NETWORK-BASED STRUCTURE OPTICAL FLOW ESTIMATION

Shu Liu

A thesis submitted for the degree of Bachelor of Engineering with Honours

The Australian National University

October 2019

Declaration

I declare that this thesis is my own original work except where otherwise indicated.

Shu Liu

24 October 219

Acknowledgements

I would like to thank my supervisor Nick Barnes for the opportunity of completing this thesis under his supervision. None of my work would have been possible without the suggestions from him. I also would like to thank Robert Mahony for holding technical meetings.

Abstract

Scene flow used to be the mainstream representation for the three-dimensional motion as it provides intuitive three-dimensional visualization of motion. Meanwhile, since the motion that is perpendicular to the image plane cannot be correctly estimated from the image plane, the accurate estimation of scene flow cannot be realized with only the monocular images. It leads to the difficulty of estimating three-dimensional motion with monocular cameras, while structure optical flow is one of the solutions to this problem.

This thesis addresses the problem of estimating structure optical flow with convolutional neural network. The contribution is threefold. First, it introduces a novel network-based method for estimating general three-dimensional motion field based on the architecture of spatial pyramid structure. Warping operation is applied in the method for bridging the motion field with an evaluable loss function. Such method acquires a comparable performance on scene flow estimation in Monkaa benchmark.

The second contribution is the investigation of the feasibility of current optical flow datasets for training a supervised structure estimator. The imbalanced proportion of motion of datasets becomes the main obstruct for estimation. At the same time, the consistency between colour and semantics also misleads the estimation result.

The final contribution is the visualization of the three-dimensional motion with a mesh plot in Cartesian coordinates. Differ from the traditional visualization that presenting optical flow and depth change independently, the mesh plot intuitively illustrates the motion in three axes by one image.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 INTRODUCTION OF THIS THESIS

Vision plays an important role in enabling an independent agent to capture the information from the surrounding scene. The static vision enables an agent to acquire a spatial perception in virtual environment, while the dynamic vision can capture the change of environment. In real life, the dynamic vision is usually related to capturing motion and it provides a continuity to the perception and helps the agent to take advantageous operation and avoid harm. In the field of robotics and computer vision, the topic of motion recognition focuses on the mechanism of dynamic vision, and it is widely used in industrial applications, such as autonomous driving.

The study of motion starts from optical flow that it known as a two-dimensional motion in image plane. Scene flow follows this technique and explores the representation of the three-dimensional motion, and the three-dimensional information is necessary for precisely estimating scene flow. Based on that, although scene flow is able to provide a demonstration of motion including both optical flow and depth change, the estimation of scene flow with a monocular camera is still ill-posed. However, structure optical flow, the scaled scene flow can be considered as a more intuitive representation of three-dimensional motion which only requires two-dimensional features for producing an accurate estimation.

In the early stage, the optical flow is used to rely on the closed-form solution, which is not robust when dealing with arbitrary motions. The development of machine learning provides a robust and accurate estimation of optical flow, which is adapted to estimating scene flow afterwards. Convolutional neural network can be considered as a new technique for computer vision. It is based on computer vision, while it can extract more complex features from images with back-propagation algorithm and convolution operation. The recent application in convolution neural network have acquired outstanding performance on motion field estimation.

In this document, we propose (1) the first deep convolutional neural network approach to estimation structure optical flow (that is called **structure flow** in following paragraphs). We also investigate (2) the feasibility of current public optical

flow datasets for training a supervised structure flow estimator in Chapter 3. Besides, we introduce (3) a new visualization for the three-dimensional motion field.

For acquiring a convincing result, we compare the estimation of scene flow with that of structure optical flow. Then, we review the existing techniques and have discussion on the result in Chapter 4. Finally, Chapter 5 includes the conclusion and some potential future works.

1.2 BACKGROUND AND MOTIVATION FOR STRUCTURE OPTICAL FLOW

In real-life, humans are able to determine the motion of surrounding objects and naturally avoid damage with one eye, while such ability does not only relies on the vision, but also requires the prior knowledge of the environment they are staying at.

In this monocular case, scene flow estimation is essentially an ill-posed problem with input monocular images as input, as the depth change cannot be apparently observed [1]. We demonstrate this problem by considering the projection of a three-dimensional motion on an image plane and present the relationship among optical flow, scene flow and structure flow.

The point X represents an independent object in a three-dimensional Cartesian coordinates. With a camera centre $\{\mathbf{C}\}$, its shape can be projected on to the image plane spanned by $\{u, v\}$, which is denoted by x. In this case, we take l_x and λ_x to represent the distance between $\{\boldsymbol{C}\}$ and define the projected shape x, the distance between the projected shape x and the object X respectively.

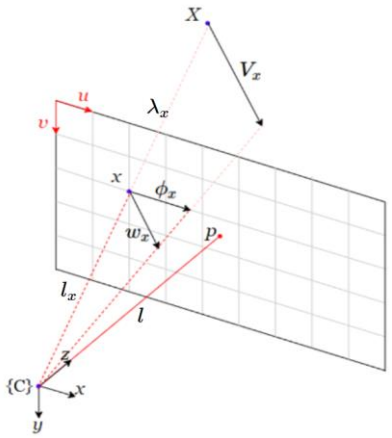


Figure 1: The motion of independent point X can be expressed as, optical flow ϕ_x, scene flow V_x or structure flow w_x

For different representations of motions, optical flow represents the apparent two-dimensional motion field in the image plane, describing by ϕ_x in Figure 1. Since optical flow cannot demonstrate the change of motion in the direction that is orthogonal to the image plane, scene flow, is then proposed for describing the three-dimensional motion with respect to the camera. In Figure 1, the scene flow of point X is same as its motion, which can be expressed by V_x. As for structure flow, it can be considered as the three-dimensional projected scene flow onto the image plane. With the notation of l_x and λ_x, structure flow can be expressed as

$$w_x = \frac{l_x}{l_x+\lambda_x} V_x.$$

Scene flow provides a representation for three-dimensional motion, but when we inversely estimate the scene flow with its projection on the image plane, the magnitude of scene flow would be ambiguous. Considering a new scene with two independent objects A and B in Figure 2, object B is larger and more distant to the image plane compared with object A. In this case, these two objects have two projections a and b, which have identical sizes on the image plane.

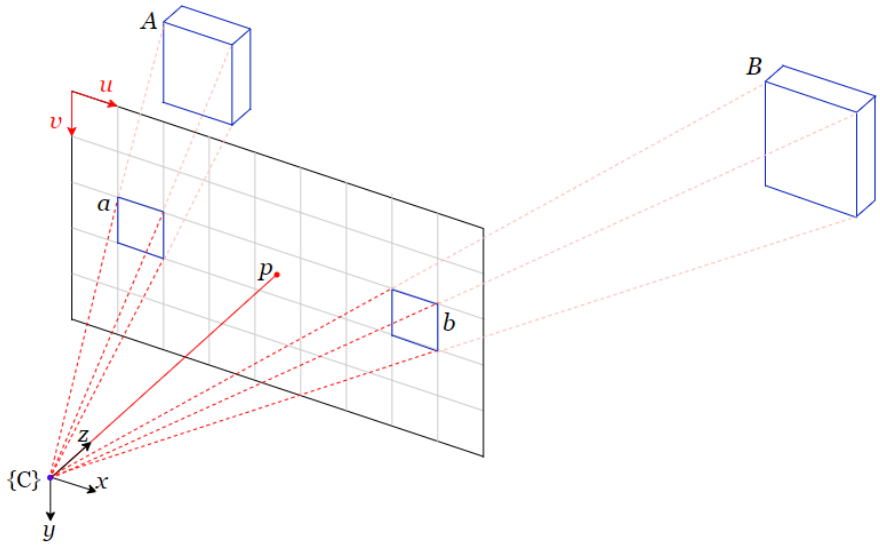


Figure 2: Independent object A and B and their projection a and b in the image plane.

In Figure 3, assuming object A and B are moving to A' and B', their projections a and b have a mirror motion that move to a' and b' respectively. If we observe the central point $\bar{A}, \bar{B}$ of the two objects, it is noted that the actual motion (scene flow) $V_{\bar{A}}$ and $V_{\bar{B}}$ cannot be reflected with the pattern on the image plane.

