Analysis of Mainstream Models Based on Deep Learning

Zhien Lai

School of Computer, Xi'an Jiaotong University, Xi'an, China

Lze14alo@stu.xjtu.edu.cn

Abstract. Optical flow estimation is a hot topic that significantly impacts the development of the computer vision field, especially in pattern recognition. With advancements in artificial intelligence technologies, deep learning has seemingly opened up a new path for optical flow estimation. This paper focuses on four of the most influential deep learning-based models: the FlowNet series, the PWC-Net series, the RAFT series, and the SPyNet series. Each category includes representative models that showcase subsequent improvements. Firstly, this paper analyzes the neural convolutional network architecture and design of the basic models in each category. Following this, the optimization principles behind the improved versions of these models are explored. Finally, the strengths and weaknesses of these models by evaluating their Average Endpoint Error (AEE) results across various datasets are summarized. The paper concludes with an overall summary of the findings, providing insights into the progress and future directions of optical flow estimation in the context of deep learning.

Keywords: Optical flow, deep learning, artificial intelligence, convolutional neural network.

1. Introduction

Optical Flow estimation is a frequently discussed topic. Optical flow method is still widely used in computer vision, image processing and other fields, including motion detection, object cutting, collision time calculation, motion compensation coding and etc. However, it is not well known to the public because it is not easy to explicitly present in applications. As the computer vision community shifts from image understanding to video understanding, people pay more and more attention to the research and application of video. As a traditional Method to estimate the motion direction of objects in video, optical flow estimation is mainly based on brightness constancy and small motion, among which the most representative methods are Lucas-Kanade Method and Horn-Schunck Method respectively. Limited by the academic level and the development of hardware and software at that time, the efficiency, robustness and accuracy of these traditional methods are still lacking. First, its computational complexity is high, especially when dealing with large-scale image or video data, it may encounter computational efficiency problems. Second, the optical flow method is more sensitive to regions with discontinuous noise and motion, and the wrong estimation of these regions may affect the accuracy of the entire optical flow field.

Fortunately, with the progress of technology, the successful application of artificial intelligence technology in many fields such as computer vision has led many scholars to focus on deep learning and try to explore ideas based on deep learning. With the introduction of FlowNet in 2015, an algorithm that marks a major breakthrough in the field of optical flow estimation, many deep-learn-based optical

estimation algorithms have emerged successively. SPYNet (Spatial Pyramid) algorithm and PWC-Net algorithm. In addition, there are many valuable algorithms based on the above improved algorithms, such as LiteFlowNet algorithm, FlowNet2.0 algorithm and IRR-PWC algorithm. Many algorithms provide a variety of options for optical flow estimation. At present, there are not enough literature on the analysis of algorithms based on artificial intelligence, and the analysis of literature focusing on the advantages and disadvantages of various algorithms based on deep-learning is not thorough enough.

This paper summarizes the methods, advantages and disadvantages of existing mainstream algorithms based on deep-learning, analyzes the potential development direction in the future and the main problems urgently needed to be solved. After sorting out and summarizing relevant literature, this paper divides the solutions to optical flow estimation based on deep learning into several categories as shown in the figure 1.
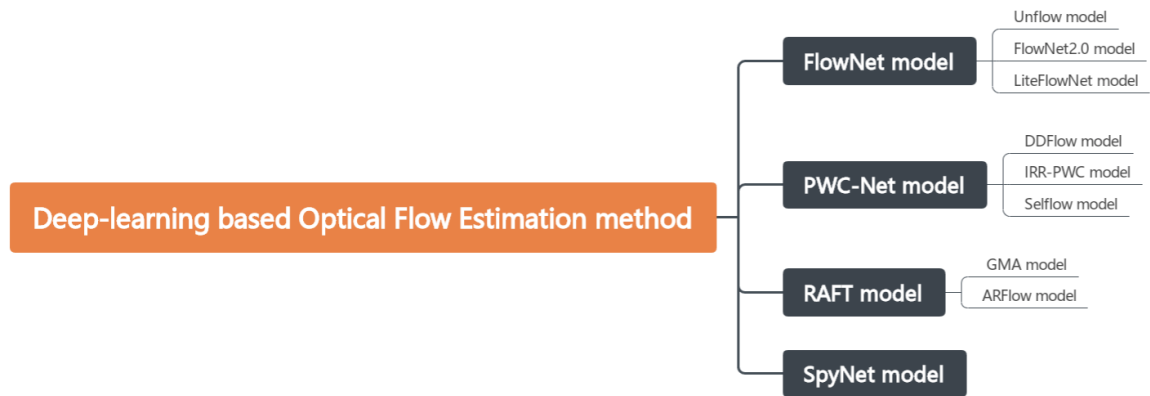

