will be a FX 8300e 3.3 GHz, due to resource limitations. Although it is a weaker CPU, most of the pipeline will be processed in the GPU through the CUDA high performance library. Thus, the performance lost due to this consumer grade CPU should be minimal. Other system requirements may vary, but they will have little to no impact on the proposed test scenarios. The code provided by Doering, Iqbal e Gall (2018) is going to be the foundation for our own code, in order to avoid possible accuracy losses due to sub-optimal coding. As observed during our literature review process, most works rely on TensorFlow[2] as their deep learning framework using Python[3] as the preferred programming language, so will we.

[1] httpl://www.nvidia.com
[2] httpl://www.tensorflow.org
[3] httpl://www.python.org

We propose two test scenarios to evaluate our model. In the first scenario, the full PastLens model will be executed on all validation and test videos provided by the PoseTrack dataset. Final scores will be compared to the results of related works and considerations will be made towards the main insights and observed deficiencies. As a second scenario for the full model, we proposed a subset of random selected videos from the validation set that differ in the number of annotated frames, duration and motion speed. For instance, we have slow grass mowing videos in contrast with fast paced dancing ones. We ran qualitative visual tests overlaying the videos with our estimated score and PAF maps. These qualitative tests focus on the video segments that our model had the best, or worst, performances.The worst sequences typically includes a large number of overlapping people and fast camera movements, or artificial perspective transitions.

5.1.2 Training and Implementation Details

Starting with data augmentation, we chose not to rely on any sort of artificial data, such as cropped or even flipped frames. The main reason is that these augmentations would have to be done in the same manner for each video as a whole, while changing randomly at each training epoch. Otherwise, the artificial data would add little to no variation to the data, redeeming this effort as not worthwhile. Thus, we judge that the impact of an erroneous data augmentation step would be more detrimental to our tests than a correct procedure would be advantageous.

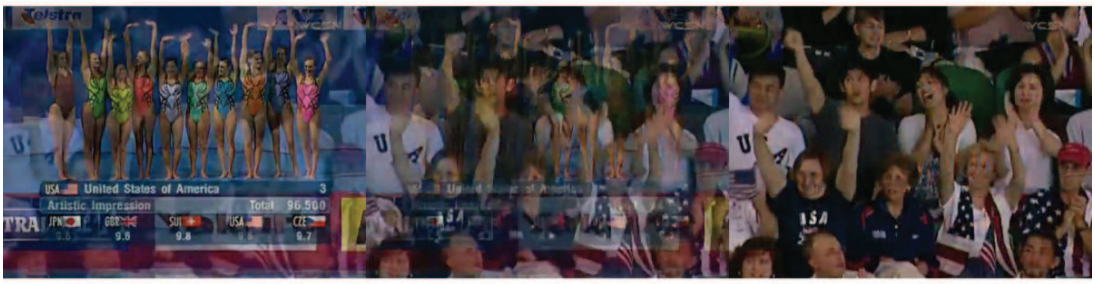
During training, regularization was applied through a dropout strategy of 80% of a random layer being "hidden" from the model during each iteration. Due to memory limitations and to avoid local minima, on top of all the regularization, we apply a stochastic gradient descent optimization (SGD) method. By SGD we refer to the method of applying the error gradients frame by frame, without any mini-batch formulation, which are the standard training method in current literature. Consequently, we do not apply any sort of batch normalization to our data. With respect to the implementation itself, we used the ADAM (KINGMA; BA, 2014) optimizer provided by the Tensorflow api, with $\epsilon = 10e - 04$. The learning rate is set to 0.0001, with a weight decay of 0.33 every 45000 steps (nearly each epoch). Following results are obtained after 5 training epochs, each epoch consisting of 43603 iterations. Combining a rather slow slope from the squared $L2$ loss with SGD consumed too much of the time available for this research project, harming the final results.

5.1.3 The PoseTrack Benchmark

PoseTrack is currently the largest dataset for multi-person pose estimation and tracking in videos. It provides video sequences with a great range of scales, number of individuals and people getting in and out of the camera field of view. This variety is further depicted in their article, thus, we do not described it any further. The authors claim that improvements on single frame estimates will inevitably improve tracking performance as well. Their data is derivative

Figure 10: Some of the challenges present in the PoseTrack data that our model is not designed for: (a) abrupt changes in viewpoint; (b) television's interface overlap and artificial transitions.

(a)



(b)

Source: Elaborated by the author.

from the MPII dataset, from which they selected a subset of frames as keyframes and added seconds of video footage around it. Each video contains between 41 and 151 frames, since the frame rate may vary. The keyframes are usually the ones with multiple articulated people engaging in diverse activities simultaneously. For bigger crowds, in which individuals are too small, such as supporters in a football match, they provide "ignore region" annotations.

The benchmark contains 540 videos totaling 66,374 frames, split into 292 videos for training, 50 for validation and another 208 for testing. On testing videos only the 30 frames from the center are annotated, while in the validation set every fourth frame is annotated, allowing long-term pose tracking. Each person has a fix ID across these frames and a bounding box surrounding its image region inferring approximate scale. As aforementioned, for each individual they annotated a set of 17 body parts. Regarding our own model, their validation set is interesting for evaluating the training step, since it gives preference to videos that include person-to-person occlusions. They also provide an evaluation server for a fair comparison to other ranked methods. It measures pose estimation accuracy in each frame separately on a hidden set o videos, while also requiring pose tracking throughout the whole video (XIAO; WU; WEI, 2018). Both scale and number of individuals per-video are completely unknown, but they enforce that they are not domain specific. As a reference for newer models, they provide baseline methods to validate the results sent to their servers and to define a lower threshold for accepting submissions. These methods were further described in Section 3.3. Most of the works submitted to their server relied on external sources to increase the appearance variability during training. These works used MS-COCO and MPII datasets for a pre-training stage, fine-tuning

their models on the PoseTrack data afterwards. It is important to notice that both pre-training datasets are composed of static images, i.e. add no additional temporal information.