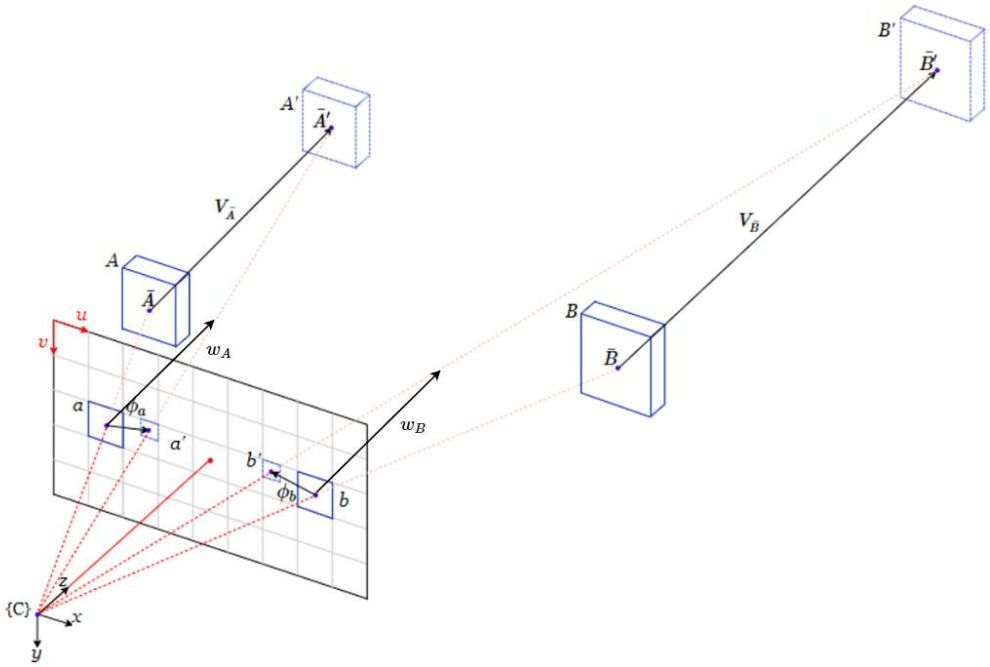


Figure 3: Before moving, objects A and B have projections a and b. After moving, A' and B' have projections a' and b'.

This example demonstrates the essential problem of scene flow estimation. Under such circumstance, structure flow becomes a solution as it is insensitive to unobservable depth change. In Figure 3, the structure flow of A and B are presented by w_A and w_B respectively. Structure flow w_A and w_B have identical shapes, and match the observation of the motion pattern on image plane.

In this case, we consider structure flow as a more practical representation of the three-dimensional motion for estimation based on monocular images compared with scene flow. And in this thesis, we present the details of discovering a new method for obtaining an accurate estimation of structure optical flow.

Chapter 2: Literature Review

This chapter begins with a background of different representations of motion fields including that of both two dimensions and three dimensions (Section 2.1). The application of network-based motion estimation and the mainstream benchmark datasets are reviewed in the Section 2.2 and Section 2.3 respectively. To this end, Section 2.4 highlights the summary and the implications based on the review.

2.1 REPRESENTATION OF VISUAL MOTION FIELD

Visual motion fields were introduced to describe the apparent motion in a visual environment [2]. Since it consists of the dynamical spatial information, which then became the main representation of motion and was widely used in the field of computer vision, such as face recognition and behaviour recognition [3] [4]. The idea of motion field was also exploited in robot control and helped develop autonomous driving in recent years [5] [6].

Meanwhile, there are quite a few challenges of estimating motion field. One critical challenge comes from the texture-less objects in the scene. Without a distinguishable texture, the identification of motion in a pixel-wise level becomes inaccurate [7]. Another challenge is caused by the variation of motion, especially the long-range motion. Due to hardware limitations, a small video sampling rate results in the long-range motion between continuous image frames [8]. Such long-range motion that lacks spatial consistency is hard to estimate. At the same time, some external factors, such as the lighting and the shadow change or the camera instability, introduce extra variable to motion estimation [9].

2.1.1 Two-dimensional motion field

Two-dimensional motion field is often referred to optical flow, which calculates two-dimensional pixel-wise motions between two consecutive frames, and outputs the location change of each pixel in x and y axis.

The method of partial derivatives was widely accepted for estimating the optical flow. Lucas-Kanade (LK) method assumes the magnitude of optical flow is locally related, which provides a local optimization method for resolving the equation of flow [10].

LK method was then followed by many researchers, such as Gang et al., who applied SVD method for improving the performance of detecting fast long range motion based on LK method [11]. At the same time, Horn-Schunck (HS) method discovered the field of estimating optical flow with global method [12].

Discrete optimization methods, including linear programming and belief propagation, considered the assignment of each pixel in optical flow map and tried to minimize the distance between the estimated result and the ground truth [13, 14, 15]. This solution has become the mainstream with the development of convolution neural network [16].

2.1.2 Three-dimensional motion field

Like two-dimensional optical flow, three-dimensional scene flow is also defined for each pixel in a reference image. However, three-dimensional scene flow consists of not only pixel-wise x-axis and y-axis motion parallel to the image plane, but also the z-axis motion which is perpendicular to the image plane [17].

Vedula et al. [18] formulated the solution of generating scene flow without assuming a rigidity of the observed scene. They applied the traditional two-dimensional optical flow to estimate scene flow. Meanwhile, since the three-dimensional motion cannot be fully reflected from the two-dimensional image, Patras et al. combined optical flow and the change of disparity for representing three-dimensional motion. Such idea is followed by Zhang et al. [19], who utilized multiple-view images and a global smoothness constraint to estimate the scene flow. The succeeding works in Gong et al. estimated disparity flow with stereoscopic images [20].

As for structure flow, it was developed for detecting road condition and was described as the scene flow scaled by the inverse depth in Adarve's work [21].

2.2 NETWORK-BASED FLOW ESTIMATION

Deep convolutional neural networks (DCNN) were firstly applied for estimating two-dimensional optical flow by Dosovitskiy et al. [16]. They proposed FlowNet that employed the hourglass-style architecture with encoder and decoder modules. Many empirical experiments were conducted based on the design of FlowNet, which resulted in improved architecture designs and more robust performance against occlusion [22, 23, 24]. SpyNet developed FlowNet and applied the idea of spatial pyramid [25], which explored a new network architecture for flow estimation. Some traditional

computer vision techniques, such as warping and cost volume, were combined with the network-based method by Sun et al., which is the state-of-the-art [26].

The estimation of three-dimensional scene flow with a network-based method was explored by Mayer et al. [28] They followed the work of Dosovirskiy et al. [16] and predicted a dense disparity map with stereo continuous frames. The succeeding work fused the independently estimated depth map and optical flow map for generating dense three-dimensional scene flow [27]. SF-Net made the trial to estimate optical flow with RGBD images [29].

2.3 DATASET FOR MOTION FLOW

The Middlebury dataset consists of RGB image pairs and the corresponding optical flow map that is calculated with high-resolution UV images. It is widely used as the benchmark for evaluating the performance of optical flow estimation [30]. The Kitti dataset focused on the application of driving, which was collected with an autonomous driving platform. It is widely accepted as it provides a large amount of recordings on both optical flow and depth measurements [5, 31, 32].

The development of virtual dataset started from McCane's work that applied synthetic sequences with optical flow and provided a metric with angular error [33]. However, the early-stage synthetic dataset is not realistic. MPI-Sintel was introduced for simulating semantically similar spatial features with the real-life scene. It was then considered as the predominating benchmark for three-dimensional motion field estimation [34]. Mayer's work produced three datasets, Monkaa, Flying chair and Driving, which have larger amount of data compared with previous datasets [28]. In addition, these three datasets contain accurate motion measurement in all three axes.

2.4 SUMMARY AND IMPLICATIONS

In general, two-dimensional optical flow and three-dimensional scene flow have been developed to present accurate visual motion fields. The estimation of these flow map also requires two-dimensional and three-dimension information respectively.

Since depth change in actual scale cannot be explicitly extracted from the monocular images. For scene flow estimation, many works tended to exploit extra information such as stereoscopic images for estimating the scene change that is perpendicular to image plane.

The structure optical flow has the same magnitude with the scaled scene flow, which is insensitive to the different scale between optical flow and depth change. In this case, it is feasible to acquire an accurate estimation of structure flow with pure monocular images [21].