Figure 1. Classification of the mainstream optical flow estimation models based on deep learning (Photo/Picture credit : Original)

2. FlowNet model

2.1. The principle of FlowNet

FlowNet is a convolutional neural network (CNN) for optical flow estimation, proposed by Dosovitskiy et al. [1].The model is trained in an end-to-end manner as an alternative to traditional methods.Trained with a large amount of composite data such as Flying Chairs, the model was able to adjust network parameters through backpropagation and gradient descent, learning to extract motion information from image pairs. The FlowNet model refines layer by layer through layer by layer sampling and deconvolution, and finally obtains high-resolution optical flow prediction (Figure 2).

(a)
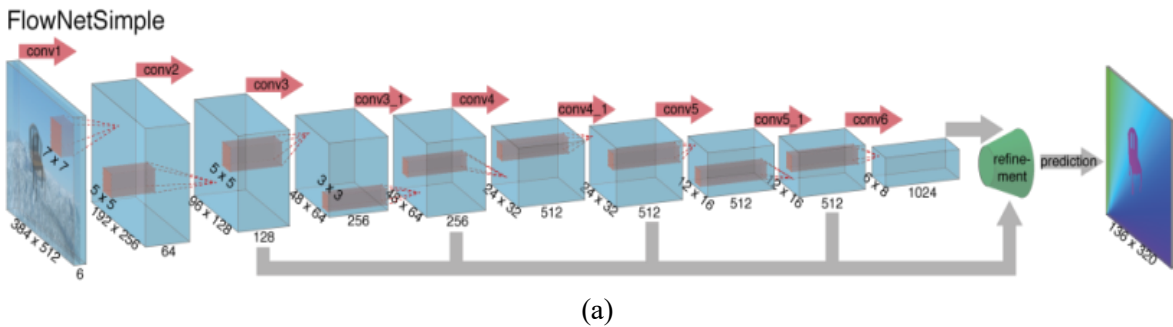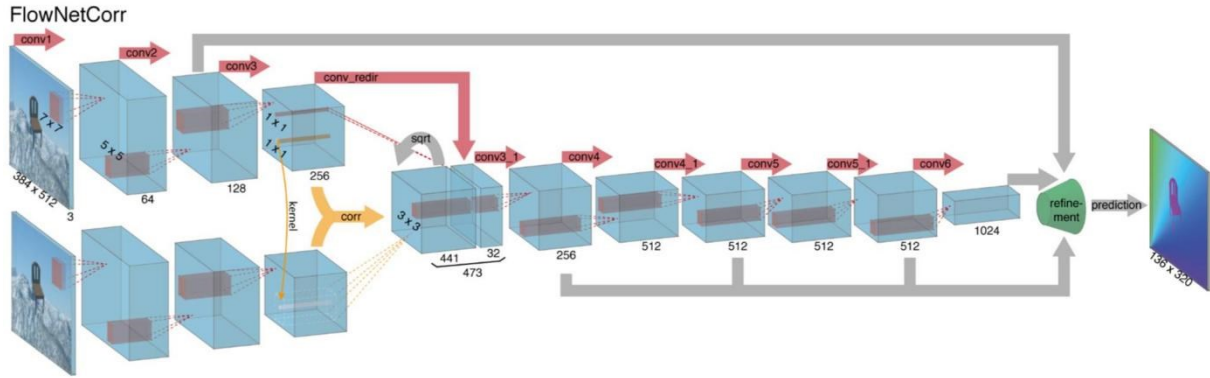
(b)

Figure 2. The two network architectures: FlowNetSimple (a) and FlowNetCorr (b) [1].

2.2. Improved model based on FlowNet

Given that FlowNet was originally an early model that was structurally easy to expand, the potential for subsequent improvements is huge. There are some successful improvement cases in changing supervision mode, convolutional neural network architecture and algorithm.

