Optical Flow Estimation with Deep Learning, a Survey on Recent Advances

Chapter · January 2020

DOI: 10.1007/978-3-030-32583-1_12

3 authors, including:

Stefano Savian
Free University of Bozen-Bolzano
3 PUBLICATIONS **7** CITATIONS
SEE PROFILE

Mehdi Elahi
University of Bergen
85 PUBLICATIONS **1,484** CITATIONS
SEE PROFILE

Optical Flow Estimation with Deep Learning, a Survey on Recent Advances

Please note that this version is not the final version of the article. Final version can be accessed here: link.springer.com/chapter/10.1007%2F978-3-030-32583-1_12

Stefano SAVIAN, Mehdi ELAHI, Tammam TILLO

Abstract One of the many components used in Biometrics is optical flow estimation. This could be due to the fact that motion is an inseparable attribute of our (visual) world and hence it is a valuable resource of data needed to tackle many real-world problems. Indeed, technologies that use object detection, motion detection, object tracking, gait recognition as well as video compression heavily rely on optical flow estimation. This chapter explores recent advances in optical flow estimation, while mainly focusing on estimation techniques based on Deep Learning (DL). In fact, recent advancements in deep learning are seemingly making a shift in the optical flow estimation research field. This chapter begins with reviewing traditional (hand-crafted) approaches, then introduces the more recent approaches, and finally gets concluded with surveying deep learning approaches.

Key words: optical flow estimation, deep learning, convolutional neural network

1 Introduction

Biometric Systems attempt to detect and identify people based on certain physiological characteristics, e.g., fingerprints and face, or even behavioural characteristics, e.g., signature and gait. In recent years, there has been a large development in

Stefano SAVIAN
Free University of Bozen - Bolzano, Piazza Domenicani 3, Bozen-Bolzano, Italy, e-mail: ssavian@unibz.it

Mehdi ELAHI
Free University of Bozen - Bolzano, Piazza Domenicani 3, Bozen-Bolzano, Italy, e-mail: meelahi@unibz.it

Tammam TILLO
Free University of Bozen - Bolzano, Piazza Domenicani 3, Bozen-Bolzano, Italy, e-mail: ttillo@unibz.it

biometric systems, thanks to advances in Deep Learning (DL). DL techniques leverage the hierarchical architectures to learn discriminative representations and have contributed to some of the top performing biometric techniques. These techniques have also fostered numerous successful real-world biometric applications, e.g., face recognition and face identification [96].

Gait recognition can be a good example of a behavioral biometrics that uses the shape and motion cues of a walking person for identification. Gait could be performed at a distance, in contrast to the other biometric approaches such as fingerprint or iris scan [53]. The shape features are captured during gait phases, while motion features get captured during the transition between these phases. Still there are challenges in gait recognition, including variations in clothing, footwear, carrying objects, complicated background and walking speed [96, 88]. Motion clues can be obtained, even without involving additional hardware (e.g., accelerometers and lidar [95]), instead by extracting motion from the captured video [53]. For example, Xiao *et. al.* [103] obtained the very good performance by explicitly using the motion information, i.e., optical flow, as the input in a simple pose estimation framework. Motion estimation involves the estimation of optical flow which is the projection of 3D motion into 2D plane of the camera. Nonetheless, the "global" motion provided by the optical flow is composed of the motion of the objects in the scene and the ego-motion, i.e., the motion of the camera.

Optical flow estimation has a long history, and much research has been carried since the pioneering methods of Horn and Shunk [30] and Lucas Kanade [51] have been published in 1981. Hence, through more than 3 decades of history, there is a massive improvement of techniques used for different aspects of optical flow estimation [1]. In the particular case of small displacements, the problem of optical flow estimation has been almost completely solved [22]. The remaining challenges can be listed as: (i) fast motion, (ii) illumination changes, (iii) occlusions, and (iv) untextured regions. In these challenges, the optical flow estimation problems become ill-posed and hard to treat, analytically. However, very recent approaches have come after the wave of DL progress which had a massive impact on this field of research. Still this chapter would take advantage of the noted survey [22] (see the footnote), in assisting to review not only the current state-of-the-art optical flow estimation techniques but also highlighting the benefits and limitations of traditional approaches. Nonetheless, in order to motivate and explain the reasons behind the success of DL based optical flow estimation, the authors will show how some classical based top performing methods resemble a deep structure a keen to convolutional neural networks.

To sum up, to obtain high-level understanding of video contents for Computer Vision tasks, it is essential to know the object status (e.g., location and segmentation) and motion information (e.g., optical flow) [17]. Hence motion estimation is a valid data source to obtain non-intrusive and remote high quality biometrics.

Optical flow estimation is a fast evolving field of computer vision and to the authors' knowledge, there is only one comperhensive review by Tu *et. al.* [93], which surveys classical and DL based methods for optical flow. Nonetheless, this chapter

[1] In a valuable work, early history of the field has been surveyed by Fortun *et. al.* [22]. This work reviews the traditional works, and does not cover the very recent progresses.

provides a (more systematic) comparison of DL methods, and gives a slightly different categorization of DL methods (compared to the related works) and introduces a novel class of Hybrid methods. Readers are kindly pointed to Tu *et. al.* survey [93] for more detailed descriptions of performance measures for optical flow, optical flow color code, and optical flow applications (besides biometrics ones).

In conclusion, this chapter briefly reviews optical flow estimation methods, which can be used mainly in biometric applications. More particularly, it focuses on how the field of optical flow is evolving, what are the benefits and limitations of DL based methods, as well as what do DL and classical approaches have in common. The rest of the chapter is organized as follows:

In Section 2, a brief introduction to DL is provided. In Section 3, optical flow estimation is discussed by mainly reviewing the traditional approaches. Then DL for optical flow approaches are described and compared. This section continues with introducing the Hybrid approaches. Section 4 provides further discussions and lists a number of applications of the optical flow estimation in biometrics. The chapter is finalized with Section 5 by providing the conclusion.

2 Deep Learning

Deep Learning (DL) is a class of signal processing architectures which consist of connecting and stacking different convolutional layers and non-linear activation functions in order to generate flexible predictive models. These models are typically tuned by the *Backpropagation* algorithm using the target information, to indicate how much the (internal) parameters should be updated [45]. Deep learning is blooming and it already enables noticeable steps forward in various engineering applications, and influencing many signal processing fields, e.g., *Image Classification, Natural Language Processing (NLP)*, and *Time Series Analysis*. A notable comprehensive overview of the Deep learning field has been authored by LeCun, Bengio, and Hinton [45]. The book from Goodfellow *et. al.* is also considered a milestone on this topic [25]. It covers the main techniques, including the *Convolutional Neural Network (CNN)*, that is heavily used in Computer Vision. There is a lot of material on Deep learning models and applications and the reader is invited to further investigate the above mentioned papers or books, if interested for more details. Here existing (generic) architectures which have been tailored to optical flow estimation are briefly reviewed. It is assumed that the reader has a basic knowledge on the field, as it is not possible to describe all the technical details.

- **The state-of-the-art CNN for image recognition**: the pioneering work on this field is LeNet from LeCun *et. al.* [46], a 7 layer CNN, used for digit recognition. Subsequently, The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [79] was introduced to foster the development of models for visual recognition. AlexNet (see Krizhevsky *et. al.* [43]) uses a similar architecture as LeNet, but it is deeper, with more filters per layer, and with stacked convolutional layers. GoogleNet, Szegedy *et. al.* [89] also known as InceptionNet is

inspired by LeNet but implemented a novel element which they called inception module, GoogleNet achieved very high performance in ILSVRC 2014. VGGNet, Simonyan *et. al.* [81] consists of 16 convolutional layers and a very uniform architecture, however, it has a massive number of parameters (i.e., 138 Millions) to train. The winner of ILSVRC 2015 is Residual Neural Network (ResNet), Kaiming He *et. al.* [29]. This architecture introduced *skip connections* between convolutional layers. Thanks to this innovative technique, they were able to train a network with 152 layers while still having lower complexity in comparison to VGGNet. It achieves a top 5 error rate beating other baselines on ImageNet dataset.

- **Fully Convolutional Networks (FCNs)**: CNN models where every layer is convolutional, FCNs obtain very good performance in image segmentation, e.g., Long *et. al.* [49]. Additionally, U-Nets are FCNs specifically designed to produce accurate segmentation even with a relatively small dataset, Ronnenberger *et.al.* [78].
- **Siamese Networks** Koch *et. al.* [41] (originally introduced by Bromley *et. al.* [10]), are a class of neural network architectures that contain two or more identical subnetworks. The subnetworks have the same configuration and share the same parameters. Parameter updating is mirrored across both subnetworks. Sharing weights across subnetworks leads to less learnable parameters, thus less tendency to overfit. Siamese Networks output is usually a one dimensional feature vector which indicates a similarity or a relationship between two comparable (input) things.
- **Generative Adversarial Networks** (GANs): originally introduced by Goodfellow *et. al.* [26] are generative algorithms that can generate new data instances by learning the distribution of the input data. Differently from the previous architecture which are discriminative, GANs can learn to produce new data. One network, called the generator, generates new data instances, while the other, the discriminator, evaluates them for authenticity.

Although the deep learning based optical flow architectures could vary dramatically, still all the techniques presented in this chapter are drawn from the above mentioned models.

3 Optical Flow

Before discussing further the optical flow, we note that, traditionally, optical flow estimation rely on two assumptions:

- *Assumption 1*: pixel intensity remains unchanged along motion trajectory
- *Assumption 2*: motion appears locally as a translation.