In the following chapters, we design a network-based method for fusing optical flow and depth change and estimating structure flow.

Chapter 3: Research Design

In this Chapter, we present the detailed design of our method. Section 3.1 and 3.2 discuss some mechanisms that can be used for estimating optical flow. Section 3.3 and 3.4 specify the algorithms for estimating two-dimensional and three-dimensional motion field. Section 3.5 and 3.6 demonstrate the design of the convolutional neural network. Some other related design details are included at the end.

3.1 MOTION EXTRACTION METHODOLOGY

FlowNet [16] formed an architecture that takes correlation layer to capture the corresponding spatial features from two feature maps. By denoting two feature maps with $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$, the correlation operation $\boldsymbol{c}((\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2))$ compares the square patches that are centred in $(\boldsymbol{x}_1, \boldsymbol{y}_1)$ on $\boldsymbol{f}_1$ and centred in $(\boldsymbol{x}_2, \boldsymbol{y}_2)$ on $\boldsymbol{f}_2$. Such operation is shown below.

$$\boldsymbol{c}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{o \epsilon [-k,k] \times [-k,k]} \langle \boldsymbol{f}_1(\boldsymbol{x}_1 + \boldsymbol{o}_x,\ \boldsymbol{y}_1 + \boldsymbol{o}_y),\ \boldsymbol{f}_2(\boldsymbol{x}_2 + \boldsymbol{o}_x,\ \boldsymbol{y}_2 + \boldsymbol{o}_y) \rangle$$

where $\boldsymbol{o}$ is a vector with elements $\boldsymbol{o}_x$ and $\boldsymbol{o}_y$. The size of patch is $2k \times 2k$. It is noted that the distance between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ should be restricted since the correlation computation is time-consuming. We take d to denote the maximal distance, which is

$$|\boldsymbol{x}_2 - \boldsymbol{x}_2| < d$$

The size of patch k and the maximal distance between patches should be specified for the correlation operation, while such parameters are highly related to datasets. For instances, k needs to have a smaller value for acquiring fine spatial feature, while d should have a larger value if the dataset has many long-range motions.

In this case, although correlation layer has been empirically proven that it can explicitly drive the motion map with the correlation; we still consider it as an inefficient mechanism as the setting of its parameters is ambiguous.

[37] applied image warping to estimate two-dimensional motion flow. Denoting $warp(\boldsymbol{f}, \boldsymbol{\phi})$ as an operation for warping feature map $\boldsymbol{f}$ with optical flow $\boldsymbol{\phi}$, the error

of $\boldsymbol{\phi}$ can be reflected by calculating the Euclidean norm between the warped first image and the second image.

$$error = \|warp(\boldsymbol{f}_1, \boldsymbol{\phi}) - \boldsymbol{f}_2\|_2$$

The error can be reduced by updating optical flow $\boldsymbol{\phi}$ with $\frac{d\boldsymbol{l}}{d\boldsymbol{\phi}}$. It can be considered as an implicit method that applies the warping operation to bridge the corresponding spatial features in two images.

Compared with the correlation layer, warping has a drawback as it cannot restrict the influence of remote spatial features, although it saves the computational cost.

3.2 SPATIAL PYRAMID STRCUTURE

The spatial pyramid structure was used to be called hierarchical correlation on motion estimation [37], and it was reintroduced in [28] that utilized convolutional neural network to generate two-dimensional optical flow map. This coarse-to-fine detection which was reflected from [25] is easy to follow when dealing with multi-scale motion estimation.

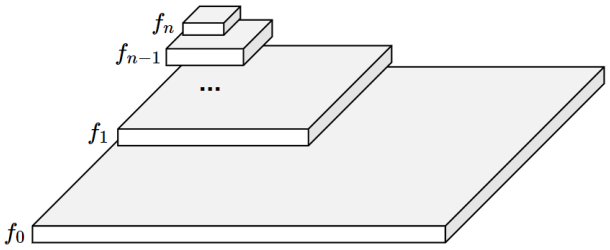


Figure 4: Demonstration of spatial pyramid structure

In general cases, the spatial pyramid structure is implemented with maximal pooling operation. Given an initial feature map $\boldsymbol{f}_0$ with size $a \times b$, the high-level feature $\boldsymbol{f}_n$ has size $(\frac{a}{2^n} \times \frac{b}{2^n})$ and it can be iteratively generated with max-pooling:

$$\begin{aligned}\boldsymbol{f_n}(x,y) = \max\{&\boldsymbol{f}_{n-1}(2x-1,2y-1),\\ &\boldsymbol{f}_{n-1}(2x-1,2y),\\ &\boldsymbol{f}_{n-1}(2x,2y-1),\\ &\boldsymbol{f}_{n-1}(2x,2y)\}\end{aligned}$$

With the feature set $\boldsymbol{F} = \{\boldsymbol{f}_0, \ldots \boldsymbol{f}_n\}$, the motion estimation can be iteratively conducted from the low-level feature to the high-level feature.

As for an inverse operation for acquiring a feature map with larger size, we applied the image interpolation from [36] to up-sample the feature map.

3.3 METHOD FOR ESTIMATING TWO-DIMENSIONAL MOTION FIELD

In the following paragraph, the notation $\boldsymbol{im}_1$ represents a source image and $\boldsymbol{im}_1{}'$ means the continuous image corresponded with $\boldsymbol{im}_1$. The optical flow map $\boldsymbol{\phi}_1$ has the same size with $\boldsymbol{im}_1$ and demonstrates the motion from $\boldsymbol{im}_1$ and $\boldsymbol{im}_1{}'$. In this case, we consider the architecture of SpyNet as baseline and formulate this processing for estimating the two-dimensional motion field from coarse to fine [28].

The corresponding image pairs with different size is iteratively generated with max-pooling operation, and the index of image pair is denoted with $i = 1, \dots, n$. Specifically, the width and length of $\boldsymbol{im}_i$ is two times smaller than those of $\boldsymbol{im}_{i-1}$. The set of image pair is denoted with $\boldsymbol{Im} = \{(\boldsymbol{im}_1, \boldsymbol{im}_1{}'), \dots (\boldsymbol{im}_n, \boldsymbol{im}_n{}')\}$.

Our estimation starts from the coarsest image pair $\boldsymbol{im}_n$ and $\boldsymbol{im}_n{}'$, and the optical flow $\boldsymbol{\phi}_{n+1}$ is initialized to be a zero map with half the length and width of $\boldsymbol{im}_n$**.** Supposing we have an optical flow estimator function $P_2((\boldsymbol{im}_i, \boldsymbol{im}_i'), \widehat{\boldsymbol{\phi}}_i)$ while i is in the range from n to 1, the function takes arbitrary image pair $(\boldsymbol{im}_i, \boldsymbol{im}_i')$ and the corresponding optical flow $\widehat{\boldsymbol{\phi}}_i$ to estimate a finer optical flow map $\boldsymbol{\phi}_i$. It is noted that though $\widehat{\boldsymbol{\phi}}_i$ and $\boldsymbol{\phi}_i$ have the dimension, while $\widehat{\boldsymbol{\phi}}_i$ is acquired by upsampling the coarser optical flow $\boldsymbol{\phi}_{i+1}$ with interpolation.

The algorithm of 2D optical flow estimation is expressed as below.