There are some disadvantages of training the PastLens model on this dataset, as shown on Figure 10. Most MPII videos are captured directly from television content, including artificial transactions, animations, logos and drastic changes on viewpoint, even between two adjacent frames. Since we force the model to estimate based solely on spatio-temporal correlation, it is prone to misinterpret these unexpected scenarios and estimate false pose configurations, even if these configurations are trivial to recognize spatially.

5.1.4 Evaluation Metrics

As for evaluation measures for static images, the multi-person pose estimation literature follows slightly modified versions of the traditional HPE metrics. Due to the large space of possible scenes and poses in which human subjects may be found, it is not feasible to build a universal dataset. This led to the creation of specialized datasets (BELAGIANNIS et al., 2016a; LÓPEZ-QUINTERO et al., 2016), making fair comparisons between them a difficult task (ROGEZ et al., 2012). However, many evaluation metrics were proposed during the last few years of research. One of the former proposed metrics, which was key to many advancements on HPE methods (YANG; RAMANAN, 2013), is the *Percentage of Correct Poses* (PCP) by Ferrari, Marin-Jimenez e Zisserman (2008). PCP measures the number of body parts location estimates where both endpoints (joints) are less than 50% of the body part's ground truth length away from its ground truth location. Sapp e Taskar (2013) propose a variant of PCP that uses Euclidian distances in pixels from the ground truth in an image scaled so the torso is 100 pixels tall. Two augmented versions of PCP are proposed in Yang e Ramanan (2013): *Probability of Correct Keypoints* (PCK) and *Average Precision of Keypoint* (APK). PCK evaluates a joint estimate as correct if the distance to its ground truth is lower than the longer side of a bounding box surrounding the individual. APK is a variation of PCK that calculates the distance relative to body parts size, since there are no ground truth bounding boxes at test time.

Andriluka et al. (2014) augment the two aforementioned evaluation measures, proposing PCPm and PCKh. PCPm stands for PCP *Mean* and removes a disadvantage of PCP: its over-penalization of foreshortened body parts. The solution proposed is to use 50% of the mean length for a given body part over all training set images. Estimates over images in which this body part is severely foreshortened will be penalized equally to an image where it appears in natural scale. For PCKh the PCK's matching threshold is modified to be 50% of the head length, making it articulation independent. Pursuing a similar solution, Ramakrishna, Kanade e Sheikh (2013) propose *Keypoint Localization Error* (KLE), calculating the Euclidean distance between estimates and the ground truth, normalized by head size in order to perceive scale changes. Of note, OKS is the measure used on the MS-COCO benchmark, which plays the same role for keypoint detection as the IoU does for object detection (LIN et al., 2014). Ionescu et al. (2014)

Table 3: Comparing our results with state-of-the-art methods on validation set.

Higher mAP is better.

Work	Online	mAP
FlowTrack	No	76.7
PoseFlow	Yes	66.5
JointFlow	Yes	66.7
Mobilenet+PAFs	Yes	23.6
PastLens	Yes	43.0

Source: Elaborated by the author.

propose *Mean per Joint Localization Error* (MPJLE) reporting the distance in pixels from the joint estimated location to its true location. Such a direct approach is possible because their dataset has markers attached to the human subjects and every image is captured on the same scale. Concerning 3D pose estimation, usually, there is a radius that defines a thresholding sphere around a 3D joint in which the estimates are considered to be correct (GANAPATHI et al., 2012). We evaluate our model frame-by-frame using PCKh, through the implementation provided by the chosen benchmark.

To measure the final accuracy of our model we follow the benchmark and use *mean Average Precision* (mAP) (PISHCHULIN et al., 2016). These spatial measures do not consider occlusion during evaluation, penalizing methods that correctly estimate occluded joints or the ones that estimate them, but do not consider them as part of their final set of poses. To provide a fair comparison, the benchmark consider estimates for occluded joints as correct if they were estimated for the correct image region even though the joints are partially or fully occluded. They also consider as correct estimates if these joints are not estimated at all. This latest characteristic may be harmful to our model, since it will try to estimate these occluded joints in every frame, while other approaches may just completely ignore them to reach higher scores.

5.2 Wrap-up and Discussion

Estimating human poses is not a trivial task due to the wide variations of body shape and clothing appearance (HOLT et al., 2011; YANG; RAMANAN, 2013; DANTONE et al., 2014; RAMAKRISHNA et al., 2014). Regarding clothes, low-level image features are not useful in unconstrained or outdoor scenarios because there is no reliable way to determine the color or texture of a person's clothing without prior knowledge. In addition, the color and texture of the background clutter can be too close to the human subject, preventing background subtraction. The clothing further interferes with the body shape, since it can be crumpled or deformed by the environment or by contact with other people.

Table 4: Comparing our results with state-of-the-art methods on test set.

Higher mAP is better.