Lets denote D as motion displacement vector in a two dimensional space and T as the *temporal* sampling step; By having D vector with the positional displacement

among 2 frames, we can compute the motion velocity vector V according to the following formula:

$$V = \lim_{T \to 0} \frac{D}{T} \tag{1}$$

$$w(x, y, t) = \Big(u(x, y, t), v(x, y, t), 1 \Big) \tag{2}$$

$$w^{\top} \nabla f(x, y, t) + \frac{\partial}{\partial t} f(x, y, t) = 0 \tag{3}$$

Eq. 3 is known as the *Optical Flow Equation*, where f is a 3 dimensional spatio-temporal field depending on the coordinates x, y and time t, and u and v are the displacements respectively along x and y axes (a detailed explanation can be found in [69]). The optical flow is defined as a 2 layers matrix with the same height and width of the input frame, where each of the two layers gives the offset of each pixel movement, where layer v is along y axis and layer u along x axis.

One of the earliest techniques proposed to solve the optical flow equation (Eq.3) are *Variational Methods*. As an example, Horn Schunck (HS) [30] approach (1981) adopted the minimization of a cost function where mean squared error with a regularization term has been sued. A work by Lukas-Kanade [51] from the same year proposed the minimization of an iterative cost function under slightly different assumptions, i.e., velocity vector V being constant in local patches, i.e. *Patch Based Methods*. However, both approaches have drawbacks. For example, the motion between 2 frames must be sufficiently small, the equations when discretized increase noise and locality assumptions result in poor motion accuracy [19].

3.1 Traditional Methods

In this section, we discuss mainly the traditional methods for optical flow estimation (see Table 1).

3.1.1 Variational Methods

One of the earliest class of optical flow [2] estimation methods were variational approaches. This class of approaches computes optical flow as the minimizer of an energy functional. One of the most effective and most simple variational methods has been developed by HS [30]. By exploiting the Brightness Constancy Equation (BCE) assumption and thus considering horizontal and vertical displacements to be sufficiently small, one can linearize the optical flow equation. Variational approaches,

[2] There is a slight difference between optical flow and flow fields, however, in computer vision these two term are usually used interchangeably [22].

hence, estimate the (optical flow) interframe displacement **w** by minimizing E, Eq. 4.

$$E(w) = \iint_{\Omega} \Big(D(u,v) + \alpha S(u,v) \Big) \mathrm{d}x\, \mathrm{d}y \tag{4}$$

Where D is a term that penalizes deviations of the data from the BCE and S is a smoothing term; Ω is the image region of locations x,y. All variational approaches aim at minimizing an energy function similar to E but exploiting different properties of the data, e.g., Gradient constancy Brox and Malik (BM) [11], higher order derivatives Papenberg *et. al.* [67], color model Mileva *et. al.* [61] and Zimmer *et. al.* [116]. Variational methods are biased towards the initialization which is usually the zero motion field because of all local minima, this approach selects the one with the smallest motion, which is not necessarily the correct solution [12].

3.1.2 Patch based methods

Methods based on Lucas Kanade (LK) approach work on the discrete domain: the two frames are divided into patches (regions) of fixed sizes, matched by minimizing a gradient. Accordingly, LK uses Newton-Raphson Technique for cost minimization. Most of the coarse-to-fine methods apply variations of LK on the frame, at different levels of granularity by looking for correspondences, first on a more large area, and then moving to smaller patches; or alternatively, by producing a rough estimate of the optical flow by downsampling the frames (see Fig. 1). Patch based methods are biased by the motion of large scale structures. The coarse-to-fine approaches have the drawback of hardly detecting small fast moving objects when the motion of bigger structures (or camera motion) has an overall high magnitude. Similarly to variational approaches, patch based methods have been improved very much, where majority of the improvements involve a different computation of the descriptor function. A descriptor is a function which is applied to all patches, used to produce a vector of similarities. Scale Invariant Feature Transform (SIFT) [50] and histogram of Oriented Gradients (HOG), as well as DAISY [91] are well-known descriptors which have been used in many computer vision field including optical flow estimation (e.g., [47]).

3.1.3 Patch based with variational refinement

A first unifying method was proposed by Brox and Malik (BM) [12] representing a big shift in performance of optical flow estimation. Brox and Malik formulate the problem of optical flow estimation in a variational refinement, but introducing an additional energy term $E_{descriptors}$ on SIFT and color. Many methods based on BM have a descriptor stage, a matching stage and a (variational) refinement stage to interpolate the optical flow in a sparse to dense manner.

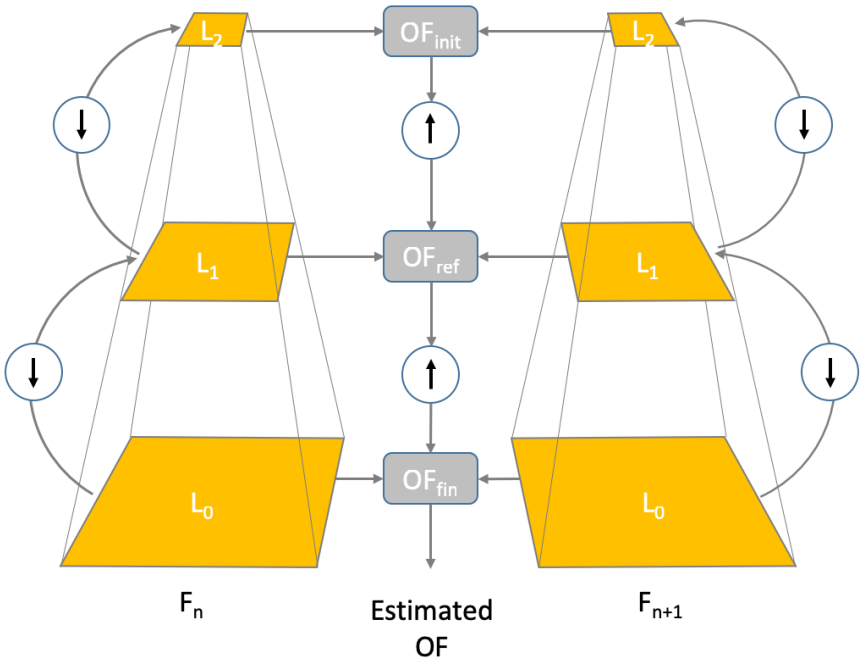

Fig. 1 Example of three layers frame pyramids. Input frames F_n, F_{n+1} are downsampled: L_0 are the input frames at native resolution, L_1, L_2 are the frames downsampled one and two times respectively. The optical flow is iteratively refined starting from the most downsampled frames. The initial optical flow (OF) is computed first at the lowest resolution L_2 (OFinit), upsampled and used as initial estimate for the next higher resolution layer, refined (OFref) and upsampled until the final high resolution layers.

For the matching stage, different techniques have been proposed over time (an overview can be found in [22]). In this part, we additionally review very recent developments of patch based with variational refinement (state-of-the-art at the time of publishing) that have been further improved by the partial integration of DL architectures, as described in Sec.3.6.

PatchMatch [7] is a general purpose computer vision algorithm to match arbitrary descriptors using K-Nearest Neighbors algorithm in a coarse-to-fine manner with random initializiation. FlowFields [4] is similar to PatchMatch for the propagation stage, but uses a kd-tree (a specific type of binary tree) to compute initial matches. Also noticeable is DiscreteFlow [60] which computes patch similarities in the discrete domain, using DAISY descriptors to find pixelwise correspondences among neighboring frames, and processing the vector of similarities with a Conditional Random Field (CRF), without coarse-to-fine optimization. Finally, FullFlow [16] also optimize an energy function as a CRF, over discrete regular grids.

A noticeable different approach, inspired by the success of Brox and Malik, is DeepFlow, Weinzaepfel et. al. [98]. DeepFlow is a variational optical flow with an additional loss function based on a deep matching algorithm on a classical variational framework which allows to integrate feature descriptors and matching. DeepFlow, is non-parametric and is based on a fine-to-coarse approach. Two consecutive frames are divided in four by four patches which are then convolved producing three dimensional response maps. The convolution operator outputs a stack of response maps providing higher values where patches have similar patterns. Subsequently, the obtained feature maps are convolved with larger patches to find coarser matches, with a structure a keen to CNN, but with no learnt parameters. This process is recursively applied to coarser patches. Finally all the feature maps, which are obtained by convolutions at different level of granularity, are processed: higher activation of the feature maps mean higher similarity among patches. A local maxima is computed to find dense matches (correspondences), which are then used to compute the optical flow. DeepFlow most important innovation is its matching algorithm, DeepMatching [76], which has been used as descriptor and matching algorithm. Compared to various classical methods mentioned using i.e. SIFT and HOG, DeepMatch obtains similar performance for small displacements, while drastically outperforming classical methods on large displacements. For these reasons, many top performing methods, make use of DeepMatch along with different variational refinement methods (see Fig. 2).

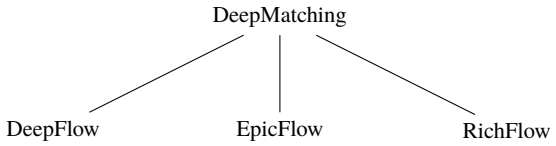

Fig. 2 Sparse to dense refinement methods applied to DeepMatching. DeepMatching algorithm is used to compute correspondences which can be refined by DeepFlow, EpicFlow or RichFlow.