Algorithm 1: The two-dimensional optical flow estimation

Input: A set of image pairs $\boldsymbol{Im} = \{(\boldsymbol{im}_1, \boldsymbol{im}_1{}'), \dots (\boldsymbol{im}_n, \boldsymbol{im}_n{}')\}$ in different sizes **Input**: Initialized zero optical flow $\boldsymbol{\phi}_{n+1}$ **Input**: Maximal level of structure spatial pyramid n **Output**: Estimated 2D optical flow $\boldsymbol{\phi}_1$
For i in range from n to 1: Up-sampled optical flow $\widehat{\boldsymbol{\phi}}_i \leftarrow$*interpolation* $(\boldsymbol{\phi}_{i+1})$ Warped image $\widetilde{\boldsymbol{Im}}_i \leftarrow$ *warp* $(\boldsymbol{im}_i', \boldsymbol{\phi}_i)$ Feature map $\boldsymbol{F} \leftarrow$ Concatenate $\{\boldsymbol{Im}_i, \widetilde{\boldsymbol{Im}}_i, \widehat{\boldsymbol{\phi}}_i\}$ in channel dimension Estimate optical flow $\boldsymbol{\phi}_i \leftarrow P_2(\boldsymbol{F})$

3.4 METHOD FOR ESTIMATING THREE-DIMENSIONAL MOTION FIELD

The novel formulation of estimation of 3D optical flow is based on the Algorithm 1 in pervious section. We have a new function $P_3((\boldsymbol{im}_{\mathrm{i}}, \boldsymbol{im}'_{\mathrm{i}}), \widehat{\boldsymbol{\phi}}_i)$ for predicting $3D$ motion. Similar to function P_2, function P_3 takes image pair $(\boldsymbol{im}_{\mathrm{i}}, \boldsymbol{im}'_{\mathrm{i}})$ and the up-sampled coarser optical flow $\widehat{\boldsymbol{\phi}}_i$ as input and predicts the 3D optical flow $\boldsymbol{V}_i$ with the same size of image $\boldsymbol{im}_i$. Since warping operation can only apply on two-dimensional image with optical flow, we extract the optical flow $\boldsymbol{\phi}_i$ from $\boldsymbol{V}_i$ for reserving the warping operation.

Algorithm 2: The three-dimensional optical flow estimation

Input: A set of image pairs $\boldsymbol{Im} = \{(\boldsymbol{im}_1, \boldsymbol{im}_1{}'), \ldots (\boldsymbol{im}_n, \boldsymbol{im}_n{}')\}$ in different sizes

Input: Initialized zero optical flow $\boldsymbol{\phi}_{n+1}$

Input: maximal level of structure spatial pyramid n

Output: Estimated 3D motion flow $\boldsymbol{V}_1$

For i in range from n to 1:

- Up-sampled optical flow $\widehat{\boldsymbol{\phi}}_i \leftarrow$ *interpolation* $(\boldsymbol{\phi}_{i+1})$
- Warped image $\widetilde{\boldsymbol{Im}}_i \leftarrow$ *warp* $(\boldsymbol{im}'_i, \boldsymbol{\phi}_i)$
- Feature map $\boldsymbol{F} \leftarrow$ Concatenate $\{\boldsymbol{Im}_i, \widetilde{\boldsymbol{Im}}_i, \widehat{\boldsymbol{\phi}}_i\}$ in channel dimension
- Estimate 3D motion flow $\boldsymbol{V}_i \leftarrow P_3(\boldsymbol{F})$
- Extracted 2D optical flow $\boldsymbol{\phi}_i$ from $\boldsymbol{V}_i$

3.5 NETWORK ARCHITECTURE FOR ESTIMATING 3D OPTICAL FLOW

The motion flow estimators P_2 and P_3 in previous section are implemented with a 6-layer convolutional neural network (CNN). Such network takes stacked features as input with arbitrary size and fixed number of channels.

Table 1: The specification of CNN module, padding operation is applied for reserving the same size of feature maps

	Layer1-input	Layer2	Layer3	Layer4	Layer5	Layer6-output
Number of layers	9	32	64	32	32	3
Kernel size	7	7	7	7	7	7

Although we considered a spatial pyramid structure that is able to handle multiple scale features, it still has some drawbacks. On the one hand, since the CNN module is updated with backpropagation, the error in each layer is accumulated leads to a huge numerical updating on the network weights. Such updating will cause an overflow problem if the pyramid structure consists of too many layers. On the other hand, some fine objects cannot be visible on the coarsest images. It would drive the network to neglect these objects when estimating their motions. Due to these reasons, we set the levels of spatial pyramid to be four ($n = 4$). And it means for each input images, they will be down-sampled for four times.

The network architecture is shown below for demonstrating the process of motion estimation. The input image pairs are down-sampled for 4 times from left to right and transmitted to the red processor module in opposite direction.

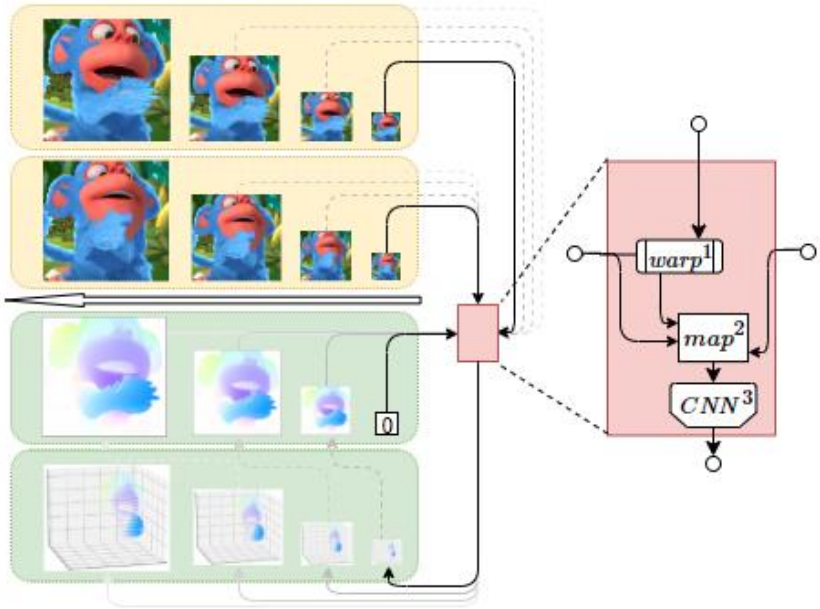


Figure 5: The architecture of network for estimating three-dimensional motion

The smallest image pair $\boldsymbol{im}_4, \boldsymbol{im}_4{}'$ and a zero optical flow map are taken for conducting the operation shown in Algorithm2. The first image $\boldsymbol{im}_4$ is warped with the initialized zero map to generated $\widetilde{\boldsymbol{Im}}_4$ (in $warp^1$ block). The warped second image $\widetilde{\boldsymbol{Im}}_4$, the second image $\widetilde{\boldsymbol{Im}}_4$ and the optical flow are concatenated to be a feature map with six channels (in map^2 block**)**. The CNN^3 module then processes the feature map and generates a three-dimensional motion field $\boldsymbol{V}_4$ as the output of the initial

iteration. Such iterative process will be ended until $\boldsymbol{V}_0$, which has the same size with original images, is generated.

3.6 LOSS FUNCTION DESIGN FOR SCENE FLOW AND STRUCTURE OPTICAL FLOW

In the previous works, there are few datasets containing accurate and dense measurement of the motion in three dimensions. In this case, we suggest generating three-dimensional structure flow data based on optical flow $\boldsymbol{\phi}$ and disparity $\boldsymbol{\mathcal{D}}$.

Given continuous input images $\boldsymbol{im}$ and $\boldsymbol{im'}$, the corresponding pixel-wise disparity maps are denoted by $\boldsymbol{\mathcal{D}}$ and $\boldsymbol{\mathcal{D}'}$. If the optical flow $\boldsymbol{\phi}$ that describes the two-dimensional motion from $\boldsymbol{im}$ to $\boldsymbol{im'}$ is provided, the ground truth of scene flow $\boldsymbol{V}_{ground}$ and structure optical flow $\boldsymbol{w}_{\boldsymbol{ground}}$ are expressed by:

$$\boldsymbol{V}_{ground} = \left\{\left[\boldsymbol{\phi}_x,\ \boldsymbol{\phi}_y,\ \boldsymbol{\mathcal{D}'} - warp(\boldsymbol{\mathcal{D}}, \boldsymbol{\phi})\right]\right\}$$

$$\boldsymbol{w}_{\boldsymbol{ground}} = \left\{\left[\frac{\boldsymbol{\phi}_x}{\boldsymbol{\mathcal{D}}}, \frac{\boldsymbol{\phi}_y}{\boldsymbol{\mathcal{D}}}, \frac{\left(\boldsymbol{\mathcal{D}'} - warp(\boldsymbol{\mathcal{D}}, \boldsymbol{\phi})\right)}{\boldsymbol{\mathcal{D}}}\right]\right\}$$

With the predicted motion field $\boldsymbol{V}_{\boldsymbol{pred}}$ and $\boldsymbol{w}_{\boldsymbol{pred}}$, the loss function for both three-dimensional motion fields is defined with Euclidean norm.

$$\boldsymbol{Loss}_V = \left|\boldsymbol{V}_{pred} - \boldsymbol{V}_{ground}\right|_2$$

$$\boldsymbol{Loss}_w = \left|\boldsymbol{w}_{pred} - \boldsymbol{w}_{\boldsymbol{ground}}\right|_2$$

3.7 DATASET SELECTION

The mainstream dataset of optical flow are summarized as below.