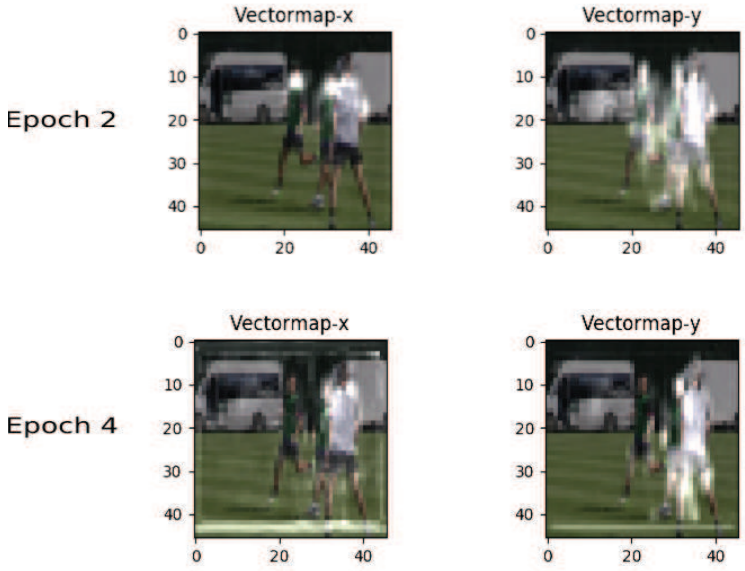
Work	Online	mAP
FlowTrack	No	74.6
PoseTrack	No	59.4
PoseFlow	Yes	63.0
JointFlow	Yes	63.3
Mobilenet+PAFs	Yes	21.6
PastLens	Yes	40.4

Source: Elaborated by the author.

Apart from clothing, appearance is severely affected by variation in body shape and visual appearance among the human population. Even regarding single-person estimation tasks, appearance may impose some challenges: arms and legs are pictured as parallel lines by many feature descriptors, but it is also a very common pattern on background clutter. Thus, when each part is considered individually, without global or multi-part contextual information, lower limbs lack defining features (PISHCHULIN et al., 2013). Wei et al. (2016) show that the detection rate of parts that have a consistent appearance, such as the head or the shoulders, is always superior to the lower limbs. This statement holds for every studied method. It is also a good justification for why most earlier methods targeted only the upper body. Rotation and positioning can also be an issue due to the overlap with other body parts, because these overlaps change visual characteristics. For instance, an elbow foreshortens when rotated, yielding different appearance cues when positioned above the head or beside the torso (YANG; RAMANAN, 2013). Orientation and viewpoint are also valid concerns, since the same pose can have multiple representations and completely different features depending on the viewpoint. As an example, the head viewed from the back does not have the most distinctive features such as eyes or mouth.

Provided that we used the PoseTrack validation set to calculate our mAP, based on the PCKh metric, we compare our accuracy to the works that also experiment over the same subset of videos. Table 3 clearly shows a gap between our results and the current state-of-the-art, by a margin of $\sim 20\%$ considering the online methods. In turn, we argue that our model is capable of better scaling, since it adds little overhead to the base CNN structure. For instance, the JointFlow method adds a whole new sequence of refinement stages on top of an additional vector field calculation between pairs of frames. Moreover, it would be required to add a forth one in case of using an additional frame as input. On the other hand, offline methods lack applicability, while keeping the worse scaling disadvantage. Another consideration is the way we collected these results: we did not have computational power to retrain the model over such a huge dataset as PoseTrack, thus, we chose the smaller version to implement (pair of frames). This choice bring on a few limitations, as the observed non-convergence on some harder subsets

Figure 11: There were hints of non-convergence on a few videos, as observed in the depicted example. At epoch 2 the model correctly detected the spatial dependency for the background individuals. Later, on epoch 4, it outputs false estimates. It may be due to the low number of free parameters (weights) available to the model during training.



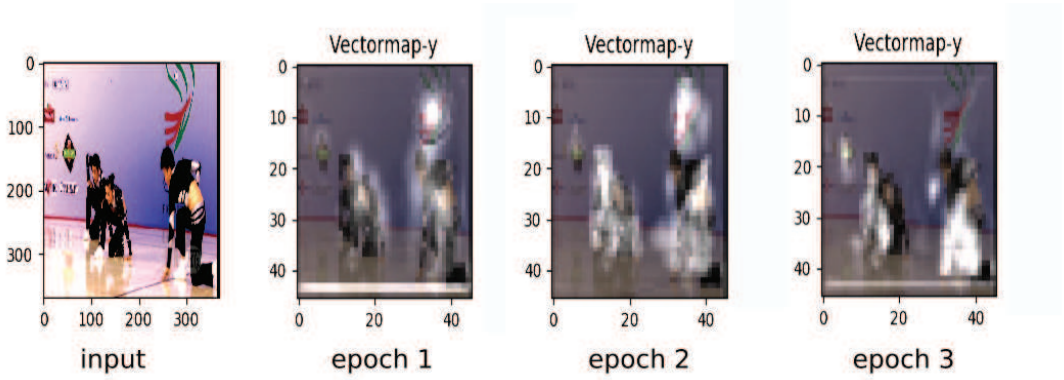
Source: Elaborated by the author, using the Tensorboard framework.

of videos (Figure 11). Hence, although our results did not reached state-of-the-art performance, all other models lack the capacity to learn spatio-temporal features direct from data, which can lead to huge improvements given enough data and computational resources.

Just looking at the obtained performance, we argue that current CNN architectures are capable of learning multi-person pose estimation through spatio-temporal features. To further validate our argument, observe that in the variant trained exclusively on spatial features, Mobilenet+PAFs, the addition of a feature merging module gave effective increases of $\sim 20\%$. Table 4 shows the results on the more challenging dataset of test videos: the difference are akin to the validation subset of videos. Notice that besides our the PastLens and Mobilenet+PAFs entrances, other results were collected by other authors, mainly Doering, Iqbal e Gall (2018).

An example of the spatio-temporal feature learning process is portrayed on Figure 12. During the early stages of training, the CNN mistakenly adds artifacts in regions that had high confidence pose candidates in the previous frame I_{t-1}. This happens because during the feature merging step multiplication, the spatial features from current frame I_t are not precise yet, therefore, almost every image region has some sort of activation potential. If the set of activations from the latest frame is already narrower towards the correct pose configuration, the final output will be a blend between the best pose candidates from each of the frames. As the training process continue, the confidence on the absence of body parts in the regions of I_t gets higher, thus, suppressing the activations from I_{t-1} that are not correlated to I_t. We use the PAFs vectormap as an example, but the same properties were observed for spatial score maps.