EpicFlow is one of the recent methods for refinement and post-processing task (Revaud *et al.* [75]) and it has been adopted by several works [104, 3, 5, 16, 32, 108]. EpicFlow is built on DeepMatch and random forests; DeepMatch is used to compute matches, while structured forest (Structured Edge Detectors SEDs Dollar *et. al.* [18]) are used to compute image edges, exploiting the local structure of edges by looking at the information gain of random forests. The additional edges information allows to further densify the sparse matches and improves the variational refinement energy function. The energy function is further improved by using geodesic distance instead of euclidean distance, obtaining a more natural model for motion discontinuities (further details in the paper [75]). EpicFlow further improves DeepFlow on large discontinuities and occluded areas, nonetheless outperforms all state of the art (classical) coarse-to-fine approaches.
Due to its performance, EpicFlow has been integrated in Robust Interpolation of Correspondences for Large displacement optical flow (RicFlow), Hu *et. al.* [31].

RicFlow is based on Epicflow in the sense that uses DeepFlow and SED to compute the flow. Additionaly to EpicFlow, the input images were segmented in superpixel, to improve the method over input noise and provide a better initialization. RicFlow was among the best performing state-of-the-art methods on Sintel *clean pass* at the time of publishing.

Table 1 Overview of handcrafted methods; the handcrafted methods mentioned here are not real-time

PAPER YEAR REF.	CONTRIBUTION	BUILDS ON	LIMITATION
DeepFlow 2013 [98]	DeepMatching Fine-to-coarse image pyramids	BM [11]	Sparse matches
EpicFlow 2015 [75]	Improved matches interpolation Used in [104][3][5][16][32][108]	DeepMatching [76], SED [18]	Input noise
RicFlow 2017 [31]	Further improved EpicFlow interpolation	EpicFlow	Strongly relies on (dense) input matches
FlowFields 2015 [4]	Binary tree for patch matching	EpicFlow	Handcrafted features for match
DenseFlow 2013 [84]	Segmentation. Fully-connected inference method	EM	Computationally Expensive
ProbFlow 2017 [97]	Predicts optical flow and uncertainty	HS, FlowFields	Small EPE improvement
DiscreteFlow 2009 [60]	Uses CRF to reduce patches search space	See [22]	Semi-dense optical flow
CPM 2016 [32]	Discrete coarse-to-fine	PatchMatch [7] EpicFlow	Small details are lost

3.2 Deep Learning Approaches

In the previous section, it has been shown that DeepFlow, one of the top performing handcrafted algorithms resembles a convolutional structure similar to deep learning models, but with no learnt parameters. Perhaps triggering a new line of research based on deep convolutional structures in the field of optical flow estimation. In this section, advances of models based on deep learning are reviewed (see Table 2).

3.2.1 Development of DL based Optical Flow Estimation

Single and Stacked Architectures

The first deep learning optical flow architecture has been introduced by Fisher *et. al.* and named FlowNet [19], FlowNet directly estimates the optical flow using a generic CNN U-Net architecture [78]. Due to the lack of data with optical flow groundtruth the authors generated a new dataset: “Flying Chairs”, Meyer *et. al.* [57], which is a synthetic dataset with optical flow ground truth. It consists of more than 20K image pairs and corresponding flow fields of 3D chair models moving with just affine motion in front of random backgrounds. This dataset is necessary for network convergence, since CNN typically has a very large number of trainable weights (tens of Millions) requiring a considerable number of input data to avoid overfitting.

The original paper proposes two slightly different model architectures FlownetS (Flownet Simple) and FlownetC (Flownet Correlation). FlownetS consists of a CNN receiving two stacked input RGB frames which are then supervisely trained on the optical flow groundtruth. Similarly, FlowNetC is also supervisely trained on the groundtruth, but instead of working with a stacked input the two frames are fed to two identical branches which are merged on a later stage by a correlation layer, the correlation layer perform cross-correlation (cost volume) between the feature maps of the two input, enabling the network to compare each patch with no additional learnt parameters (at the correlation layer).

Both networks, upsample feature maps by upconvolutions at the output side of the network to increase the resolution of the computed optical flow, degraded by the stacked convolutions and pooling layers at the contractive side of the network (see Fig. 3). The expanding part of the network is composed of “upconvolutional" layers: unpooling and deconvolution. There are 4 upconvolution layers in the refinement part and for computational reasons the flow is finally upsampled to full resolution by bilinear upsampling. Skip connections are used to connect layers on the contractive part to the expanding (refinement) part providing additional information of flow level features at the upsampling stage.

Data augmentation is very important for model generalization. Augmentation has been performed in place both to the image pair and to the grountruth flow fields, it includes geometric transfomations, gaussian noise, multiplicative color and additive brightness changes. Finally, the authors trained the network by minimizing the squared error on endpoint error EPE. As said previously, EPE is the euclidean distance between network estimates and groundtruth flow. For further details on EPE and other optical flow metrics please refer to Tu *et. al.* [93]. Training on EPE is not optimal for small displacements, as the euclidian distance only gives information on the error magnitude while omitting error direction information, however it allows the architecture to perform well in case of large displacements such as in the Sintel benchmark. One important discovery is that although FlowNetS and FlowNetC have been trained on synthetic data, they can also perform well on natural scenarios. However, the main drawback is the low accuracy in case of small and simple movements which instead are conditions where traditional methods perform well.

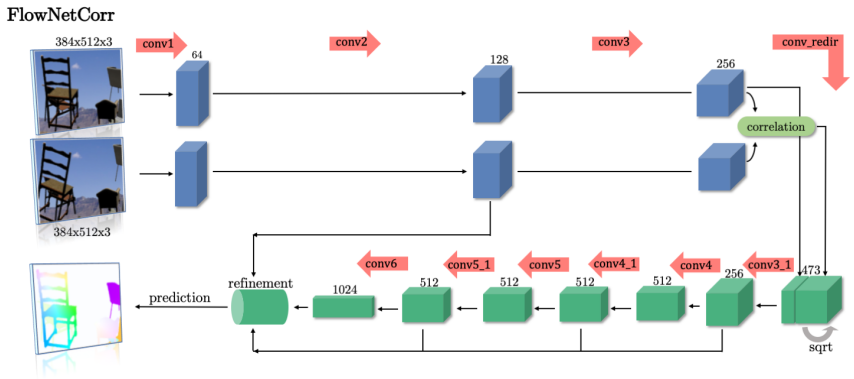

Fig. 3 FlowNetC architectures [19].

A second drawback is the over-smoothed flow fields produced, and the fact that a variational post-processing stage is required, as noticed by Hui *et. al.* [34].

To overcome FlowNet limitations, Ilg *et. al.* proposed FlowNet 2.0 [37], which stacks FlownetC and FlownetS and new designed FlowNet-SD (Small Displacement) to perform the flow estimation. Flownet-SD (see Fig. 4), has larger input feature maps and is trained on a dataset with small displacement, ChairsSDHom. Furthermore, after each subnetwork the flow is warped and compared with the second image and the error is fed to a fusion network that takes as inputs the estimated flows, the flows magnitudes and the brightness error after warping. The fusion network contracts and expands to full resolution producing the final flow fields.

Due to the large size of FlowNet 2.0, i.e., around 38 Millions parameters (see Table 4), its subnetworks have been trained sequentially by training one subnetwork while freezing the weights of the others. Moreover, the authors have generated a new more realistic dataset, Things3D [37] to be robust to untextured regions and to produce flow magnitude histograms close to those of the UCF101 [83] dataset, which is composed of real sequences. A very important finding is the impact of training schedule on network performance. Solely training on the more complex Things3D is worse than using the simpler FlyingChairs, and training on a mixture of FlyingChairs and Things3D also does not lead to better performance. The order of presenting the data affects model accuracy; the best schedule is training on FlyingChairs and finetuning on Things3D. Also subnetworks FlowNetS and FlowNetC can benefit of around 20-30 % of improvement when trained with the above mentioned schedule. The authors conjecture is that FlyingChairs allows the network to learn color matching and that the refinement with Things3D improves performance under realistic scene lighting. Flownet 2.0 outperforms EpicFlow, and obtain state of the art performance on Sintel *final pass*, at the time of publishing.

FlowNet and FlowNet 2.0 are important milestones of optical flow estimation and serve as building block for other methods. Ilg *et. al.* modified FlowNetC to estimate the confidence interval on the estimated optical flow in [36]. Xian *et. al.* [102] use FlowNetS and add a multi-assumption loss function (brightness constancy,

gradient constancy and image-driven smoothness assumption) in the expanding part during the network training. On the contrary, FlowNet 3 (Ilg *et. al.* [38]) further improves FlowNet 2.0, by taking out the small displacement network, removing explicit brightness error and, add residual connections in the stack based on [66]. They also modified the stack and in particular FlowNetC to jointly compute forward and backward flow consistency and estimate occlusions. The authors demonstrate that efficient occlusions estimates come at no extra cost.

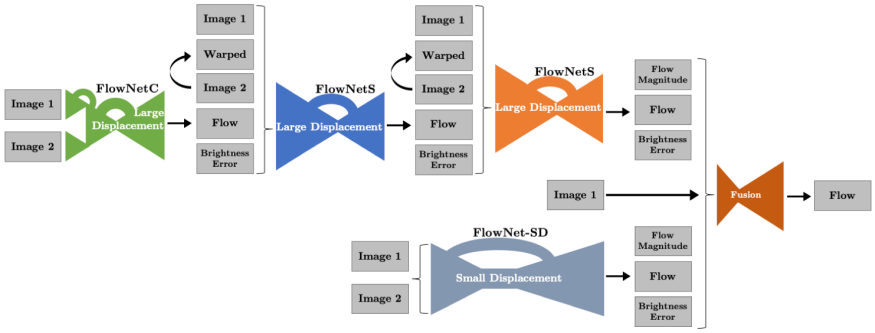

Fig. 4 FlowNet 2.0 architecture [37], which consists of FlowNetS and FlowNetC stacked.