UnFlow is an unsupervised optical flow estimation model proposed by Meister et al in 2018 to solve the problem of dependence on large supervised data sets for optical flow estimation tasks [2]. The network architecture of UnFlow is largely inherited from FlowNet. Both of them adopt the architecture based on convolutional neural networks, and the processing methods are similar. UnFlow proposes to estimate optical flow through unsupervised learning, which is the key innovation, thereby reducing the reliance on synthetic data and enhancing its generalization ability in real-world scenarios.

Unflow's self-supervised learning pattern reduces its reliance on labeled data. However, it does not improve the estimation accuracy of FlowNet. FlowNet2.0 is also an improved version of FlowNet, which was basically based on structure of FlowNet, aiming to improve estimation accuracy [3]. Firstly, The most significant innovation in FlowNet2.0 is its stacked architecture, cascading multiple sub-networks. This design is to make each network refine the optical flow predictions produced by the previous network, leading to progressively more accurate estimations. Secondly, to adjust the optical flow estimation at different stages, FlowNet2.0 employed a variety of data augmentation techniques during training. These included random cropping, scaling, rotation, and color augmentations. By introducing these variations, the network was exposed to a broader range of motion patterns, lighting conditions, and object sizes during training. In addition, FlowNet2.0 proposes a residual learning mechanism to further improve the performance of the model by learning the difference between the current predicted result and the real optical flow.

LiteFlowNet, designed from the perspective of power consumption, which is different from the ideas of Unflow and FlowNet2.0, is an efficient optical flow estimation network designed with a focus on reducing computation while maintaining high accuracy, which gives it good applicability in resource-constrained environments [4].The design concept and architecture of LiteFlowNet is inspired by the FlowNet family, which uses a multi-level pyramid structure to gradually refine the optical flow estimation.When the optical flow is refined, the edge sensing module is added, which can consider the edge features of the image when estimating the optical flow, helping to reduce the error of the edge region.

3. PWC-Net model

3.1. The principle of PWC-Net model

PWC-Net is an excellent model that uses convolutional neural networks to extract features from cost bodies for optical flow estimation and balances efficiency and accuracy [5].The key design feature is to

construct a multi-level image pyramid to deal with optical flow estimation. The input image pair (I1, I2) is passed through a convolutional neural network to extract feature pyramids at multiple resolutions. Starting from the lowest resolution level, the estimated optical flow from the previous layer is used to warp the feature map of I2. This warping aligns the features of I2 with those of I1, helping to reduce the difference between the two and make the flow estimation more accurate. A cost volume is constructed by comparing the warped features of I2 with the features of I1. The cost volume represents the matching costs for different pixel displacements and serves as a key input for flow estimation. This is typically done by computing the correlation or inner product between feature vectors. At each pyramid level, a small convolutional neural network (Flow Estimation Network) uses the cost volume and contextual features to estimate the optical flow for that resolution. The flow is then refined at higher resolutions. This iterative process gradually improves the accuracy of the flow field, starting from a coarse flow estimate at low resolution to a finer estimate at high resolution. Figure 3 shows the PWC-Net warps features of the second image using the upsampled flow, computes a cost volume, and process the cost volume using CNNs. Both post-processing and context network are optional in each system. The arrows indicate the direction of flow estimation and pyramids are constructed in the opposite direction. Please refer to the text for details about the network
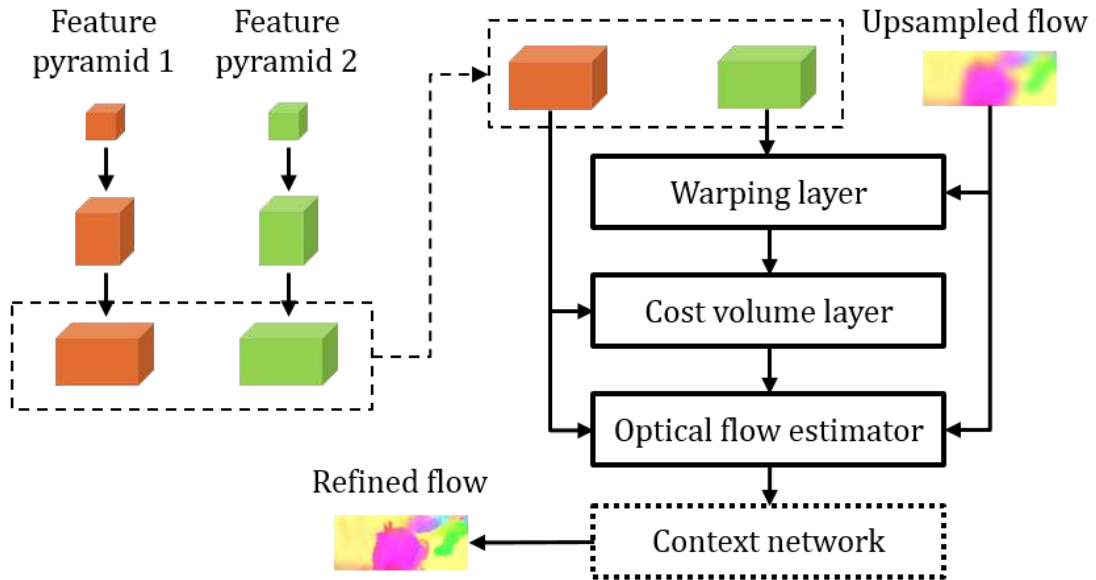
Figure 3. Feature pyramid and refinement at one pyramid level by PWC-Net. [5].