Table 2: The summary of the optical flow dataset

	Effective frames[1]	Virtual	Precision	Motion proportion[2] $r_x : r_y : r_z$	Scene category
Monkaa	8640	True	Accurate	-60 : 0.83 : 1	Action anime
Driving	1098	True	Accurate	-4.7 : 1.1 : 1	Driving
Sintel	1041	True	Inaccurate z	/	Action anime
Kitti	191136	False	Sparse	/	Driving
Virtual Kitti	8640	True	Inaccurate z	3.47 : 4.21 : 1	Driving

[1] Each set of paired image and its structure optical flow is considered as an efficient frame

[2] The proportion of motion is in the sequence of x, y and z, while the positive values indicate leftward, upward and looming (i.e., moving towards the camera) direction respectively.

There are various optical flow datasets, but most of them are not suitable for estimating three-dimensional structure flow. One of the problems is the low accuracy of measurement on the ground truth of optical flow. This occurs in Kitti, Virtual Kitti and Sintel datasets. For Kitti and Virtual Kitti, the depth measurements of the fine object would be missing in some frames, which makes these objects have abnormal depth change in the annotation map.

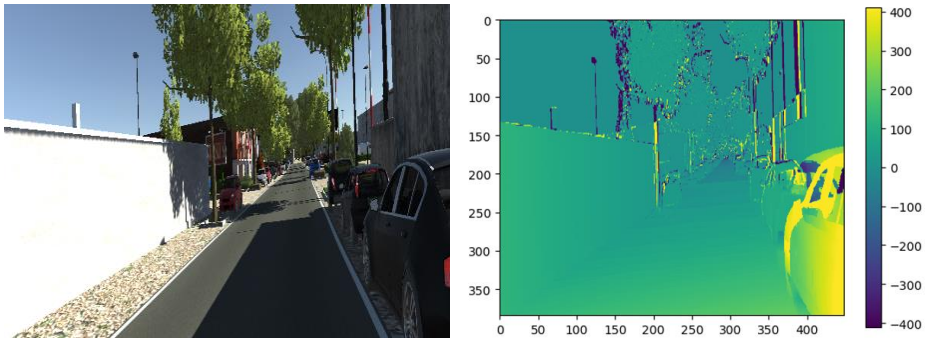
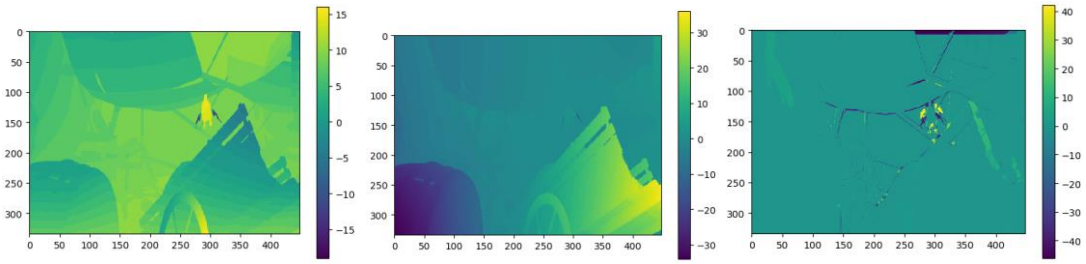


Figure 6: The training image and its annotated depth change in Virtual Kitti. The tree trunks and poles have abnormally large or small value due to the missing of their depth in either the first or the second frame

In Sintel, the optical flow value is stored as float, while the depth value is stored as integer. It causes an ambiguous representation of the pixel-wise depth change.



(a) Structure flow in x axis (b) Structure flow in y axis (c) Structure flow in z axis

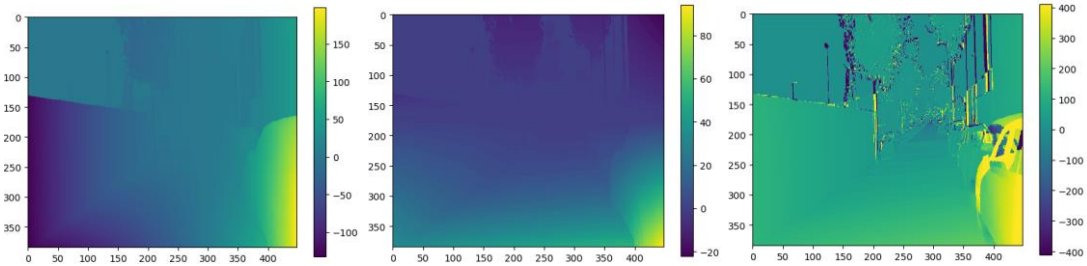
Figure 7: The annotation of structure flow of the Sintel dataset. It is noted the structure flow of the front object in z-axis (c) does not have a continuous pattern, as it should have a more uniform motion like (a) and (b).

Meanwhile, another problem is the imbalance of the magnitude of motion in datasets. The proportion the summed motion in each axis $r_x : r_y : r_z$ is calculated with

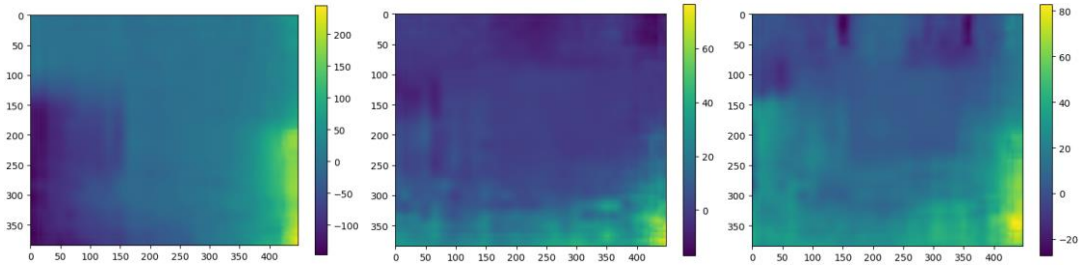
$$r_x : r_y : r_z = \frac{\sum \boldsymbol{\phi}_x}{\min\left(\sum \boldsymbol{\phi}_x + \sum \boldsymbol{\phi}_y + \sum \boldsymbol{\phi}_z\right)} : \frac{\sum \boldsymbol{\phi}_y}{\min\left(\sum \boldsymbol{\phi}_x + \sum \boldsymbol{\phi}_y + \sum \boldsymbol{\phi}_z\right)} : \frac{\sum \boldsymbol{\phi}_z}{\min\left(\sum \boldsymbol{\phi}_x + \sum \boldsymbol{\phi}_y + \sum \boldsymbol{\phi}_z\right)}$$

Such motion proportion of Monkaa, driving and Virtual Kitti are recorded in the Table2. This result shows the significant imbalance of motion among three axes occurs in all three datasets, which means the motion in one or two of the axes will be predominant.

One example of imbalance in Virtual Kitti is shown in Figure 8. We train our model using the training set of Virtual Kitti and made estimation of the testing images. The estimated results of x and y axes have a similar pattern and magnitude with the ground truth, while the result in z-axis is not comparable due to the imbalance of motion proportion.



(a) Structure flow in x axis (b) Motion in y axis (c) Motion in z axis



(a) Estimation in x axis (b) Estimation in y axis (c) Estimation in z axis

Figure 8: Estimated structure flow of Virtual Kitti testing images with a model trained on Virtual Kitti training set. The performance of estimation in z-axis is not comparable with those in the other two axes

The influence of the imbalance is not vital for the method that estimates the optical flow and disparity independently. However, since our model estimates the motion in all three dimensions, the balanced magnitude of dataset is necessary.

3.8 MOTION VISUALIZATION

The current method of flow visualization presents the optical flow with a colour cycle in HSV colour space. Since HSV space can only fit a two-dimensional vector space, the optical flow and the depth change have to be presented separately.