Coarse-to-Fine Iterative Refinement

The first coarse-to-fine end-to-end approach is proposed by Ranjan *et. al.*, proposing Spatial Pyramid Network (SpyNet) [71], which combines the classical coarse-to-fine approach for optical flow estimation with deep neural networks. At each level of the pyramid a CNN is trained independently, meaning that each level of the pyramid (independently) deals with motion within a certain boundary of displacement. This architecture allows SpyNet to perform at different magnitudes of displacement. In fact, since every level of the pyramid deals with a fixed range of motion, the final optical flow is produced by iteratively up-sampling the coarse optical flow estimates and warping it with higher resolution frames. Hence, at each level of the pyramid L, a pixel motion at the top level corresponds to 2^{L-1} pixels at the full resolution [87] (see Fig. 1). Coarse-to-fine iteration benefit the estimates as there is no need for a full computation of the cost function, which is a bottleneck for real-time optical flow estimation.

The authors show how SpyNet improves the results of FlowNet which perform badly in case of small movements, while obtaining the same performance in case of large motion. Nonetheless, SpyNet has 96 % less weights compared to FlowNet, consisting "only" of 1 Million parameters, one order of magnitude less than FlowNet and two compared to FlowNet 2.0, (Sec. 3.3 for further insights on the number of

weights).

Following this line of research, Sun *et. al.* [86] presented PWC-Net. PWC-net is a pyramidal coarse-to-fine CNNs based for optical flow estimation. PWC-net uses pyramid, warping, and cost volume (see Fig 5). The network is trained end-to-end in a similar manner as SpyNet, however with some differences: i) SpyNet warps frames with coarse estimates of the flow, while PWC-net warps feature maps, ii) SpyNet feeds CNN with frames while PWC-net inputs a cost volume iii) PWC-net data augmentation do not include Gaussian noise (more details in Sec. 3.4). PWC-net image pyramids are end-to-end learnable and the cost volume is produced exploiting FlowNetC correlation layer. Finally PWC-net uses a context network to exploit contextual information for refinement. PWC-net is outperforming all methods to date in the challenging Sintel *final pass* 4.2. It is the first time that an end-to-end

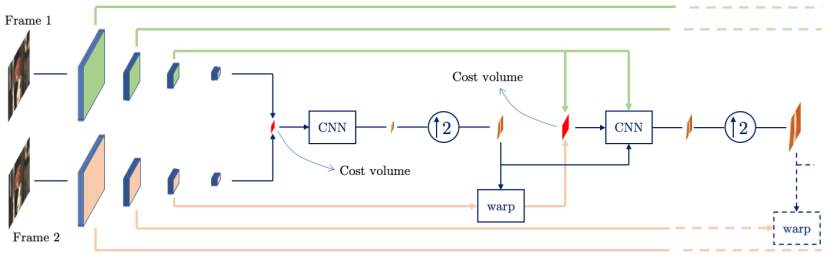

Fig. 5 PWC-net architecture [86].

method outperforms well-engineered and fine-tuned traditional methods. At the time of writing PWC-net is still the top performing two frames optical flow estimator on Sintel *final* and it is used as building block for the top performing multi-frame optical flow estimators Neoral *et. al.* [62] and Ren *et. al.* [74]. Finally, Inspired by PWC-net, Fang *et. al.* [21], Hui *et. al.* [34] and [21] proposed similar methods, but with lower performance.

3.2.2 Other Important DL based Methods

Multi Objective

Optical flow is an important piece of information for motion segmentation and action recognition. Ilg *et. al.* already show how the flow fields generated by FlowNet 2.0 matches and outperforms classical optical flow methods, when plugged into the CNN temporal stream of Symonyan *et. al.* [81]. This configuration produces high level action recognition. The state-of-the-art performance has been obtained also in

motion segmentation, by plugging the optical flow of FlowNet 2.0 to [40] and [65]. Moreover, Cheng *et. al* [17] further elaborate the idea and propose one unified Deep Learning framework, called SegFlow, for joint estimation of object segmentation and optical flow. SegFlow architecture has two branches: FlowNetS for the optical flow and a residual CNN, Resnet-101 [29]. Feature maps are merged between two branches at the final layers. Training is done iteratively: weights are initialized according to FlowNetS and ResNet-101. When optimizing the segmentation branch, the optical flow branch weights are frozen. The segmentation is trained on the DAVIS dataset [68], with additional affine data augmentation. Similarly, when training the optical flow, the segmentation branch is fixed and weights are only updated in the flow network using optical flow datasets with groundtruth: Sintel, KITTI, Monkaa and Driving, Sec. 3.4. Both networks benefit from each other and the results are state-of-the-art for both segmentation and optical flow estimation, with accuracy doubled compared to FlowNet.

Indirect-supervised and Semi-supervised Methods

Indirect-supervised approaches treat the optical flow estimation as a frame reconstruction problem. Although these unsupervised methods do not need to use large datasets as training samples, the overall accuracy is slightly inferior to the supervised approaches. These methods still need to be trained with groundruth data to tune their weights, but they do not need specifically optical flow groundtruth data to model the optical flow. Instead they use a *proxy task* i.e. frame synthesis. All of the methods below rely on frame warping which is a differentiable operation and allow backpropagation for network tuning (see Table 3). Differences among indirect-supervised are rather small and mostly involve different constrains on the cost function: photometric or geometrical error function. Thus a common optical flow estimation pipeline is: i) let the network estimate the flow fields, ii) warp the frame with the flow fields, iii) measure the photometric loss between the synthetized frame and the groundtruth [100].

Ahmadi and Patras [1] presented a method for training a CNN using the UCF101 dataset [83] for motion estimation without explicit optical flow groundtruth data, instead of exploiting the optical flow equation, Eq. 3, similarly to traditional HS. The architecture proposed is very similar to FlowNetS and it has been trained on the real scenario dataset UCF101. Performance on Sintel are very close to FlowNetS and FlowNetC. Yu et al. [39] also design a network similar to FlowNetS, with a photometric loss function. Zhu [114] *et. al.* also use FlowNetS trained on *photometric loss*, but initialize the learning on proxy ground truth provided by FlowFields [4]. Ren et al. [73] also train FlowNetS on frame interpolation error, but with an additional nonlinear data term. Niklaus *et. al.* [64] jointly perform interpolation and flow estimation, with results comparable to FlownetS. Long et al. [48] train a CNN for optical flow estimation by interpolating frames, with some more minor differences. An U-net is trained to synthesize the middle frame in the training phase. Afterwards, the frame

Table 2 Overview of deep learning methods. $*$ =top among methods working with frame pairs

PAPER YEAR REF.	CONTRIBUTION	BUILDS ON	LIMITATION
ContinualFlow 2018 [62]	Top Performing (Sintel final pass) Occlusion estimates Multiple frame	PWC net, GRU	Not top on Sintel *clean*
MFF 2018 [74]	Multiple frame	PWC net,	Not top on Sintel *clean*
FlowNet 2015 [19]	First DL model	U-net	Artifacts for small motions Oversmoothed flow
FlowNet2.0 2016 [37]	Stacked FlowNetS and FlowNetC Improved training schedule	FlowNetS, FlowNetC	High number of weights
FlowNet3 2018 [38]	FlowNetC with FWD/BCK consistency check	ResNet, FlowNet2.0	High number of weights
FlowNetH 2018 [36]	Confidence Measures	FlowNetS-C	Focus only on confidence estimates
SpyNet 2017 [71]	End-to-end Coarse-to-fine on frames	Coarse-to-fine, U-net	Piecewise training
PWC-Net 2018 [86]	Top on Sintel *final** Coarse-to-fine on features Cost volume layer	SpyNet FlowNetC	Not top on Sintel *clean*
SegFlow 2017 [17]	Segmentation and optical Flow estimation	FlowNetS, ResNet	Two times EPE w.r.t. 1st ranked
Xiang *et. Al.* 2018 [102]	Traditional priors on cost function	FlowNetS	Small improvement
Fang *et. al.* 2018 [21]	Two branch CNN lightweight	Coarse-to-fine approach	Not tested on Sintel *final*

correspondences are obtained directly through the same network backpropagation, a process called analysis by synthesis [109]. The network uses triplets of consecutive frames, the first and last are used as input and the middle frame serves as groundtruth. Network output is two matrices of gradients with respect to the input; the gradients are obtained from the network through backpropagation, which produces sensitivity maps for each interpolated pixel. However the backpropagation pass is computationally expensive and, especially in unstructured or blurry regions the derivatives are not necessarily well located.Zhu and Newsam [115] extend DenseNet architecture [33] by adding Dense connectivity to FlowNetS layers, however the network accuracy is two times worsen if compared to the original FlowNetS. Meister *et. al.* uses

a loss based on the CENSUS transform [28] and check forward and backward flow consistency, explicity integrating occlusion reasoning.

Other approaches instead use *geometrical reasoning* for self-supervision. Alletto *et. al.* [2], trained a network to estimate a global homography and a second network to estimate the residual flow after warping using the homography. The method has been validated on KITTI, with performance similar to FlowNetS. Vijayanarasimhan *et. al.* presented SfM-Net [94] which decomposes scene motion in terms of scene and object depth, camera motion and 3D object rotations and translations. Given a sequence of frames, SfM-Net predicts depth, segmentation, camera and rigid object motions, and converts those into motion fields. Wulff *et. al.* [100] noticed that not always the flow fields can be learnt by photometric error due to untextured regions and lack of context information. They trained a temporal interpolation network on frame synthesis on large set of videos without involving any prior assumption and fine tune the network on groundtruth data from KITTI and Sintel. The explicit use of groundtruth data drastically improves performance, and the architecture outperforms FlowNetS and SpyNet, at the cost of not being fully unsupervised. Similar to self supervision, Lai *et. al.* [44] use GAN. They used a discriminator network trained on optical flow groundtruth. The discriminator is used in adversarial loss to learn the structural patterns of the flow warp error without making assumptions on brightness constancy and spatial smoothness. Once the discriminator network is trained, the network can be trained in any dataset, providing a loss for unsupervised training.