3.2. Improved models based on PWC-Net

PWC-Net is a milestone model in the field of optical flow estimation because it effectively introduces a multi-level pyramid structure on the basis of convolutional neural network. Therefore, in addition to the improvement ideas mentioned above in FlowNet, there are also some ideas aimed at improving the multi-level pyramid structure, especially for the local data dependency problem of PWC-Net.

The reliance on labeled data when training PWC-Net models is a noteworthy issue, and some models have improved the problem by changing the supervision mode. DDFlow and SelFlow are good examples, both of which learn through a self-supervised pattern. DDFlow is an self-supervised learning model for optical flow estimation [6]. Its network architecture is developed on the basis of PWC-Net, and an innovative double-ended supervision framework is proposed to effectively reduce the errors caused by occlusion and motion ambiguity in optical flow field. DDFlow algorithms also use pseudo-supervised data instead of real labels, avoiding the need to manually label data, and can be trained using large amounts of unlabeled data.

SelfFlow, of which its improvements are similar to DDFlow, is a self-supervised learning model, which relies on the geometric relationship between images for optical flow estimation [7]. By

constructing self-supervised loss function, dependence on real labeled optical flow is avoided. The Selflow model uses convolutional neural network to extract the feature representation of the image and feed it into the optical flow estimation network to predict the optical flow between the images.Then the error of image reconstruction is optimized. Finally, the consistency between forward optical flow and reverse optical flow is used to constrain optical flow prediction. A multi-scale mechanism is used to optimize optical flow prediction at different resolution levels.

The common problems with DDFlow, Selflow, and PWC-Net is that estimations are not accurate enough, even though we know Selflow is doing its best, especially when confronting occlusions and luminosity variations. The improvement of IRR-PWC is mainly regarding the accuracy, particularly in the case of more complex motion [8]. It shares similar designs in features extraction and cost volume construction, and introduces the idea of iterative residual optimization on this basis. A deep convolutional network is used to extract the multi-level pyramid features of the input image to improve the optical flow estimation. For each pyramid feature, the optical flow estimation module is used to roughly estimate and refine the optical flow field layer by layer. The algorithm introduces an iterative residual optimization module to gradually correct the previously estimated optical flow.Each iteration is optimized based on the error (residual) of the previous estimate, generating a new optical flow prediction.