Figure 12: Spatio-temporal features learning process. The image is part of a video containing fast paced dancing, from which this specific frame was captured right after a prominent downward move. Notice the artifact mixing up both frames during epoch 1. As the training advanced, it learned to use these past features and ended up approximating the expected correlation, without temporal interference.

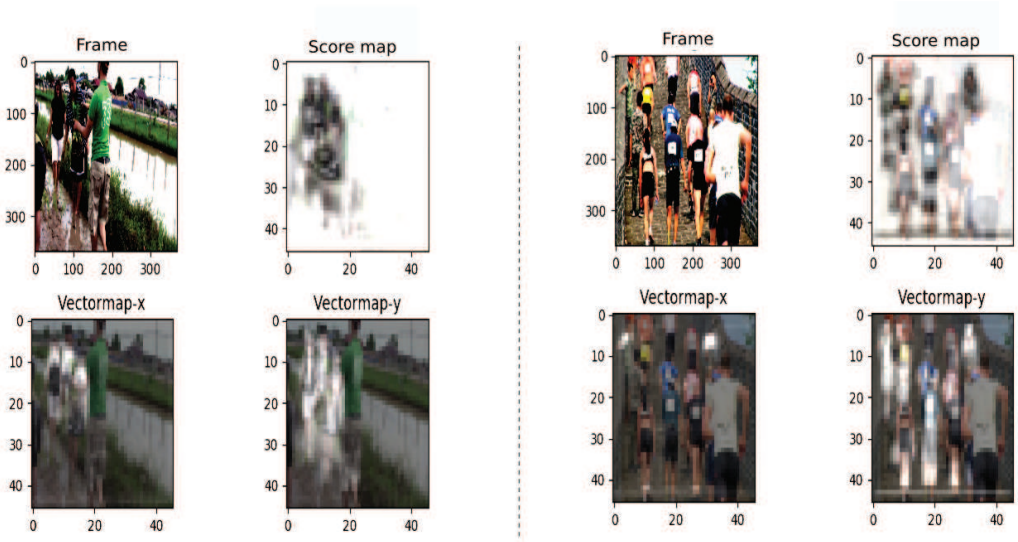


Source: Elaborated by the author, using the Tensorboard framework.

The color scheme of score maps in all the following figures is reversed to differentiate it from the body part assignment features learned. Another recurrent deficiency rise when there are a foreground person way closer to the camera than the other individuals. Figure 13 exemplify these cases, which were observed in two distinct forms: it either estimates everything as expected, missing the scaled individual completely, or it misses horizontal association between parts besides the individual itself. Even on state-of-the-art models, small variations in scale on the same individual may result in huge differences for the final pose estimates (KE et al., 2018). In addition, scale is also dependent on individual body shapes and foreshortening, thus, simply defining a global scale beforehand is not a valid alternative to make the model scale-invariant (YANG et al., 2017). Researchers train their models to be invariant to scale and number of individuals and to cope with person-to-person occlusion. However, a single model may require too much capacity to deal with all this variance (NEWELL; DENG, 2016; VARADARAJAN; DATTA; TICKOO, 2017). Since our model is rather small, this may be another negative consequence of limiting the number of free parameters in exchange for performance. We could not decisively conclude why this happened, but it became a target for future research.

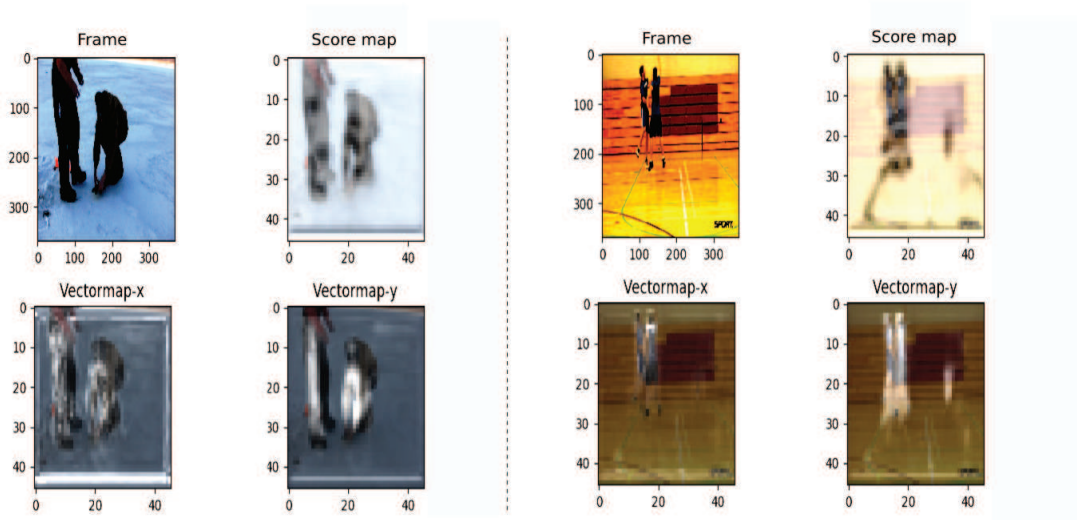
A drawback of other methods that map long sequences of frames is the susceptibility to error accumulation, i.e. a misleading pose from a former frame can wrongly penalize following correct pose candidates (SHI et al., 2017). Further, persons may be visible for short periods and then stay partially or fully occluded for a long time interval. Therefore, an ideal model would have to impose temporal consistency within varying time windows (ANDRILUKA; ROTH; SCHIELE, 2008). The PastLens model was able to capture motion at different time intervals and frequencies without relying on offline processing. Figure 14 showcase coherent pose estimation across two outlier cases: staying still for longer periods and falling from a jump motion.