Yin *et. al.* presented GeoNet [110] exploiting geometric relationships extracted over the predictions of depth, rigid and non rigid parts and then combined as an image reconstruction loss. They separate static and dynamic scene parts. Depth maps and camera poses are regressed respectively and fused to produce the rigid flow. Furthermore, the second stage is fulfilled by the ResFlowNet, i.e. Residual FlowNet using the output from the rigid structure reconstructor, to predict the corresponding residual signal for handling dynamic objects flow fields. The final flow is a combination of the rigid and non rigid estimated flow with an additional geometric constrain. Similarly to GeoNet, Ranjan *et. al.* [72] propose a framework for estimation of depth, camera motion, optical flow and segmentation using neural networks that act as adversaries, competing to explain pixels that correspond to static or moving regions, and as collaborators through a moderator network that assigns pixels to be either static or dynamic. This and GeoNet are among the best unsupervised methods, however their performance are not comparable to supervised and classical methods and their benchmarks are mostly from the automotive domain, e.g. KITTI .

3.3 Deep Learning Networks Comparison

One of the challenges of deep learning models is to limit the number of network parameters to avoid overfitting and reduce memory footprint. Unlike Stacked approaches, *deep* coarse-to-fine SpyNet and PWC-net do not need to deal with large

Table 3 Overview of indirect-supervised learning methods. ∗ = similar architecture

PAPER YEAR REF.	CONTRIBUTION	BUILDS ON	LIMITATION
GeoNet 2018 [111]	Rigid and non rigid motion	ResNet-50	Automotive domain
Ahmadi *et. al.* 2016 [1]	Photometric loss	HS, coarse-to-fine	Train on DeepFlow
Jason *et. al.* 2016 [39]	Photometric loss and smoothness	FlowNetS	Automotive domain
MIND 2016 [48]	Analysis by synthesis	FlowNetS*	Results only on Sintel train
Ren *et. al.* 2017 [73]	Photmetric loss	FlowNetS*	Low performance
DenseNet 2017 [115]	Extends DenseNet	DenseNet [33]	Large memory footprint
Zhu *et. al.* 2017 [114]	FlowFields proxy groundtruth	FlowFields	Train on FlowFields
Yang *et. al.* 2018 [107]	FlowNet 2.0 proxy groundtruth	FlowNet 2.0	Rely on FlowNet 2.0, explicit ground truth
Wulff *et. al.* 2018 [100]	Fine tune on groundtruth data	MIND*	Groundtruth
UnFlow 2017 [58]	Occlusion estimates	FlowNetsC	Results only on Sintel Train
Ranjan *et. al.* 2018 [72]	Multi-objective (segmentation)	FlowNetC	Custom split for evaluation
Lai *et. al.* 2018 [44]	GAN applied to optical flow	FlownetS	Uses groundtruth
TransFlow 2017 [2]	L1 norm (Charbonnier)	FlowNetS	Automotive domain
SfM-Net 2017 [94]	Depth, occlusion mask estimation, photometric error	FlowNetS*	Not tested on Sintel

motions thanks to the image pyramid for the first and feature pyramid for the latter. It has been shown that coarse-to-fine image and feature pyramids require less weights, and at the same time lead to state-of-the-art performance for PWC-net. LiteFlownet

and other minor coarse-to-fine models have not been included as they are not the first published or top performing methods.

Principles	**FlowNetS**	**FlowNetC**	**FlowNet2**	**SpyNet**	**PWC-Net**
Pyramid	-	3-level	3-level	Image	6-level
Warping	-	-	Image	Image	Feature
Cost volume	-	single level large range	single level large range	-	multi-level small range
#parameters (M)	38.67	39.17	162.49	1.2	8.75
Memory (MB)	154.5	156.4	638.5	9.7	41.1
Forward (ms)	11.40	21.69	84.80	-	28.56

Table 4 Supervised Deep Learning model comparison. Coarse-to-fine approaches require less parameters and lead to state-of-the-art performance. Data source [87]

3.4 Optical Flow Datasets

It has been already discussed that it is difficult to obtain the proper data for training deep optical flow models. For this reason, a fundamental contribution of FlowNet and FlowNet 2.0 approaches are the (computer graphics) datasets that have been released to train the networks: FlyingChairs and FlyingThings3D (see Sec.3.2.1).

While very useful, however, it is not still clear how to generate more data that can generalize well on real world videos. In a recent follow-up paper, Mayer *et. al.* [56] performs an in-depth analysis of what are the characteristics of good training datasets, as research is shifting from proposing models to generate abundant data for supervised learning. There are further findings reported in the paper: the authors discovered that i) artificially rendered data can well generalize on real videos, ii) if training with a single dataset, complex lighting and post processing effects worsen the performance, iii) training on different datasets with an increasing level of complexity leads to best performance.

In the following, we briefly describe the training dataset that, to the best of our knowledge, are the largest with dense optical flow groundtruth:

- **FlyingChairs** [19] is a synthetic dataset which consists of more than 22K image pairs and their corresponding flow fields. Images show renderings of 3D chair models moving in front of random backgrounds from Flickr [3]. Motions of both the chairs and the background are purely planar. FlyingChairs2 contains additional minor modalities [35].
- **ChairsSDHom** [37] is a synthetic dataset of image pairs with optical flow ground truth. ChairsSDHom is a good candidate for training networks on small displacements, it is designed to train networks on untextured regions and to produce flow

[3] `https://www.flickr.com/`

magnitude histograms close to those of the UCF101 dataset. ChairsSDHom2 contains additional minor modalities which are discussed in [35].

- **FlyingThings3D** [57] is a dataset rendered of image pairs from 3D models (randomly shaped polygons and ellipses) with simple but structured background. Foreground objects follow linear trajectories plus additional non-rigid deformation in 3D space.
- **Sintel** has 1K training frames drawn from the entire video sequence of an open source movie. Sintel is not sufficient to train a network from scratch [19], however, it can be used for fine-tuning in the context of deep learning. This dataset is mostly used as challenging benchmark for evaluation of large displacement optical flow (Sec 4.1).
- **Monkaa** Mayer *et. al.* [57] contains 8,5K frames, and it is drawn from the entire video sequence of a cartoon, which is similar to Sintel, but more challenging. Monkaa contains articulated non-rigid motion of animals and complex fur.
- **Playing for Benchmarks** Richter *et. al.* [77] is based on more than 250K high-resolution video frames, all annotated with ground-truth data for both low-level and high-level vision tasks, including optical flow. Ground-truth data (for variety of tasks) is available for every frame. The data was collected while driving, riding, and walking a total of 184 kilometers in diverse ambient conditions in a realistic virtual world.
- **KITTI2012** Geiger *et. al.*, [24] contains almost 200 frames of stereo videos of road scenes from a calibrated pair of cameras and lidar mounted on a car. While the dataset contains real data, the acquisition method restricts the ground truth to static parts of the scene thus the main motion is given by the ego-motion of the camera [57]. KITTI2015 Menze *et. al.* [59] (800 frames) is obtained by fitting 3D models of cars to the point clouds. However, the ground truth optical flow is sparse.
- **Driving** [57] contains more than 4K frames of virtual scenes in an naturalistic, dynamic street setting from the viewpoint of a driving car, made to resemble the KITTI datasets.

It is worth noting that, FlyingChairs and FlyingThings3D contain well textured background. By ablation studies Meyer *et. al.* [56] discovered that background textures help to better perform on unseen datasets, and yield to best results on Sintel even though the motion where they have been trained is unnatural. The mentioned dataset are large enough to train deep CNN with just some additional data augmentation. Differently, Monkaa contains very difficult motion on repetitive and monotonous texture which have been found to be counterproductive for training.

3.5 Training Schedule and Data Augmentation

In Sec. 3.2.1, it has been shown that just by retraining FlowNetC with a new schedule, on FlyingChairs followed by the more refined FlyingThings3D it is possible to improve its performance by a 20-30% underlying the importance of training schedule.

Nonetheless, Deqin *et. al.* in [87] further demonstrate this concept by retraining PWC-net and FlowNetC. They further increase the accuracy of PWC-net by 10% and show that it is possible to further improve FlowNetC by 56% solely by retraining the network with their new training schedule and a smoother data augmentation (i.e., no additive Gaussian noise), outperforming the more complex FlowNet2 by 5%.

These results show that if trained improperly, a good model can perform poorly. Meaning that a fair comparison of deep learning models should consider the same training datasets and scheduling in order to disentangle model and training data contributions.

3.6 Hybrid Methods

This class of optical flow estimation techniques integrate end-to-end learnt approaches with traditional architectures (see Table 5). Two main branches of hybrid methods can be identified:

- **deep feature based**: obtained by partially integrating the flow estimation pipeline with CNN. In this context cost volume, matching or descriptors are obtained by deep learning while other building blocks are traditional, i.e. variational refinement.
- **scene understanding**: CNN are used to differentiate frames regions based on object properties or semantics. This information is integrated with prior knowledge on the motion field, e.g. motion is prominent on foreground objects while the background has a smoother and more linear motion.

3.6.1 Feature Based

Hybrid Deep learning patch based methods make use of learned matching functions [112] [113] [52]. These architectures have been adopted to extract and match descriptors for optical flow. The most relevant examples are PatchBatch [23], Deep DiscreteFlow [27], DCFlow [104], and Exploiting Semantic Information and Deep Matching for Optical Flow [3] (which also integrates semantic information, and is discussed in Sec. 3.6.2). These methods exploit learned matching functions which are then integrated into handcrafted methods.