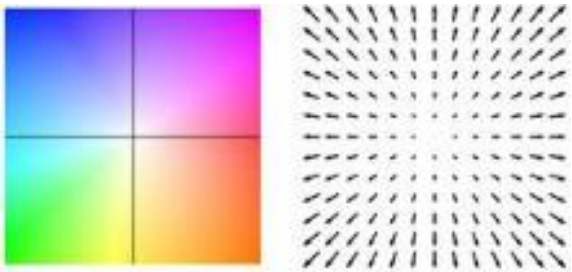


Figure 9: The correlation between optical flow and the colour in HSV space

Although such method accurately displays the motion field, we cannot intuitively compare the magnitude of optical flow and pixel-wise depth change. In this case, we suggest a three-dimensional visualization of motion with mesh plot that combines optical flow and depth change. Since generating a mesh plot is time-consuming, the three-dimensional motion map is downsampled to a sparse manner for presenting.

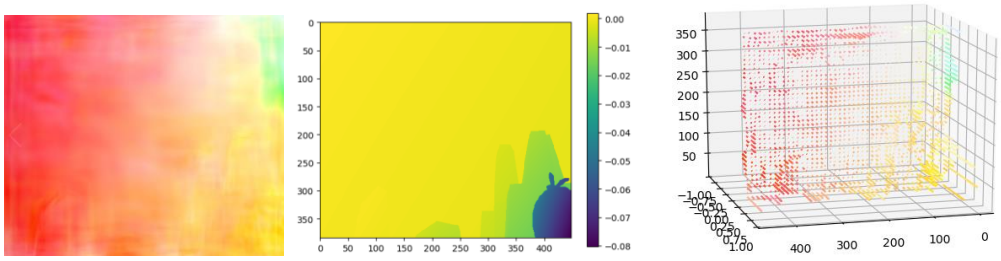


Figure 10: Separate visualization (left: optical flow, mid: depth change) verse downsampled combined visualization

3.9 SETTING OF IMPLEMENTATON

Our structure optical flow estimation is applied on the Monkaa and Driving datasets [38]. At the same time, the estimation of scene flow is also produced for comparing with structure flow. We also evaluate the scene flow estimation performance of our method on standard scene flow benchmark.

With the discussion in Section 3.5, our method includes the estimation of the motion field at four different sizes and we assume its input RGB image has size 448×334. In this case, the size in different level of spatial pyramid is given as below.

Table 3: The size of images in different dimensions

	Level 1 input	Level 2	Level 3	Level output
Size of image	448×334	224×168	112×84	56×42

The model is implemented with Pytorch [38]. It is trained with learning rate $a_1 = 0.001$ and SGD optimization in the first 100 epochs, we then update the learning rate $a_2 = 0.0001$ to let the model converge. The entire process requires around 180 epochs, which is completed in 41 hours on a device with i7-6700k CPU and GTX1080 GPU.

The training set and testing set separation is based on the scenes in datasets. For each scene, we randomly select 10% image pairs for testing, while the rest images pairs are considered as training set.

As for the image argumentation, we take image rotation and random cropping for enhancing the performance of our model. For image rotation, we learn the implementation from [25] and assume rotation range in $[-5°, 5°]$. Such rotated images are randomly cropped to be patches with size 448×334. This image patches are then normalized using a calculated mean and standard deviation based on ImageNet [39].

Chapter 4: Results

4.1 RESULT DEMONSTRATION

We train our model for estimating structure flow and scene flow independently. In this case, we take average end-point-error for evaluating the error between estimation and ground truth in three dimensions.

Despite the three-dimensional motion estimation, we also extract and evaluate two-dimensional motion map from the estimated structure flow and scene flow on the testing set. The experimental results of Monkaa and Driving are show in Table4.

Table 4: Comparison of the average end-point-error of results on Monkaa benchmark

	Structure flow				Scene flow			
	3D motion		2D motion		3D motion		2D motion	
	Train	Test	Train	Test	Train	Test	Train	Test
Monkaa	3.41	7.14	-	5.62	2.74	5.33	-	4.32
Driving	1.74	2.94	-	2.23	1.63	2.81	-	2.22

The two model are trained with three-dimensional structure and scene flow, while the evaluation of testing set is conducted on both two dimensions and three dimensions.

We do not conduct fine-tuning for enhancing the performance since there is no other dataset available for estimating both structure flow and scene flow. The result in Table4 shows that the training on Monkaa is difficult for acquiring a lower loss, it also implies a fact that there is a consistency existing between the there-dimensional and two-dimensional errors.

Table 5: the comparison among supervised two-dimensional optical flow estimation with RGBD input on Monkaa

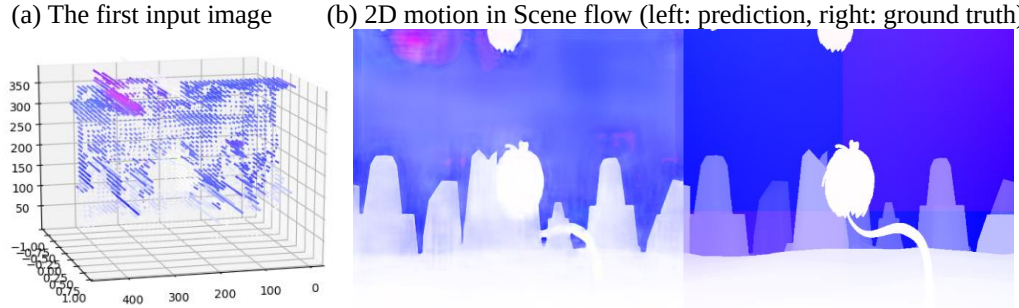
	Optical flow 2D motion		Scene flow 3D motion	
	Train	Test	Train	Test
PD-flow[40]*	43.62	-	-	-
SRSF[1]*	21.81	-	-	-
Sun et al.[42]*	19.54	-	-	-
SF-Net [29]*	4.91	-	-	-
Ours	-	**4.32**	**2.74**	**5.33**

*Take RGBD input, while the exact size of training image is not mentioned.

Since there is no Monkaa benchmark for scene flow, instead, we compare the two-dimensional optical flow extracted from scene flow estimation with other works in Table5. Here, although our model aims to minimize the error of three-dimensional motion and takes only RGB images as input, it still acquires great performance on estimating optical flow compared with the methods employing RGBD images.

With such result, though currently there is no benchmark for structure flow, we still think our model achieves a comparable performance of scene flow estimation.

(a) The first input image (b) 2D motion in Scene flow (left: prediction, right: ground truth)

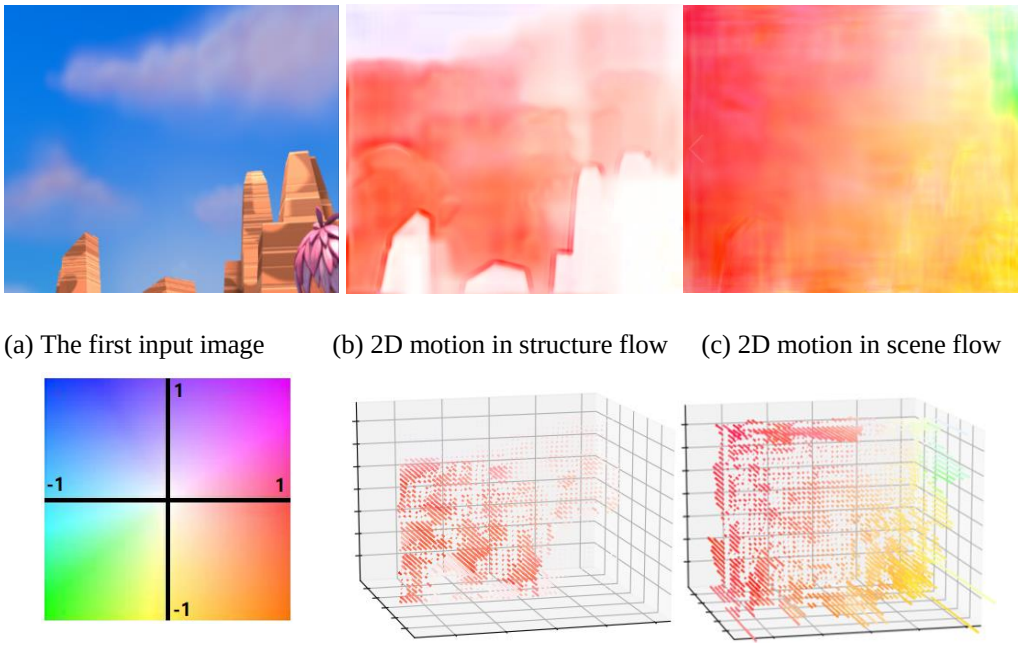


(c) Structure flow (c) 2D motion in structure flow (left: prediction, right: ground truth)

Figure 11: The demonstration of estimated result, both scene flow and structure flow have good performance

4.2 DISCUSSION ON STRUCTURE FLOW AND SCENE FLOW

Structure flow is mathematically defined as the scaled scene flow with depth, so that we intuitively consider it forms a motion representation that is less sensitive to depth change. One of the representative example is shown below.