Figure 13: Persistent failure case: foreground people with a bigger scale than the other individuals are missed, while background ones return accurate estimations on both body part detection score maps and assignment vector fields. Nonetheless, in some cases, as the one depicted in the right, the horizontal vector fields are inconsistent as well.



Source: Elaborated by the author, using the Tensorboard framework.

Figure 14: Success cases obtained after the last training epoch. The one in the left is a fairly static sequence with little to no movement, while the right one depicts a jump scene during a basketball match. Although the frame rate and motion patterns are completely different, the PastLens model is capable of detecting spatio-temporal features in both scenes.



Source: Elaborated by the author, using the Tensorboard framework.

In summary, our feature merging module had a meaningful impact on the Mobilenet+PAFs pose estimator. Yet, more important is the fact that there was no guidance towards the way it should approach pose sequences, it learned everything from data. This is where our cost-efficient statement comes from: while other models squeezed as much performance as possible from their CNN structures, we just modified the way our inputs (receptive fields) are encoded and already improved naive approaches (single-frame) quite significantly.

6 CONCLUSION

This work answered the proposed research question: *How would a model that enforces spatio-temporal dependencies and learns how to encode temporal features from data perform on the task of estimating poses of multiple interacting persons in videos?* It shows that the model is indeed capable of tackling such task, by improving upon approaches for static images. Furthermore, it has a wide range of parameters to be tuned and new architectures to be adapted for, which is in itself promising. As a takeaway, poses can be more accurately estimated based on the temporal dependency of body parts in sequences of frames, than through spatial features calculated frame-by-frame.

6.1 Contributions

Tracking and estimating human poses by itself is a challenging task, since humans may present a wide variety of pose configurations. Adding background clutter and multiple individuals into the mix, generating long-term occlusions, severely increases the complexity of this task. State-of-the-art methods explore deep learning and complex graphical models to map this pose space, augmenting the spatial evidence with temporal consistency through motion features calculated across multiple video frames. It is hard to develop a method to simultaneously allow a long time interval and keep real-time performance. Trying to fill this gap, we developed the PastLens model, which instead of handcrafting arbitrary representations to capture temporal features, learn such representations from data. Our results show that merging spatial and temporal features in the mid layers may be a rather novel, although efficient, way to improve the encoding of temporal features for pose estimation, even on multi-person scenarios. The expected scientific contribution of our model to the multi-person pose estimation literature is:

(i) A cost-efficient alternative to impose temporal consistency to the HPE pipeline, through receptive field increase only, letting the temporal features' representation to be learned from data.

Additionally to this major contribution, we consider that our model achieved impactful accuracy increases without drastic changes to the base CNN, allowing the model to be easily attached to novel architectures. The following section brings a few limitations on our proposed test scenarios and some future directions to the continuity of this research project.

6.2 Limitations and Future Works

Most project decisions were made in order to reduce the scope of our work. Nonetheless, a few limitations arise from these decisions. Our pose estimates are always calculated based on motion and frame pairs disparity, therefore, it performs better when the input pair represents

some significant spatio-temporal change. Furthermore, we do not have enough time to train each external CNN from ground up for every proposed task in our model, which could improve results by itself. Consequently, we may have some difficulty in scenes where interactions or clutter are particular to a domain, given that all our imported training procedures were not domain specific. For instance, medical devices in operation rooms may have unique features that are unexpected by our model and may interfere in estimates accuracy, or sports' footage may include unexpected motion patterns. As future works, we propose addressing the limitations we pointed out and trying out more variations of our model, increasing the number of input frames to further stretch receptive fields, permuting the feature merging module along other layers, or even using multiple merging steps to cross-correlate spatio-temporal features.

The scale deficiency mentioned in the last chapter intrigued us and may be addressed in the near future. Another interesting trend is estimating the direct correspondence of 3D body meshes to 2D image evidence in real time (GÜLER; NEVEROVA; KOKKINOS, 2018). It is a fundamental step towards more interactive virtual reality games. Enhancing our model to deal with such a challenging task would be an enlightening path, since adding a dimension to the output will surely lead to unexpected results.

REFERENCES

ANDRILUKA, M.; IQBAL, U.; MILAN, A.; INSAFUTDINOV, E.; PISHCHULIN, L.; GALL, J.; SCHIELE, B. PoseTrack: A benchmark for human pose estimation and tracking. **CoRR**, [S.l.], v. abs/1710.10000, 2017.

ANDRILUKA, M.; PISHCHULIN, L.; GEHLER, P.; SCHIELE, B. 2D Human Pose Estimation: new benchmark and state of the art analysis. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2014. **Anais...** [S.l.: s.n.], 2014.

ANDRILUKA, M.; ROTH, S.; SCHIELE, B. People-tracking-by-detection and people-detection-by-tracking. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2008., 2008. **Anais...** [S.l.: s.n.], 2008. p. 1–8.

BELAGIANNIS, V.; AMIN, S.; ANDRILUKA, M.; SCHIELE, B.; NAVAB, N.; ILIC, S. 3d pictorial structures revisited: multiple human pose estimation. **IEEE transactions on pattern analysis and machine intelligence**, [S.l.], v. 38, n. 10, p. 1929–1942, 2016.

BELAGIANNIS, V.; WANG, X.; SHITRIT, H. B. B.; HASHIMOTO, K.; STAUDER, R.; AOKI, Y.; KRANZFELDER, M.; SCHNEIDER, A.; FUA, P.; ILIC, S. et al. Parsing human skeletons in an operating room. **Machine Vision and Applications**, [S.l.], v. 27, n. 7, p. 1035–1046, 2016.

BERTINETTO, L.; VALMADRE, J.; HENRIQUES, J. F.; VEDALDI, A.; TORR, P. H. S. Fully-Convolutional Siamese Networks for Object Tracking. In: COMPUTER VISION – ECCV 2016 WORKSHOPS, 2016, Cham. **Anais...** Springer International Publishing, 2016. p. 850–865.