4. Recurrent All-Pairs Field Transforms (RAFT) model

4.1. The principle of RAFT

RAFT model is a deep learning method for optical flow estimation that introduces a deep convolutional neural network [9]. RAFT uses a CNN encoder to extract features from two input frames, generating high-resolution feature maps. Each pixel corresponds to a high-dimensional feature vector, which will be used for subsequent optical flow estimation. The model computes the similarity between all feature points from both frames, constructing a 4D cost volume that records the matching cost for every pixel with all possible pixels in the other frame. This ensures that the model can identify the best matches between the two images. Then an initial rough optical flow field is generated, typically initialized to zero optical flow, assuming no motion between the frames at the start. The Recurrent Update Module (using GRU) iteratively refines the optical flow estimate. Each iteration combines the current optical flow estimation with the cost volume and feature maps, progressively correcting the flow field. This process is repeated multiple times, gradually approaching the true optical flow. Before final output, the optical flow field undergoes bilinear interpolation for fine-tuning, capturing subtle motion changes and enhancing the overall precision (Figure 4).
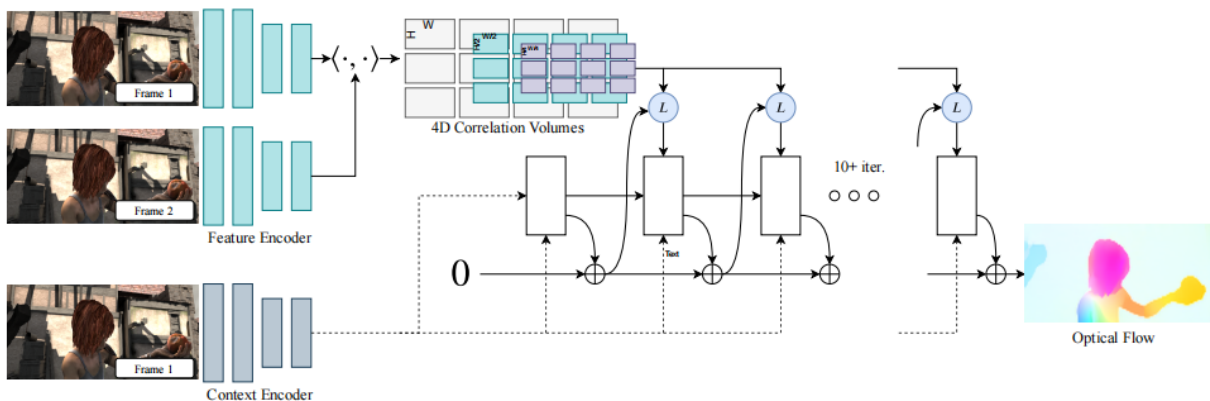
Figure 4. RAFT model [9]

4.2. Improved models based on RAFT

The RAFT model is known for its exaggerated estimation accuracy but this does not mean that it has no potential for improvement as its computing cost and dependence on local correlation features remain problematic. The following two models effectively improve RAFT from these two aspects.

GMA reduces the dependence on local pixel correlation through Global Motion Aggregation, which is an important improvement over RAFT model [10]. The core idea is to introduce Global Motion Aggregation on the basis of RAFT model to capture global motion information. GMA calculates the global similarity between each pixel and all other pixels in the image, forming a global attention matrix, allowing the model to capture both local and global motion information. By aggregating motion cues from across the entire image, GMA can better estimate optical flow in challenging scenarios, such as large object motions or complex motion patterns.

ARFlow's approach to improvement is slightly different from GMA's. Augmented Reality Flow (ARFlow) is a model for optical flow estimation [11]. Data enhancement and regularization technologies are added to RAFT model, and self-supervised learning and robustness improvements are introduced. ARFlow introduces data augmentation techniques during training, such as spatial transformations (scaling, cropping, and rotation) applied to the input images. This enables the model to learn motion consistency across varying perspectives and transformations. Another key innovation in ARFlow is the introduction of a cycle consistency loss, which ensures that the optical flow estimation remains consistent when computing forward and backward flow.

5. SPyNet model

SPyNet is a deep learning model for optical flow estimation, proposed in 2017, that uses a pyramid-like architecture for step-by-step optical flow estimation [12]. Although SPyNet itself was a relatively early model for optical flow estimation, its design philosophy laid the foundation for a series of subsequent improved models, with PWC-Net and LiteFlowNet in particular making significant progress on its foundation.SPyNet is also widely used in more complex models, such as some of the design inspiration for RAFT, reflecting the importance of SPyNet in the field of optical flow estimation. So it is analyzed as a separate part.