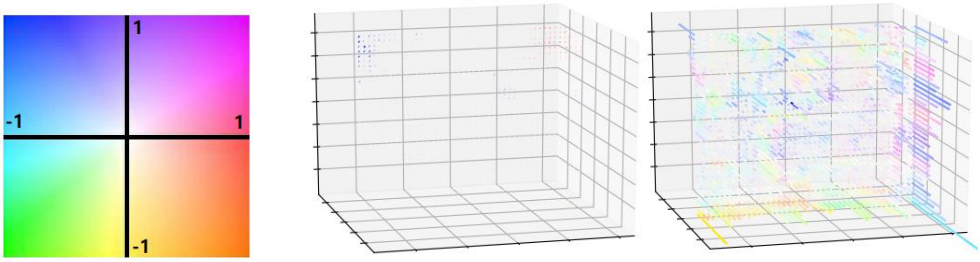
(a) The first input image (b) 2D motion in structure flow (c) 2D motion in scene flow

(d) 2D flow colour map (e) 3D motion in structure flow (f) 3D motion in scene flow

Figure 12: The demonstration of estimated result, structure flow captures more features while scene flow does not

With a camera motion that is rotating around the mountain, the objects include mountain and flower are almost static compared with the moving clouds. In this case, we cannot identify the boundary of the front objects in the motion field captured by scene flow, as the subtle motion of front object and the large motion of the background are mixed. On the contrary, structure flow shows a recognizable front object as their motion is scaled by their depth.

(a) The first input image (b) 2D motion in structure flow (c) 2D motion in scene flow



(d) 2D flow colour map (f) 3D motion in structure flow (f) 3D motion in scene flow

Figure 13: The demonstration of estimated result, scene flow is orderless

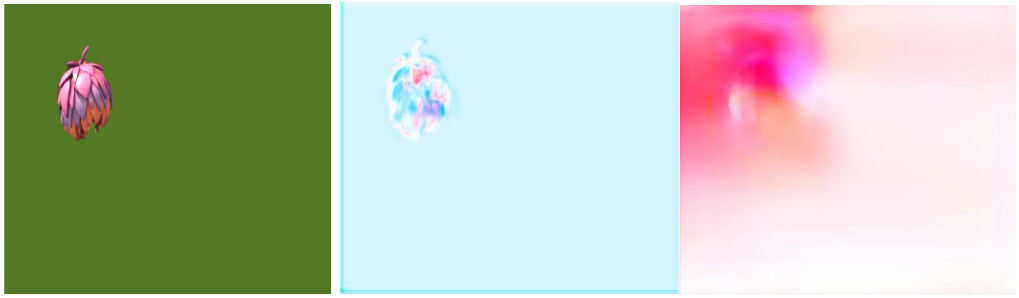
As for another example (Figure 13), the camera conducts a motion with small magnitude, which can be correctly observed in structure map. Meanwhile, the pixel-wise motion in the scene flow map is orderless. One possible reason for such result might be the environment is hard for a scene flow model to learn.

With the examples above, we believe structure optical flow has advantage for presenting the three-dimensional motion in some certain cases, though it still have drawbacks.

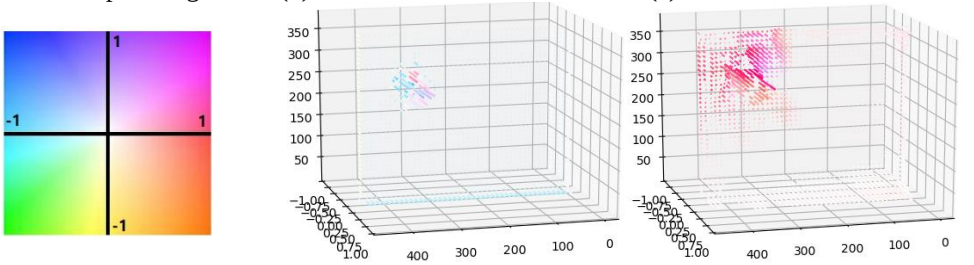
4.3 DISCUSSION ON COLOR INFLUENCE

We observed that there is a consistency existing between the colour and object semantics in Monkaa. For instance, sky is the only object indicates blue in all the scenes. In this case, we designed experiments for validating the performance of the model that is trained on Monkaa in a pure background scene.

We extracted the colour of the grass and the sky in Monkaa as the background in the first and the second experiment respectively. By moving the snipped pink flower rightward for 5 pixels, we conducted both structure flow and scene flow estimation for this motion.



(a) The first input image (b) 2D motion in structure flow (c) 2D motion in scene flow

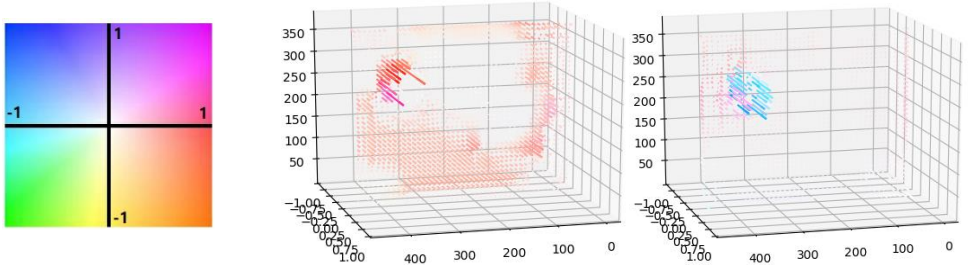


(d) 2D flow colour map (e) 3D motion in structure flow (f) 3D motion in scene flow

Figure 14: The demonstration of estimated result with a green background

For experiment with pure grass green background, we can observe a small motion occurs on the background on both structure flow and scene flow map. As for the motion representation of the pink flower, structure flow map shows a clear boundary but a wrong direction of motion, while scene flow predicts a correct direction but an uncertainty around the flower boundary.

(a) The first input image (b) 2D motion in structure flow (c) 2D motion in scene flow



(d) 2D flow colour map (e) 3D motion in structure flow (f) 3D motion in scene flow

Figure 15: The demonstration of estimated result with a blue background

The result in blue background is totally different. Although structure flow estimates the correct direction of motion of the flower, its map incorrectly reflects a moving background. While scene flow gives an opposite estimation on the motion direction of the flower.

Over all, we can find colour has a significant influence on both structure flow and scene flow estimation, while structure flow has worse performance against the blue background.

In Monkaa, green grass is usually static and blue sky usually occurs in a dynamic manner, which indicates a consistency between the motion and the colour. Besides, with the recorded proportion of motion in Table2 (-60:0.8:1), we can find rightward motion is more frequent, which may explain why the estimated structure flow shows the background moving rightward.

In this case, we can hypothesize that although our model acquires a low end-point-error value, it does not estimate the motion fully based on the pixel-wise position change, instead the relationship between motion and colour is utilized.

Chapter 5: Conclusions

In this thesis, we reviewed the pervious three-dimensional motion estimation approaches and compared it with an innovative representation – structure flow. With the inspiration from existing works, we developed a convolutional neural network method with spatial pyramid structure and warping mechanism for estimating three-dimensional motion. We investigated the available mainstream optical flow datasets and discussed the feasibility of fitting a structure flow model based on these existing current datasets. For validating the performance of structure flow estimation, we did prediction on both structure flow and scene flow.

When it comes to three-dimensional motion, pervious works usually estimated and illustrated optical flow and disparity change independently as they have different units. Since the pixels in structure flow map reserve the same ratio of magnitude in three dimensions with the actual motion, we applied three-dimensional mesh plot for visualizing this motion field in one image. It contributes to an intuitively comparison among the magnitude of motion in all three axes.