As discussed in Sec. 3.1.2 different approaches have been developed to obtain descriptors and aggregate information from local matches. However handcrafted patch based optical flow estimators are limited by the computational cost of computing a 4D cost volume [16] or by number of pixelwise flow proposals at the initialization stage [60] [7]. In these cases integrating learned convolutional networks on the handcrafted pipeline lead to better accuracy and orders of magnitude faster inference. DCFlow [104] is inspired by [16] and constructs a cost volume by using a 4 Layer CNN, this cost volume produces a feature space for all pixels in a way that matching scores are then computed by a simple internal product in this space, refined

by EpicFlow post processing. Learning feature embedding and matches with a CNN allows the method to be more resilient to patch appearance changing and make large search space computationally feasible, nonetheless, the dimensionality of the feature space allows to find a trade off between computational cost and performance while drastically requiring less parameters (around 130K) if compared with fully learned methods.

Deep DiscreteFlow independently train a context network with a large receptive field size on top of a local network using dilated convolutions on patches. It performs feature matching by comparing each pixel in the reference image to every pixel in the target image: matching points on a regular grid in the reference image to every pixel in the other image, yielding a large tensor of forward matching costs, similarly to DiscreteFlow a CRF is used for flow refinement. ProFlow is also a MultiFrame method and is discussed in Sec. 3.7.

CPM [32] + RichFlow [31] + Maurer *et. al.* [55] ⟶ ProFlow [54]

Flow Fields [4] ⟶ Flow Fields CNN [5]

FullFlow [16] ⟶ DCFlow [104]

DiscreteFlow [60] ⟶ Deep DiscreteFlow [27]

PatchMatch [7] ⟶ PatchBatch [23]

Yamaguchi *et. al.* [106] + EpicFlow ⟶ ESIDM [3]

3.6.2 Domain Understanding

This section refers to methods which exploit high level semantic of the scene to obtain prior information on the optical flow. These methods classify the scene into different regions of similar motion and apply an optimized optical flow model to each region depending on motion characteristics. These models are known in literature as "layered models". A good example is Optical Flow with Semantic Segmentation and Localized layers (SOF), Sevilla-Lara *et. al.* [80]. The authors classify scenes into "Things" (rigid moving objects), Planes (planar background) and Stuff. A different model is then adapted for each of the three classes to refine DiscreteFlow, Sec. 3.1.2. Segmentation is performed with CNN by using DeepLab [15]. Focus is on the estimation of "Things" by applying layered optical flow only in the regions of interest ("Things" can be considered foreground). SOF is based on Sun *et. al.* [85] a primer for the idea of embedding semantic information into flow estimation. A similar idea

was proposed by Sai *et. al.* [92] using fully connected models for segmentation jointly with a variational approach for optical flow, however their evaluation is limited to frame interpolation for optical flow and the segmentation dataset is limited. Instead, Bai *et. al.* [3] exploit semantic information along with deep siamese networks to estimate matches and thus the optical flow. This method is neither fully end-to-end nor fully handcrafted, but uses siamese networks to perform the optical flow on the foreground, and uses a patch based epipolar method, Yamaguchi *et. al.* [105], to compute the optical flow on the background. In this way, the authors exploits siamese CNN for patch extraction and matching in areas with complex movement, where neighbouring frames are fed to each branch of a siamese network to extract features and the two siamese branches are then combined with a product layer to generate a matching score for each possible displacement. For the background the authors uses handcrafted methods for better performance on small and simple motion. This is an important contribution as integration of learnt functions along with handcrafted features allow the method to overcome the weaknesses of both traditional methods (complex movements) and DL based methods (small displacements). However, this method has been developed to work in the context of autonomous driving where the scene is typically composed of a static background and a small number of traffic participants which move "rigidly".

Finally, Behl *et. al.*[8] exploits the semantic cues and geometry to estimate the rigid motion between frames more robustly and leads to improved results compared to all baselines. CNNs are trained on a newly annotated dataset of stereo images and integrated into a CRF-based model for robust 3D scene flow estimation, this work obtains the lowest outlier percentage in KITTI2015 for non-occluded regions. Similarly to Bai *et. al.* [3], Wulff *et. al.* presented MR-Flow (Mostly Rigid-Flow) [101] which uses CNN to produce a semantic rigidity probability score across different regions also taking into account that some objects are more likely to move than others. This score is combined with additional motion cues to obtain an estimate of rigid and independently moving regions. A classical unconstrained flow method is used to produce a rough flow estimate. After that, the information on rigid structures and the initial optical flow are iterated and jointly optimized. Currently MR-Flow ranks first place in Sintel *clean pass* (see Sec. 4.1).

3.7 Multi-frame Methods

To the author knowledge multi-frame methods have been explored since 1991 with the work of Black and Andan [9]. Currently, multiframe methods are the top performing in the Sintel benchmark *final pass* and second best in Sintel *clean pass*. As already mentioned for the *final pass* case the methods are based on PWC-net and thus fully learnable: Neoral *et. al.* [62] and Ren *et. al.* [74].

Table 5 Overview of hybrid learning methods. $*$ =top among methods working with frame pairs.

PAPER YEAR REF.	CONTRIBUTION	BUILDS ON	LIMITATION
DCFlow 2017 [104]	CNN to produce cost function	EpicFlow	Long inference
PatchMatch 2017 [32]	Siamese Networks with new loss function	FlowFields	Long inference
SOF 2016 [80]	Semantic Segmentation. Different models for different layers.	DiscreteFlow, DenseFlow[84]	Not tested on Sintel
MR-Flow* 2017 [101]	CNN produce rigidty score. Iterative refinement	DiscreteFlow	Long inference
Guney *et. al.* 2016 [27]	Local and Context siamese networks	DiscreteFlow	Piece-wise training
Bai *et. al.* 2016 [3]	Siamese CNN. Exploit segmentation. Epipolar Flow	SOF, Epicflow for refinement.	Automotive domain
PatchBatch 2016 [23]	Siamese CNN	PatchMatch, EpicFlow	Long inference
Behl *et. al.* 2017 [8]	Stereo frames. CNNs, CRF	Tatarchenko *et. al.* [90]	Automotive domain
Maurer *et. al.* 2018 [54]	CNN trained during inference, multiframe, Best on SINTEL *clean*.	CPM [32], RichFlow	Long inference

Proflow, Maurer *et. al.* [54] obtains the second best score in Sintel *clean pass*. Proflow is based on Coarse-to-fine Patch Match (CPM) Hu *et. al.* [32], RichFlow and as additional refinement for the matches, Maurer *et. al.* [55]. Finally ProFlow uses a CNN trained online (during the estimation) on forward and backward flow to obtain a sparse to dense motion field. The model is learnt in-place and makes this method quite different from the others in this chapter.

4 Discussion

4.1 Flow Estimation Benchmarks and Performance Assessment

Optical flow evaluation is a difficult task for many reasons: optical flow information is difficult to obtain for real scenarios and artificial scenes might not be as challenging as natural videos. Optical flow benchmarks count just few datasets. The Middlebury benchmark [6] is composed of sequences partly made of smooth deformations, but also involving motion discontinuities and motion details. Some sequences are synthetic, and others were acquired in a controlled environment allowing to produce ground truth for real scenes. However the dataset is limited to few sequences and its challenges have almost been completely overcomed by modern methods [22]. For this reason, a new dataset has been generated: MPI-sintel evaluation benchmark [14]. Sintel is drawn from a short computer rendered movie, it counts around 1500 frames with optical flow groundtruth. Two thirds of the dataset is given for training and the rest is used for evaluation. Sintel is a challenging benchmark including fast motion, occlusions and non-rigid objects. There are in parallel more optical flow benchmarks that have been released to evaluate optical flow, KITTI 2012 [24] which consists of moving camera in static scenes, KITTI 2015 [59] extended to dynamic scenes, large motion, illumination changes and occlusions, and HD1K dataset [42], but they are tailored for the automotive domain. Thus, for applications not related to the automotive domain the most common benchmark is Sintel. Error measures such as photometric error on frame interpolation sequences can be misleading as not necessarily photometric error correspond to optical flow error (see Sec. 3.2.2). Moreover optical flow estimation face several different challenges: small displacement, large displacement, light change and occlusions [99]. To correctly assess performance all these factors must be taken into account, but this is hard to catch with a single metric. Thus, performance are measured using different metrics: (i) EPE all; (ii) EPE matched (EPE on non occluded regions); (iii) EPE unmatched; (iv) d0-10, d10-60, d60-140, which are average endpoint error in regions within the indicated displacement range taking only matched pixels into account; (v) s0-10, s10-40, s40+, which are average endpoint errors in regions moving within the specified speed range per frame. The overall ranking is a combination of the previous metrics, evaluated both for “clean” pass (no change in light) and “final” pass (change in light, strong atmospheric effects, motion blur, camera noise). [4]

4.2 Optical Flow Estimation Ranking

As explained above, it is very difficult to rank optical flow estimation methods because performace cannot be accurately assessed by a single metric or a single

[4] An in-depth explanation on how Sintel dataset was generated is given in [14] and [13].

scenario. For this reason we believe it is more important to cluster methods based on their application domain.