The core of SPyNet's design is to estimate optical flow step by step through an image pyramid. This method means that the input image pairs are progressively narrowed down to pyramid levels of different resolutions, each level refining the lower resolution optical flow estimates. The input image is first downsampled through the Gaussian pyramid, and each scale estimates the optical flow independently. Second, at the bottom of the pyramid, the network estimates optical flow from the smallest resolution. This estimate is then upsampled to a higher resolution and further refined at a higher level. This process is carried out layer by layer until the original resolution (Figure 5). A relatively small CNN was then used to estimate the optical flow, which has a smaller model scale compared to other more complex optical flow models such as FlowNet and RAFT.
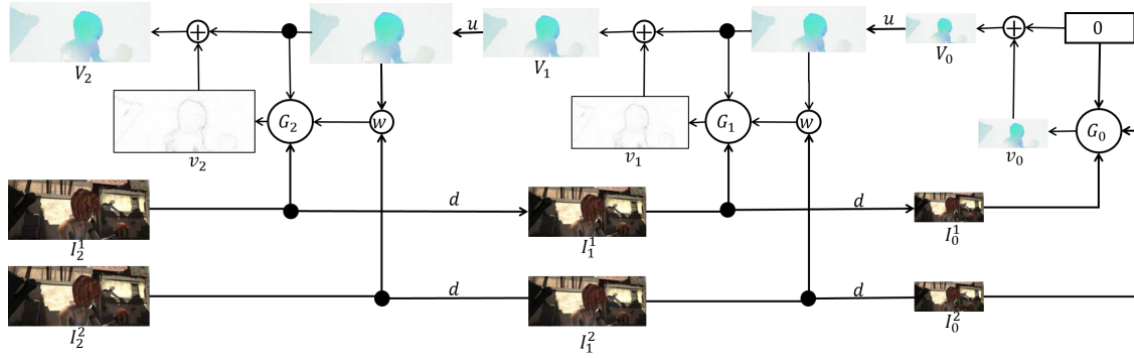
Figure 5. Inference in a 3-Level Pyramid Network: The network G0 computes the residual flow v0 at the highest level of the pyramid (smallest image) using the low resolution images {I01,I02}. At each pyramid level, the network Gk computes a residual flow vk which propagates to each of the next lower levels of the pyramid in turn to finally obtain the flow V2 at the highest resolution [12].

6. Analysis of advantages and disadvantages of various models

The advantages and disadvantages of various optical flow estimation models mentioned above are analyzed and summarized, which mainly covers the evaluation of computing speed, computing cost, efficiency, estimation accuracy and complexity of architecture (Table 1).

Table 1. Advantages and disadvantages of various optical flow estimation models

Category	Model	Advantages	Disadvantages
FlowNet-based	FlowNet	1. Fast speed, real-time estimation of optical flow can be achieved 2. Relatively simple structure which is easy to do expansion	1. The accuracy is relatively low. It's difficult to deal with complex, occluded and small movements 2. Poor performance at dealing with big motion
	UnFlow	1. A wide range of application scenarios	1. Poor performance when processing complex scenes or fast-moving objects 2. Not robust enough when dealing with noise, luminosity variations or dynamic objects
	LiteFlowNet	1. More efficient and being more suitable for devices with limited resources 2. Higher accuracy in processing small objects and complex motion scenes	1. Its structure is relatively complex, and the process of model training and parameter adjustment is relatively complex 2. Poor performance when dealing large motions
	FlowNet2.0	1. High accuracy 2. Compared with the original FlowNet, it can handle complex motion scenes better, especially in high resolution scenes and detail processing	1. More stringent requirements on computing costs and memory usage 2. The running time is long, and it is difficult to achieve real-time processing
PWC-Net-based	PWC-Net	1. Able to handle large range of motion effectively 2. Able to do the estimation in local area and avoid the high calculation cost caused by global calculation	1. Weak perception of global motion and poor performance in scenes with long distance movements or large parallax changes 2. Poor performance in motion estimation of small objects 3. Unable to deal stably with occlusion and changes in lighting
	DDFlow	1. There is no need for large-scale labeled data sets, which reduces the dependence on labeled data and makes it more widely applicable 2. Computationally efficient, suitable for processing large data sets or devices with limited resources	1.Relatively low accuracy 2. Be prone to bias when dealing with luminosity variation, noise or irregular texture.

Table 1. (continued).