We benchmarked the estimated scene flow result in the Monkaa dataset. It acquires low EPE value in two-dimensional motion by taking only RGB images as input. We then discussed the performance of both structure flow and scene flow. Finally, we notice that the consistency between colour and motion has significant influence on the three-dimensional motion estimation.

So far, our main contributions include the method for conducting structure flow training based on the optical flow and disparity measurements. We designed a network-based method with appropriate mechanism for estimating general dense three-dimensional motion field as well as proposed a visualization of motion field in three-dimensional Cartesian coordinates.

Our model shows a great performance on scene flow estimation in Monkaa. It utilizes RGB images as input and acquires a comparable on optical flow estimation compared with previous methods that took RGBD images. Such observation leads many options for our future works. Firstly, it is worth determining how the training of three-dimensional motion helps optical flow estimation. We can validate it by applying

a same model for estimating optical flow and scene flow, and then compare their performance. Secondly, our architecture of model still can be improved, as there are many efficient mechanisms, such as the combination of both forward and backward warping, which worth to be taken for estimating three-dimensional motion. Moreover, with proper regularization and three-dimensional warping operation, it is feasible to discover an unsupervised method.

As for another perspective, it is noted that the current optical flow datasets is a main limitation for our experiment. Structure flow is theoretically good at estimating depth change compared with scene flow, while there is few of scenes in Monkaa consists of conspicuous depth change. At the same time, the imbalanced magnitude of motion in different axes as well as the imbalanced colour also result in a bias in Monkaa dataset. In this case, it is worth creating a new dataset for dealing with these problems.

Although structure flow and scene flow have a similar mathematic expression, with the Figure 11 and Figure 12, we still find that their estimated maps can reflect different behaviours with the same input images. This observation leads us to a hypothesis that a model may be able to estimate better three-dimensional motion if it has learnt both structure flow and scene flow. This hypothesis would lead to a multi-task learning problem like [43].

In conclusion, the structure flow provides a new view of three-dimensional motion that is different from scene. Our work can be considered as a baseline for the development of structure flow and we believe the idea of structure flow should benefit the motion estimation in future.

Bibliography

[1] Julian Quiroga Sepulveda. Scene Flow Estimation from RGBD Images. Computer Vision and Pattern Recognition. Universit é de Grenoble, (2014)

[2] Warren, David H., and Edward R. Strelow, eds. Electronic spatial sensing for the blind: contributions from perception, rehabilitation, and computer vision. Vol. 99. Springer Science & Business Media, (2013).

[3] Zhao, Wenyi, et al. "Face recognition: A literature survey." ACM computing surveys. 35.4 (2003): 399-458.

[4] Hu, Weiming, et al. "A survey on visual surveillance of object motion and behaviors." IEEE Transactions on Systems, Man, and Cybernetics, Part C 34.3 (2004): 334-352.

[5] Geiger, Andreas, Philip Lenz, and Raquel Urtasun. "Are we ready for autonomous driving." Computer Vision and Pattern Recognition. (2012)

[6] Sun, Zehang, George Bebis, and Ronald Miller. "On-road vehicle detection: A review." IEEE Transactions on Pattern Analysis & Machine Intelligence 5. (2006): 694-711.

[7] Barron, John L., David J. Fleet, and Steven S. Beauchemin. "Performance of optical flow techniques." International journal of computer vision 12.1. (1994): 43-77.

[8] Rubinstein, Michael, Ce Liu, and William T. Freeman. "Towards longer long-range motion trajectories." (2012).

[9] Sun, Deqing, et al. "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018).

[10] Lucas, Bruce D., and Takeo Kanade. "An iterative image registration technique with an application to stereo vision." (1981): 674.

[11] Gang, Zhao, Wang Xiaoli, and Wang Lirong. "Motion analysis and research of local navigation system for Visual-impaired person based on improved LK optical flow." 2012 Fifth International Conference on Intelligent Networks and Intelligent Systems. IEEE, (2012).

[12] Horn, Berthold KP, and Brian G. Schunck. "Determining optical flow." Artificial intelligence 17.1-3 (1981): 185-203.

[13] Glocker, Ben, et al. "Dense image registration through MRFs and efficient linear programming." Medical image analysis12.6 (2008): 731-741.

[14] Ben-Ezra, Moshe, Shmuel Peleg, and Michael Werman. "Real-time motion analysis with linear-programming." Proceedings of the Seventh IEEE International Conference on Computer Vision. Vol. 2. IEEE, (1999).

[15] Felzenszwalb, Pedro F., and Daniel P. Huttenlocher. "Efficient belief propagation for early vision." International journal of computer vision 70.1 (2006): 41-54.

[16] Dosovitskiy, Alexey, et al. "Flownet: Learning optical flow with convolutional networks." Proceedings of the IEEE international conference on computer vision. (2015).

[17] Zhang, Ye, and Chandra Kambhamettu. "On 3D scene flow and structure estimation." Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Vol. 2. IEEE, (2001).

[18] Vedula, Sundar, et al. "Three-dimensional scene flow." Proceedings of the Seventh IEEE International Conference on Computer Vision. Vol. 2. IEEE, (1999).

[19] Waxman, Allen M., and James H. Duncan. "Binocular image flows: Steps toward stereo-motion fusion." IEEE Transactions on Pattern Analysis and Machine Intelligence 6. (1986): 715-729.

[20] Gong, Minglun, and Yee-Hong Yang. "Disparity flow estimation using orthogonal reliability-based dynamic programming." 18th International Conference on Pattern Recognition Vol. 2. IEEE, (2006).

[21] Juan David Adarve. “Real-time Visual Flow Algorithms for Robotic Applications”. Australian National University, PhD dissertation. (2017).

[22] Ilg, Eddy, et al. "Occlusions, motion and depth boundaries with a generic network for disparity, optical flow or scene flow estimation." Proceedings of the European Conference on Computer Vision. (2018)

[23] Ilg, Eddy, et al. "Flownet 2.0: Evolution of optical flow estimation with deep networks." Proceedings of the IEEE conference on computer vision and pattern recognition. (2017).

[24] Li, Xiaoxiao, and Chen Change Loy. "Video object segmentation with joint re-identification and attention-aware mask propagation." Proceedings of the European Conference on Computer Vision. (2018).

[25] Ranjan, Anurag, and Michael J. Black. "Optical flow estimation using a spatial pyramid network." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017).

[26] Sun, Deqing, et al. "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018).

[27] Yin, Zhichao, and Jianping Shi. "Geonet: Unsupervised learning of dense depth, optical flow and camera pose." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018).

[28] Mayer, Nikolaus, et al. "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016).

[29] Qiao, Yi-Ling, et al. "SF-Net: Learning scene flow from RGB-D images with CNNs." (2018).

[30] Baker, Simon, et al. "A database and evaluation methodology for optical flow." International Journal of Computer Vision 92.1. (2011): 1-31.

[31] Menze, Moritz, and Andreas Geiger. "Object scene flow for autonomous vehicles." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015).

[32] Geiger, Andreas, et al. "Vision meets robotics: The KITTI dataset." The International Journal of Robotics Research32.11. (2013): 1231-1237.

[33] McCane, Brendan, et al. "On benchmarking optical flow." Computer Vision and Image Understanding 84.1. (2001): 126-143.

[34] Bhoi, Amlaan. "Monocular Depth Estimation: A Survey.". 2019.

[35] Zhu, Zheng, et al. "End-to-end flow correlation tracking with spatial-temporal attention." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018).

[36] Dwight B, "Better Image Scaling", Virgina Tech, (2012)

[37] Brox, Thomas, et al. "High accuracy optical flow estimation based on a theory for warping." European conference on computer vision. Springer, Berlin, Heidelberg, (2004).

[38] Paszke, Adam, et al. "Automatic differentiation in pytorch." (2017).

[39] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. (2016).

[40] Jaimez, Mariano, et al. "A primal-dual framework for real-time dense RGB-D scene flow." 2015 IEEE international conference on robotics and automation. IEEE, (2015).

[42] Sun, Deqing, Erik B. Sudderth, and Hanspeter Pfister. "Layered RGBD scene flow estimation." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015).

[43] Siam, Mennatullah, et al. "Motion and Appearance Based Multi-Task Learning Network for Autonomous Driving." (2017).