Nonetheless, in this fast changing research field it is more important to underline what are the characteristics that give substantial improvement. Currently, the best optical flow methods for Sintel *clean pass* are Deep Learning Hybrid Methods exploiting Domain Knowledge (see 3.6.2), the best method is MR-Flow, followed by hybrid multiframe ProFlow. Sun *et. al.* [87] conjecture that this is because they exploit traditional methods to refine motion boundaries which are perfectly aligned in the clean pass. Differently, on Sintel *final pass* the post processing effects cause severe problems to existing traditional methods. In this challenging situation Deep Learning coarse-to-fine methods (see Sec. 3.2.1) obtain the best performance. PWC-net is the top performing two frame optical flow method, outperformed only by multi-frame methods applying PWC-net to multiple frames. Moreover it is believed that the *final pass* is more challenging and realistic because it is corrupted by motion blur, atmospheric changes and noise. However, PWC-net as others coarse-to-fine methods may fail on small and rapidly moving objects, due to the coarse-to-fine refinement (see Sec. 3.1.2).

4.3 Biometrics Applications of Optical Flow

There are several applications of optical flow estimation in the biometrics field. In this section, we briefly discuss most notable applications.

In [82] optical flow estimation has been adopted for the Face Recognition task. The results were promising, e.g., for particular sub-task of distinguishing real people from their image. Human Pose Estimation and Tracking is another application where optical flow has been applied and has shown promising results [103]. Accordingly, it has been shown that the use of optical flow, based on pose propagation and similarity measurement, can result in substantially superior outcome compared to baselines. The task of single person pose estimation has been addressed by several researchers, where pose estimation accuracy has been enhanced by considering optical flow [70]. In a more recent work [20], the authors extended this research line and addressed a more difficult task, i.e., Multi-People Tracking (MPT).

Another example of application of optical flow in biometrics is Action Recognition. In a recent work [63], the authors have jointly estimated the optical flow while recognising actions with convolutional neural networks capturing both appearance and motion in a single mode. The result has shown that this model significantly improves Action Recognition accuracy in comparison to the baseline. In [53] the authors proposed a multi-task CNN model that receives (as input) a sequence of optical flow channels and uses them for computing several biometric features (such as identity, gender and age).

It is worth noting that, these were some examples of the applications of optical flow estimation in biometrics. While we mentioned some important works, however,

there could be further works coming up in recent time, using more advanced optical flow estimation techniques for the particular application in biometrics.

5 Conclusion

In this book chapter, we have provided a survey on the state-of-the-art in optical flow estimation with a focus on Deep Learning (DL) methods. We have conducted a comprehensive analysis and classified a wide range of techniques, along with an identified, descriptive and discriminative dimension, i.e., whether the techniques are based on DL, or they are traditional hand-crafted. We have reported the similarities and differences between DL and traditional methods. We believe that this systematic review on optical flow estimation can help to better understand and use the methods, hence providing a practical resource for the practitioners and researchers in the field of biometrics. In addition to that, we described and listed datasets for optical flow estimation, commonly employed by the research community, and also discussed some important issues that have to be considered for a proper evaluation procedure.

It is worth noting that optical flow estimation is a mature, while still growing, research field and can be seen as a multi-disciplinary area. This research area partially overlaps with a broad range of topics, such as Signal Processing, Computer Vision, and Machine Learning. Hence, our book chapter by no means can be all-inclusive, and indeed it focuses mainly on DL methods that are of practical importance for biometric research.

References

1. Aria Ahmadi and Ioannis Patras. Unsupervised convolutional neural networks for motion estimation. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 1629–1633. IEEE, 2016.
2. Stefano Alletto, Davide Abati, Simone Calderara, Rita Cucchiara, and Luca Rigazio. Transflow: Unsupervised motion flow by joint geometric and pixel-level estimation. *arXiv preprint arXiv:1706.00322*, 2017.
3. Min Bai, Wenjie Luo, Kaustav Kundu, and Raquel Urtasun. Exploiting semantic information and deep matching for optical flow. In *European Conference on Computer Vision*, pages 154–170. Springer, 2016.
4. Christian Bailer, Bertram Taetz, and Didier Stricker. Flow fields: Dense correspondence fields for highly accurate large displacement optical flow estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 4015–4023, 2015.
5. Christian Bailer, Kiran Varanasi, and Didier Stricker. Cnn-based patch matching for optical flow with thresholded hinge embedding loss. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, page 7, 2017.
6. Simon Baker, Daniel Scharstein, JP Lewis, Stefan Roth, Michael J Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *International Journal of Computer Vision*, 92(1):1–31, 2011.

7. Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics (ToG)*, 28(3):24, 2009.
8. Aseem Behl, Omid Hosseini Jafari, Siva Karthik Mustikovela, Hassan Abu Alhaija, Carsten Rother, and Andreas Geiger. Bounding boxes, segmentations and object coordinates: How important is recognition for 3d scene flow estimation in autonomous driving scenarios? In *International Conference on Computer Vision (ICCV)*, volume 6, 2017.
9. Michael J Black and Padmanabhan Anandan. Robust dynamic motion estimation over time. In *CVPR*, volume 91, pages 296–203, 1991.
10. Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. In *Advances in neural information processing systems*, pages 737–744, 1994.
11. Thomas Brox, Christoph Bregler, and Jitendra Malik. Large displacement optical flow. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 41–48. IEEE, 2009.
12. Thomas Brox and Jitendra Malik. Large displacement optical flow: descriptor matching in variational motion estimation. *IEEE transactions on pattern analysis and machine intelligence*, 33(3):500–513, 2011.
13. D Butler, Jonas Wulff, G Stanley, and M Black. Mpi-sintel optical flow benchmark: Supplemental material. In *MPI-IS-TR-006, MPI for Intelligent Systems (2012*. Citeseer, 2012.
14. Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *European Conference on Computer Vision*, pages 611–625. Springer, 2012.
15. Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.
16. Qifeng Chen and Vladlen Koltun. Full flow: Optical flow estimation by global optimization over regular grids. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4706–4714, 2016.
17. Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang. Segflow: Joint learning for video object segmentation and optical flow. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 686–695. IEEE, 2017.
18. Piotr Dollár and C Lawrence Zitnick. Structured forests for fast edge detection. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1841–1848. IEEE, 2013.
19. Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2758–2766, 2015.
20. Matteo Fabbri, Fabio Lanzi, Simone Calderara, Andrea Palazzi, Roberto Vezzani, and Rita Cucchiara. Learning to detect and track visible and occluded body joints in a virtual world. *arXiv preprint arXiv:1803.08319*, 2018.
21. Meiyuan Fang, Yanghao Li, Yuxing Han, and Jiangtao Wen. A deep convolutional network based supervised coarse-to-fine algorithm for optical flow measurement. In *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2018.
22. Denis Fortun, Patrick Bouthemy, and Charles Kervrann. Optical flow modeling and computation: a survey. *Computer Vision and Image Understanding*, 134:1–21, 2015.
23. David Gadot and Lior Wolf. Patchbatch: a batch augmented loss for optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4236–4245, 2016.
24. Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.
25. Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

26. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
27. Fatma Güney and Andreas Geiger. Deep discrete flow. In *Asian Conference on Computer Vision*, pages 207–224. Springer, 2016.
28. David Hafner, Oliver Demetz, and Joachim Weickert. Why is the census transform good for robust optic flow computation? In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 210–221. Springer, 2013.
29. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
30. Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.
31. Yinlin Hu, Yunsong Li, and Rui Song. Robust interpolation of correspondences for large displacement optical flow. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (to appear)*, 2017.
32. Yinlin Hu, Rui Song, and Yunsong Li. Efficient coarse-to-fine patchmatch for large displacement optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5704–5712, 2016.
33. Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, volume 1, page 3, 2017.
34. Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8981–8989, 2018.
35. E. Ilg, T. Saikia, M. Keuper, and T. Brox. Occlusions, motion and depth boundaries with a generic network for disparity, optical flow or scene flow estimation. In *European Conference on Computer Vision (ECCV)*, 2018.
36. Eddy Ilg, Ozgün Cicek, Silvio Galesso, Aaron Klein, Osama Makansi, Frank Hutter, and Thomas Brox. Uncertainty estimates and multi-hypotheses networks for optical flow. In *European Conference on Computer Vision (ECCV)*, 2018.
37. Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2017.
38. Eddy Ilg, Tonmoy Saikia, Margret Keuper, and Thomas Brox. Occlusions, motion and depth boundaries with a generic network for disparity, optical flow or scene flow estimation. In *European Conference on Computer Vision (ECCV)*, 2018.
39. J Yu Jason, Adam W Harley, and Konstantinos G Derpanis. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. arxiv preprint. *arXiv preprint arXiv:1608.05842*, 2016.
40. Margret Keuper, Bjoern Andres, and Thomas Brox. Motion trajectory segmentation via minimum cost multicuts. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3271–3279, 2015.
41. Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, volume 2, 2015.
42. Daniel Kondermann, Rahul Nair, Katrin Honauer, Karsten Krispin, Jonas Andrulis, Alexander Brock, Burkhard Gussefeld, Mohsen Rahimimoghaddam, Sabine Hofmann, Claus Brenner, et al. The hci benchmark suite: Stereo and flow ground truth with uncertainties for urban autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 19–28, 2016.
43. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
44. Wei-Sheng Lai, Jia-Bin Huang, and Ming-Hsuan Yang. Semi-supervised learning for optical flow with generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 354–364, 2017.

45. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.
46. Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
47. Ce Liu, Jenny Yuen, and Antonio Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE transactions on pattern analysis and machine intelligence*, 33(5):978–994, 2011.
48. Gucan Long, Laurent Kneip, Jose M Alvarez, Hongdong Li, Xiaohu Zhang, and Qifeng Yu. Learning image matching by simply watching video. In *European Conference on Computer Vision*, pages 434–450. Springer, 2016.
49. Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
50. David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.
51. Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *In IJCAI81*, pages 674–679, 1981.
52. Wenjie Luo, Alexander G Schwing, and Raquel Urtasun. Efficient deep learning for stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5695–5703, 2016.
53. Manuel J Marín-Jiménez, Francisco M Castro, Nicolás Guil, F de la Torre, and R Medina-Carnicer. Deep multi-task learning for gait-based biometrics. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 106–110. IEEE, 2017.
54. Daniel Maurer and Andrés Bruhn. Proflow: Learning to predict optical flow. *arXiv preprint arXiv:1806.00800*, 2018.
55. Daniel Maurer, Michael Stoll, and Andrés Bruhn. Order-adaptive and illumination-aware variational optical flow refinement. In *Proceedings of the British Machine Vision Conference*, pages 9–26, 2017.
56. Nikolaus Mayer, Eddy Ilg, Philipp Fischer, Caner Hazirbas, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. What makes good synthetic training data for learning disparity and optical flow estimation? *International Journal of Computer Vision*, pages 1–19, 2018.
57. Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
58. Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. *arXiv preprint arXiv:1711.07837*, 2017.
59. Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3061–3070, 2015.
60. Moritz Menze, Christian Heipke, and Andreas Geiger. Discrete optimization for optical flow. In *German Conference on Pattern Recognition*, pages 16–28. Springer, 2015.
61. Yana Mileva, Andrés Bruhn, and Joachim Weickert. Illumination-robust variational optical flow with photometric invariants. In *Joint Pattern Recognition Symposium*, pages 152–162. Springer, 2007.
62. Michal Neoral, Jan Šochman, and Jiří Matas. Continual occlusions and optical flow estimation. *arXiv preprint arXiv:1811.01602*, 2018.
63. Joe Yue-Hei Ng, Jonghyun Choi, Jan Neumann, and Larry S Davis. Actionflownet: Learning motion representation for action recognition. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1616–1624. IEEE, 2018.
64. Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. *arXiv preprint arXiv:1708.01692*, 2017.
65. Peter Ochs, Jitendra Malik, and Thomas Brox. Segmentation of moving objects by long term video analysis. *IEEE transactions on pattern analysis and machine intelligence*, 36(6):1187–1200, 2014.

66. Jiahao Pang, Wenxiu Sun, Jimmy SJ Ren, Chengxi Yang, and Qiong Yan. Cascade residual learning: A two-stage convolutional neural network for stereo matching. In *ICCV Workshops*, volume 7, 2017.
67. Nils Papenberg, Andrés Bruhn, Thomas Brox, Stephan Didas, and Joachim Weickert. Highly accurate optic flow computation with theoretically justified warping. *International Journal of Computer Vision*, 67(2):141–158, 2006.
68. Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 724–732, 2016.
69. Béatrice Pesquet-Popescu, Marco Cagnazzo, and Frédéric Dufaux. Motion estimation techniques. *TELECOM ParisTech*, 2016.
70. Tomas Pfister, James Charles, and Andrew Zisserman. Flowing convnets for human pose estimation in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1913–1921, 2015.
71. Anurag Ranjan and Michael J. Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
72. Anurag Ranjan, Varun Jampani, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Adversarial collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. *arXiv preprint arXiv:1805.09806*, 2018.
73. Zhe Ren, Junchi Yan, Bingbing Ni, Bin Liu, Xiaokang Yang, and Hongyuan Zha. Unsupervised deep learning for optical flow estimation. In *AAAI*, volume 3, page 7, 2017.
74. Zhile Ren, Orazio Gallo, Deqing Sun, Ming-Hsuan Yang, Erik B Sudderth, and Jan Kautz. A fusion approach for multi-frame optical flow estimation. *arXiv preprint arXiv:1810.10066*, 2018.
75. Jerome Revaud, Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. Epicflow: Edge-preserving interpolation of correspondences for optical flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1164–1172, 2015.
76. Jerome Revaud, Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. Deepmatching: Hierarchical deformable dense matching. *International Journal of Computer Vision*, 120(3):300–323, 2016.
77. Stephan R. Richter, Zeeshan Hayder, and Vladlen Koltun. Playing for benchmarks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2232–2241, 2017.
78. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
79. Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.
80. Laura Sevilla-Lara, Deqing Sun, Varun Jampani, and Michael J Black. Optical flow with semantic segmentation and localized layers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3889–3898, 2016.
81. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
82. Maciej Smiatacz. Liveness measurements using optical flow for biometric person authentication. *Metrology and Measurement Systems*, 19(2):257–268, 2012.
83. Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.
84. D. Sun, J. Wulff, E. B. Sudderth, H. Pfister, and M. J. Black. A fully-connected layered model of foreground and background flow. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2451–2458, June 2013.

85. Deqing Sun, Jonas Wulff, Erik B Sudderth, Hanspeter Pfister, and Michael J Black. A fully-connected layered model of foreground and background flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2451–2458, 2013.
86. Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. *arXiv preprint arXiv:1709.02371*, 2017. preprint, original paper is published on CVPR, June 2018.
87. Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Models matter, so does training: An empirical study of cnns for optical flow estimation. *arXiv preprint arXiv:1809.05571*, 2018.
88. Kalaivani Sundararajan and Damon L. Woodard. Deep learning for biometrics: A survey. *ACM Comput. Surv.*, 51(3):65:1–65:34, May 2018.
89. Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.
90. Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Multi-view 3d models from single images with a convolutional network. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 322–337, Cham, 2016. Springer International Publishing.
91. Engin Tola, Vincent Lepetit, and Pascal Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *IEEE transactions on pattern analysis and machine intelligence*, 32(5):815–830, 2010.
92. Yi-Hsuan Tsai, Ming-Hsuan Yang, and Michael J Black. Video segmentation via object flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3899–3908, 2016.
93. Zhigang Tu, Wei Xie, Dejun Zhang, Ronald Poppe, Remco C Veltkamp, Baoxin Li, and Junsong Yuan. A survey of variational and cnn-based optical flow techniques. *Signal Processing: Image Communication*, 72:9–24, 2019.
94. Sudheendra Vijayanarasimhan, Susanna Ricco, Cordelia Schmid, Rahul Sukthankar, and Katerina Fragkiadaki. Sfm-net: Learning of structure and motion from video. *arXiv preprint arXiv:1704.07804*, 2017.
95. Changsheng Wan, Li Wang, and Vir V Phoha. A survey on gait recognition. *ACM Computing Surveys (CSUR)*, 51(5):89, 2018.
96. Mei Wang and Weihong Deng. Deep face recognition: A survey. *arXiv preprint arXiv:1804.06655*, 2018.
97. Anne S Wannenwetsch, Margret Keuper, and Stefan Roth. Probflow: Joint optical flow and uncertainty estimation. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 1182–1191. IEEE, 2017.
98. Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, and Cordelia Schmid. Deepflow: Large displacement optical flow with deep matching. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1385–1392. IEEE, 2013.
99. J. Wulff, D. J. Butler, G. B. Stanley, and M. J. Black. Lessons and insights from creating a synthetic optical flow benchmark. In A. Fusiello et al. (Eds.), editor, *ECCV Workshop on Unsolved Problems in Optical Flow and Stereo Estimation*, Part II, LNCS 7584, pages 168–177. Springer-Verlag, October 2012.
100. Jonas Wulff and Michael J Black. Temporal interpolation as an unsupervised pretraining task for optical flow estimation. *arXiv preprint arXiv:1809.08317*, 2018.
101. Jonas Wulff, Laura Sevilla-Lara, and Michael J Black. Optical flow in mostly rigid scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, page 7. IEEE, 2017.
102. Xuezhi Xiang, Mingliang Zhai, Rongfang Zhang, Yulong Qiao, and Abdulmotaleb El Saddik. Deep optical flow supervised learning with prior assumptions. *IEEE Access*, 6:43222–43232, 2018.
103. Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. *arXiv preprint arXiv:1804.06208*, 2018.

104. Jia Xu, René Ranftl, and Vladlen Koltun. Accurate optical flow via direct cost volume processing. *arXiv preprint arXiv:1704.07325*, 2017.
105. Koichiro Yamaguchi, David McAllester, and Raquel Urtasun. Robust monocular epipolar flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1862–1869, 2013.
106. Koichiro Yamaguchi, David McAllester, and Raquel Urtasun. Efficient joint segmentation, occlusion labeling, stereo and flow estimation. In *European Conference on Computer Vision*, pages 756–771. Springer, 2014.
107. Guorun Yang, Zhidong Deng, Shiyao Wang, and Zeping Li. Masked label learning for optical flow regression. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 1139–1144. IEEE, 2018.
108. Yanchao Yang and Stefano Soatto. S2f: Slow-to-fast interpolator flow. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
109. Ilker Yildirim, Tejas D Kulkarni, Winrich A Freiwald, and Joshua B Tenenbaum. Efficient and robust analysis-by-synthesis in vision: A computational framework, behavioral tests, and modeling neuronal representations. In *Annual conference of the cognitive science society*, volume 1, 2015.
110. Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2018.
111. Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
112. Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.
113. Jure Zbontar and Yann LeCun. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17(1-32):2, 2016.
114. Yi Zhu, Zhenzhong Lan, Shawn Newsam, and Alexander G Hauptmann. Guided optical flow learning. *arXiv preprint arXiv:1702.02295*, 2017.
115. Yi Zhu and Shawn Newsam. Densenet for dense flow. In *Image Processing (ICIP), 2017 IEEE International Conference on*, pages 790–794. IEEE, 2017.
116. Henning Zimmer, Andrés Bruhn, Joachim Weickert, Levi Valgaerts, Agustín Salgado, Bodo Rosenhahn, and Hans-Peter Seidel. Complementary optic flow. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 207–220. Springer, 2009.