BOGO, F.; KANAZAWA, A.; LASSNER, C.; GEHLER, P.; ROMERO, J.; BLACK, M. J. Keep it SMPL: automatic estimation of 3d human pose and shape from a single image. In: EUROPEAN CONFERENCE ON COMPUTER VISION, 2016. **Anais...** [S.l.: s.n.], 2016. p. 561–578.

BOURDEV, L.; MALIK, J. Poselets: body part detectors trained using 3d human pose annotations. In: COMPUTER VISION, 2009 IEEE 12TH INTERNATIONAL CONFERENCE ON, 2009. **Anais...** [S.l.: s.n.], 2009. p. 1365–1372.

CAO, Z.; SIMON, T.; WEI, S.-E.; SHEIKH, Y. Realtime multi-person 2d pose estimation using part affinity fields. **arXiv preprint arXiv:1611.08050**, [S.l.], 2016.

CHEN, L.; WEI, H.; FERRYMAN, J. A survey of human motion analysis using depth imagery. **Pattern Recognition Letters**, [S.l.], v. 34, n. 15, p. 1995–2006, 2013.

CHEN, Y.; WANG, Z.; PENG, Y.; ZHANG, Z.; YU, G.; SUN, J. Cascaded Pyramid Network for Multi-Person Pose Estimation. **arXiv preprint arXiv:1711.07319**, [S.l.], 2017.

CHERIAN, A.; MAIRAL, J.; ALAHARI, K.; SCHMID, C. Mixing body-part sequences for human pose estimation. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2014. **Proceedings...** [S.l.: s.n.], 2014. p. 2353–2360.

DANTONE, M.; GALL, J.; LEISTNER, C.; VAN GOOL, L. Body parts dependent joint regressors for human pose estimation in still images. **IEEE transactions on pattern analysis and machine intelligence**, [S.l.], v. 36, n. 11, p. 2131–2143, 2014.

DOERING, A.; IQBAL, U.; GALL, J. Joint Flow: temporal flow fields for multi person tracking. **arXiv preprint arXiv:1805.04596**, [S.l.], 2018.

EICHNER, M.; MARIN-JIMENEZ, M.; ZISSERMAN, A.; FERRARI, V. 2d articulated human pose estimation and retrieval in (almost) unconstrained still images. **International journal of computer vision**, [S.l.], v. 99, n. 2, p. 190–214, 2012.

FANG, H.; XIE, S.; TAI, Y.-W.; LU, C. RMPE: regional multi-person pose estimation. **2017 IEEE International Conference on Computer Vision (ICCV)**, [S.l.], p. 2353–2362, 2017.

FELZENSZWALB, P. F.; HUTTENLOCHER, D. P. Pictorial structures for object recognition. **International journal of computer vision**, [S.l.], v. 61, n. 1, p. 55–79, 2005.

FERRARI, V.; MARIN-JIMENEZ, M.; ZISSERMAN, A. Progressive search space reduction for human pose estimation. In: COMPUTER VISION AND PATTERN RECOGNITION, 2008. CVPR 2008. IEEE CONFERENCE ON, 2008. **Anais...** [S.l.: s.n.], 2008. p. 1–8.

GANAPATHI, V.; PLAGEMANN, C.; KOLLER, D.; THRUN, S. Real-time human pose tracking from range data. In: EUROPEAN CONFERENCE ON COMPUTER VISION, 2012. **Anais...** [S.l.: s.n.], 2012. p. 738–751.

GIRDHAR, R.; GKIOXARI, G.; TORRESANI, L.; PALURI, M.; TRAN, D. Detect-and-Track: efficient pose estimation in videos. **CoRR**, [S.l.], v. abs/1712.09184, 2017.

GIRSHICK, R.; SHOTTON, J.; KOHLI, P.; CRIMINISI, A.; FITZGIBBON, A. Efficient regression of general-activity human poses from depth images. In: COMPUTER VISION (ICCV), 2011 IEEE INTERNATIONAL CONFERENCE ON, 2011. **Anais...** [S.l.: s.n.], 2011. p. 415–422.

GÜLER, R. A.; NEVEROVA, N.; KOKKINOS, I. DensePose: dense human pose estimation in the wild. **CoRR**, [S.l.], v. abs/1802.00434, 2018.

HE, K.; GKIOXARI, G.; DOLLÁR, P.; GIRSHICK, R. B. Mask R-CNN. **CoRR**, [S.l.], v. abs/1703.06870, 2017.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2016. **Proceedings...** [S.l.: s.n.], 2016. p. 770–778.

HOLT, B.; ONG, E.-J.; COOPER, H.; BOWDEN, R. Putting the pieces together: connected poselets for human pose estimation. In: COMPUTER VISION WORKSHOPS (ICCV WORKSHOPS), 2011 IEEE INTERNATIONAL CONFERENCE ON, 2011. **Anais...** [S.l.: s.n.], 2011. p. 1196–1201.

HOLTE, M. B.; TRAN, C.; TRIVEDI, M. M.; MOESLUND, T. B. Human Pose Estimation and Activity Recognition From Multi-View Videos: comparative explorations of recent developments. **IEEE Journal of Selected Topics in Signal Processing**, [S.l.], v. 6, n. 5, p. 538–552, Sept 2012.

HOWARD, A. G.; ZHU, M.; CHEN, B.; KALENICHENKO, D.; WANG, W.; WEYAND, T.; ANDREETTO, M.; ADAM, H. MobileNets: efficient convolutional neural networks for mobile vision applications. **CoRR**, [S.l.], v. abs/1704.04861, 2017.

INSAFUTDINOV, E.; ANDRILUKA, M.; PISHCHULIN, L.; TANG, S.; LEVINKOV, E.; ANDRES, B.; SCHIELE, B.; CAMPUS, S. I. ArtTrack: articulated multi-person tracking in the wild. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2017. **Anais...** [S.l.: s.n.], 2017. v. 4327.