	IRR-PWC	1. Better performance in more complex scenes	1. The introduced iterative update mechanism increases the calculation cost
	Selflow	1. Good performances on large-scale data sets, ADAPTS to real-world applications without labels, and reduces training costs 2. Robustness improved for occlusion and dynamic scenes	1. Accuracy is lower than most supervised learning models, and estimation accuracy is lower than IRR-PWC and FlowNet2.0 (even with improvements) when dealing with fast motion, complex lighting and occlusion areas.
RAFT-based	RAFT	1. Able to accurately capture subtle changes in the optical flow field 2. All pixel pairs in the scene can be comprehensively compared, thus enhancing the global perception of the model and performing well in precision.	1. Large computing costs and memory requirements, which makes RAFT unsuitable for resource-constrained devices or real-time application scenarios. 2. The number of iteration updates of RAFT is fixed and cannot be dynamically adjusted according to specific scenarios. In complex motion scenarios, the fixed number of iterative steps may not be enough to ensure the convergence of the model, thus affecting the accuracy of optical flow estimation.
	GMA	1. Able to capture the large range and complex movement in the scene 2. Able to deal with the estimation of optical flow in the blocked area and avoid the estimation error caused by local motion information.	1.High computing costs and memory requirements. 2.More complex structure than RAFT , which increases the complexity of model design and the difficulty of training.
	ARFlow	1. Little reliance on large-scale labeled data sets, being suitable for data-scarce scenarios, low training costs. 2. Improved accuracy and robustness of the model in dealing with complex motion scenes.	1. Compared with supervised models of the same period, the estimation accuracy is still insufficient 2. When the illumination conditions change significantly, the accuracy of the optical flow estimation may decrease
	SPyNet	1. Making estimation of motions at different scales in a better efficiency 2. The network structure is relatively simple, the number of parameters is small, so the reasoning speed is fast, the computing resource requirements are low, and it is more suitable for running on resource-limited devices.	1. The handling of very large displacement scenarios is poor. Its pyramid level is limited and cannot effectively capture a large range of motion details. 2. Unable to effectively capture fine motion and high-frequency details in complex scenes, resulting in poor performance in complex scenes. 3. It is easy to make errors in the estimation of optical flow for complex illumination changes or non-rigid object motion. 4. If the estimate at the lower resolution layer is not accurate enough, the estimate at the higher resolution level will be affected, resulting in error accumulation

6.1. Analysis of results

The estimation results of various large models of optical flow estimation mentioned above, including AEE (Average endpoint error) under Sintel Final datasets and Fl-all under KITTI 15 datasets. The

smaller the values of AEE and Fl-all, the more accurate the estimation of the model '-' indicates that the model has not been tested on this datasets (Table 2).

Table 2. The estimation results of various large models

Category	Model	Sintel Final AEE	KITTI 15 Fl-all
FlowNet-based	FlowNetS	7.22	-
	FlowNetC	7.88	-
	UnFlow (C)	10.22	11.11%
	LiteFlowNet	6.09	10.24%
	FlowNet2.0	5.74	11.48%
PWC-Net-based	PWC-Net	5.04	9.60%
	DDFlow	7.40	14.29%
	IRR-PWC	4.58	7.65%
	Selflow	6.57	14.19%
RAFT-based	RAFT	2.86	5.27%
	GMA	2.47	4.93%
	ARFlow	5.67	11.79%
SPyNet	SPyNet	8.43	-

The first and fourth categories of models are relatively early, aiming to demonstrate the feasibility of using convolutional neural networks for optical flow estimation. Although their average AEEs on the Sintel Final dataset are higher (7.43 and 8.43 respectively), they laid a foundation for subsequent research. Improvements to FlowNet include UnFlow (unsupervised learning), FlowNet2.0 (improved accuracy to 5.74 AEE), and LiteFlowNet (balancing computational cost and accuracy). SPyNet, as an early lightweight model, focuses on low computational cost but has a higher AEE.

The second category of models uses feature pyramid structures and cost volume to combine sparse and multi-scale representations. PWC-Net demonstrates a good balance between accuracy and efficiency (Sintel: 5.04 AEE, KITTI 15: 9.60% Fl-all). IRR-PWC further improves accuracy through a recursive residual module (Sintel: 4.58 AEE, KITTI 15: 7.65% Fl-all). DDFlow and Selflow use self-supervised learning to reduce dependency on labeled data, but with lower accuracy (Sintel: 7.40/6.57 AEE, KITTI 15: 14.29%/14.19% Fl-all).

The third category represents the latest advancements, with RAFT setting a high-accuracy benchmark (Sintel: 2.86 AEE, KITTI 15: 5.27% Fl-all). GMA improves large motion handling by incorporating a global motion aggregation module (Sintel: 2.47 AEE, KITTI 15: 4.93% Fl-all). ARFlow attempts to integrate unsupervised learning, though its accuracy is slightly lower than RAFT and GMA, it still outperforms previous unsupervised and some supervised models.

Differences in AEE results across models reflect trade-offs between accuracy, computational efficiency, and application scenarios. Modern models like RAFT and GMA lead in accuracy but require more computational resources, while SPyNet and LiteFlowNet are better suited for scenarios with limited resources or real-time computation needs.

7. Conclusion

This paper discusses many models based on deep learning, and summarizes the advantages and disadvantages of them by analyzing their architectures, design concepts and AEE results on various datasets. Nowadays, various deep -learning-based models have greatly promoted the development of optical flow estimation. Due to the high efficiency of deep learning, although the accuracy of each model is slightly different, it generally performs very well, and the mainstream models with outstanding accuracy are RAFT and FlowNet2.0. On this basis, there are some models sacrifice some accuracy in exchange for the reduction of resource requirements, striking a equilibrium between performance and

efficiency, such as LiteFlowNet, DDFlow, SPyNet, etc. However, it is clear that the optical flow estimation field still has great prospects. First of all, some models still rely on the assumption of luminosity consistency, local feature extraction, etc., leading to poor performance in the face of common issues about occlusion, big motion, noise, and complex luminosity variations. Secondly, for unsupervised or self-supervised models, they cannot enjoy low reliance on labeled data to make them more widely applicable while maintaining high accuracy of estimation. Selflow inspired us to introduce some loss functions to improve accuracy.Therefore, it is hoped that future researchers can further improve the underlying algorithm of optical flow estimation models to deal with complex situations in the real world, and continue to study how to improve the estimation accuracy of unsupervised models as much as possible

References

[1] Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Brox, T. (2015). Flownet: Learning optical flow with convolutional networks. *In Proceedings of the IEEE international conference on computer vision* (pp. 2758-2766).

[2] Meister, S., Hur, J., & Roth, S. (2018, April). Unflow: Unsupervised learning of optical flow with a bidirectional census loss. *In Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1).

[3] Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., & Brox, T. (2017). Flownet 2.0: Evolution of optical flow estimation with deep networks. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2462-2470).

[4] Hui, T. W., Tang, X., & Loy, C. C. (2018). Liteflownet: A lightweight convolutional neural network for optical flow estimation. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8981-8989).

[5] Sun, D., Yang, X., Liu, M. Y., Kautz, J. (2018). Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8934-8943).

[6] Liu, P., King, I., Lyu, M. R., Xu, J. (2019, July). Ddflow: Learning optical flow with unlabeled data distillation. *In Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 8770-8777).

[7] Hur, J., Roth, S. (2019). Iterative residual refinement for joint optical flow and occlusion estimation. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5754-5763).

[8] Liu, P., Lyu, M., King, I., Xu, J. (2019). Selflow: Self-supervised learning of optical flow. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4571-4580).

[9] Teed, Z., Deng, J. (2020). Raft: Recurrent all-pairs field transforms for optical flow. *In Computer Vision–ECCV 2020: 16th European Conference*, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16 (pp. 402-419). Springer International Publishing.

[10] Jiang, S., Campbell, D., Lu, Y., Li, H., & Hartley, R. (2021). Learning to estimate hidden motions with global motion aggregation. *In Proceedings of the IEEE/CVF international conference on computer vision* (pp. 9772-9781).

[11] Liu, L., Zhang, J., He, R., Liu, Y., Wang, Y., Tai, Y., Huang, F. (2020). Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6489-6498).

[12] Ranjan, A., & Black, M. J. (2017). Optical flow estimation using a spatial pyramid network. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4161-4170).