INSAFUTDINOV, E.; PISHCHULIN, L.; ANDRES, B.; ANDRILUKA, M.; SCHIELE, B. Deepercut: a deeper, stronger, and faster multi-person pose estimation model. In: EUROPEAN CONFERENCE ON COMPUTER VISION, 2016. **Anais...** [S.l.: s.n.], 2016. p. 34–50.

IONESCU, C.; PAPAVA, D.; OLARU, V.; SMINCHISESCU, C. Human3.6m: large scale datasets and predictive methods for 3d human sensing in natural environments. **IEEE transactions on pattern analysis and machine intelligence**, [S.l.], v. 36, n. 7, p. 1325–1339, 2014.

IQBAL, U.; GALL, J. Multi-Person Pose Estimation with Local Joint-to-Person Associations. **CoRR**, [S.l.], v. abs/1608.08526, 2016.

IQBAL, U.; MILAN, A.; GALL, J. Pose-Track: joint multi-person pose estimation and tracking. **CoRR**, [S.l.], v. abs/1611.07727, 2016.

JADERBERG, M.; VEDALDI, A.; ZISSERMAN, A. Speeding up Convolutional Neural Networks with Low Rank Expansions. **CoRR**, [S.l.], v. abs/1405.3866, 2014.

KE, L.; CHANG, M.; QI, H.; LYU, S. Multi-Scale Structure-Aware Network for Human Pose Estimation. **CoRR**, [S.l.], v. abs/1803.09894, 2018.

KINGMA, D. P.; BA, J. Adam: a method for stochastic optimization. **arXiv preprint arXiv:1412.6980**, [S.l.], 2014.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 2012. **Anais...** [S.l.: s.n.], 2012. p. 1097–1105.

LADICKY, L.; TORR, P. H.; ZISSERMAN, A. Human pose estimation using a joint pixel-wise and part-wise formulation. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2013. **Anais...** [S.l.: s.n.], 2013. p. 3578–3585.

LEBEDEV, V.; GANIN, Y.; RAKHUBA, M.; OSELEDETS, I. V.; LEMPITSKY, V. S. Speeding-up Convolutional Neural Networks Using Fine-tuned CP-Decomposition. **CoRR**, [S.l.], v. abs/1412.6553, 2014.

LIN, T.-Y.; MAIRE, M.; BELONGIE, S.; HAYS, J.; PERONA, P.; RAMANAN, D.; DOLLÁR, P.; ZITNICK, C. L. Microsoft coco: common objects in context. In: EUROPEAN CONFERENCE ON COMPUTER VISION, 2014. **Anais...** [S.l.: s.n.], 2014. p. 740–755.

LIU, Z.; ZHU, J.; BU, J.; CHEN, C. A survey of human pose estimation: the body parts parsing based methods. **Journal of Visual Communication and Image Representation**, [S.l.], v. 32, p. 10–19, 2015.

LÓPEZ-QUINTERO, M. I.; MARÍN-JIMÉNEZ, M. J.; MUÑOZ-SALINAS, R.; MADRID-CUEVAS, F. J.; MEDINA-CARNICER, R. Stereo Pictorial Structure for 2D articulated human pose estimation. **Machine Vision and Applications**, [S.l.], v. 27, n. 2, p. 157–174, 2016.

LOWE, D. G. Distinctive image features from scale-invariant keypoints. **International journal of computer vision**, [S.l.], v. 60, n. 2, p. 91–110, 2004.

MAO, X.; SHEN, C.; YANG, Y.-B. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 2016. **Anais...** [S.l.: s.n.], 2016. p. 2802–2810.

MEHTA, D.; SRIDHAR, S.; SOTNYCHENKO, O.; RHODIN, H.; SHAFIEI, M.; SEIDEL, H.-P.; XU, W.; CASAS, D.; THEOBALT, C. VNect: real-time 3d human pose estimation with a single rgb camera. **arXiv preprint arXiv:1705.01583**, [S.l.], 2017.

NEUBECK, A.; VAN GOOL, L. Efficient non-maximum suppression. In: PATTERN RECOGNITION, 2006. ICPR 2006. 18TH INTERNATIONAL CONFERENCE ON, 2006. **Anais...** [S.l.: s.n.], 2006. v. 3, p. 850–855.

NEWELL, A.; DENG, J. Associative Embedding: end-to-end learning for joint detection and grouping. **CoRR**, [S.l.], v. abs/1611.05424, 2016.

NEWELL, A.; YANG, K.; DENG, J. Stacked hourglass networks for human pose estimation. In: EUROPEAN CONFERENCE ON COMPUTER VISION, 2016. **Anais...** [S.l.: s.n.], 2016. p. 483–499.

OPTIMIZATION, G. Inc.,"Gurobi optimizer reference manual," 2015. **URL: http://www. gurobi. com**, [S.l.], 2014.

PAPANDREOU, G.; ZHU, T.; CHEN, L.; GIDARIS, S.; TOMPSON, J.; MURPHY, K. PersonLab: person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. **CoRR**, [S.l.], v. abs/1803.08225, 2018.

PAPANDREOU, G.; ZHU, T.; KANAZAWA, N.; TOSHEV, A.; TOMPSON, J.; BREGLER, C.; MURPHY, K. Towards Accurate Multi-person Pose Estimation in the Wild. **arXiv preprint arXiv:1701.01779**, [S.l.], 2017.

PARK, E.; HAN, X.; BERG, T. L.; BERG, A. C. Combining multiple sources of knowledge in deep CNNs for action recognition. In: IEEE WINTER CONFERENCE ON APPLICATIONS OF COMPUTER VISION (WACV), 2016., 2016. **Anais...** [S.l.: s.n.], 2016. p. 1–8.

PISHCHULIN, L.; ANDRILUKA, M.; GEHLER, P.; SCHIELE, B. Strong appearance and expressive spatial models for human pose estimation. In: IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION, 2013. **Proceedings...** [S.l.: s.n.], 2013. p. 3487–3494.

PISHCHULIN, L.; INSAFUTDINOV, E.; TANG, S.; ANDRES, B.; ANDRILUKA, M.; GEHLER, P. V.; SCHIELE, B. Deepcut: joint subset partition and labeling for multi person pose estimation. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2016. **Proceedings...** [S.l.: s.n.], 2016. p. 4929–4937.

RAMAKRISHNA, V.; KANADE, T.; SHEIKH, Y. Tracking human pose by tracking symmetric parts. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2013. **Proceedings...** [S.l.: s.n.], 2013. p. 3728–3735.

RAMAKRISHNA, V.; MUNOZ, D.; HEBERT, M.; BAGNELL, J. A.; SHEIKH, Y. Pose machines: articulated pose estimation via inference machines. In: EUROPEAN CONFERENCE ON COMPUTER VISION, 2014. **Anais...** [S.l.: s.n.], 2014. p. 33–47.

REVAUD, J.; WEINZAEPFEL, P.; HARCHAOUI, Z.; SCHMID, C. Deep Convolutional Matching. **CoRR**, [S.l.], v. abs/1506.07656, 2015.

ROGEZ, G.; RIHAN, J.; ORRITE-URUÑUELA, C.; TORR, P. H. Fast human pose detection using randomized hierarchical cascades of rejectors. **International Journal of Computer Vision**, [S.l.], v. 99, n. 1, p. 25–52, 2012.

SAPP, B.; TASKAR, B. Modec: multimodal decomposable models for human pose estimation. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2013. **Proceedings...** [S.l.: s.n.], 2013. p. 3674–3681.

SARAFIANOS, N.; BOTEANU, B.; IONESCU, B.; KAKADIARIS, I. A. 3d human pose estimation: a review of the literature and analysis of covariates. **Computer Vision and Image Understanding**, [S.l.], v. 152, p. 1–20, 2016.

SHI, Q.; DI, H.; LU, Y.; LV, F.; TIAN, X. Video pose estimation with global motion cues. **Neurocomputing**, [S.l.], v. 219, p. 269–279, 2017.

SHOTTON, J.; SHARP, T.; KIPMAN, A.; FITZGIBBON, A.; FINOCCHIO, M.; BLAKE, A.; COOK, M.; MOORE, R. Real-time human pose recognition in parts from single depth images. **Communications of the ACM**, [S.l.], v. 56, n. 1, p. 116–124, 2013.

SIGAL, L.; BLACK, M. J. Guest editorial: state of the art in image-and video-based human pose and motion estimation. **International Journal of Computer Vision**, [S.l.], v. 87, n. 1-2, p. 1, 2010.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. **arXiv preprint arXiv:1409.1556**, [S.l.], 2014.

SONG, J.; WANG, L.; VAN GOOL, L.; HILLIGES, O. Thin-slicing network: a deep structured model for pose estimation in videos. **ArXiv170310898 Cs**, [S.l.], 2017.

SUN, M.; KOHLI, P.; SHOTTON, J. Conditional regression forests for human pose estimation. In: COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2012 IEEE CONFERENCE ON, 2012. **Anais...** [S.l.: s.n.], 2012. p. 3394–3401.

TAYLOR, J.; SHOTTON, J.; SHARP, T.; FITZGIBBON, A. The vitruvian manifold: inferring dense correspondences for one-shot human pose estimation. In: COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2012 IEEE CONFERENCE ON, 2012. **Anais...** [S.l.: s.n.], 2012. p. 103–110.

TOKOLA, R.; CHOI, W.; SAVARESE, S. Breaking the Chain: liberation from the temporal markov assumption for tracking human poses. **2013 IEEE International Conference on Computer Vision**, [S.l.], p. 2424–2431, 2013.

TOMPSON, J. J.; JAIN, A.; LECUN, Y.; BREGLER, C. Joint training of a convolutional network and a graphical model for human pose estimation. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 2014. **Anais...** [S.l.: s.n.], 2014. p. 1799–1807.

TOSHEV, A.; SZEGEDY, C. Deeppose: human pose estimation via deep neural networks. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2014. **Proceedings...** [S.l.: s.n.], 2014. p. 1653–1660.

VARADARAJAN, S.; DATTA, P.; TICKOO, O. A Greedy Part Assignment Algorithm for Real-time Multi-person 2D Pose Estimation. **CoRR**, [S.l.], v. abs/1708.09182, 2017.

WEI, S.-E.; RAMAKRISHNA, V.; KANADE, T.; SHEIKH, Y. Convolutional pose machines. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2016. **Proceedings...** [S.l.: s.n.], 2016. p. 4724–4732.

XIAO, B.; WU, H.; WEI, Y. Simple Baselines for Human Pose Estimation and Tracking. **CoRR**, [S.l.], v. abs/1804.06208, 2018.

XIU, Y.; LI, J.; WANG, H.; FANG, Y.; LU, C. Pose Flow: efficient online pose tracking. **CoRR**, [S.l.], v. abs/1802.00977, 2018.

YANG, W.; LI, S.; OUYANG, W.; LI, H.; WANG, X. Learning Feature Pyramids for Human Pose Estimation. **CoRR**, [S.l.], v. abs/1708.01101, 2017.

YANG, Y.; RAMANAN, D. Articulated human detection with flexible mixtures of parts. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, [S.l.], v. 35, n. 12, p. 2878–2890, 2013.

ZHENG, L.; YANG, Y.; HAUPTMANN, A. G. Person Re-identification: past, present and future. **CoRR**, [S.l.], v. abs/1610.02984, 2016.