Unsupervised Learning of Optical Flow With Patch Consistency and Occlusion Estimation

Zhe Ren[a], Junchi Yan[a,b,*], Xiaokang Yang[a], Alan Yuille[c], Hongyuan Zha[a,d]

[a] *MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai 200240, China*
[b] *Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China*
[c] *Departments of Cognitive Science and Computer Science, Johns Hopkins University, Baltimore, MD 21218-2608 USA.*
[d] *School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA, 30332 USA.*

Abstract

Recent works have shown that deep networks can be trained for optical flow estimation without supervision. Based on the photometric constancy assumption, most of these methods adopt the reconstruction loss as the supervision by point-based backward warping. Inspired by the traditional patch matching based approaches, we propose a patch-based consistency to improve the vanilla unsupervised learning method [1]. Instead of only comparing the corresponding pixel intensity, we locate the correspondence by using the image patches with census transform, which is more robust for the illumination variation and occlusion. Moreover, a novel parallel branch is devised to estimate a soft occlusion mask jointly in an unsupervised way. The mask is adopted to weight our patch-based consistency loss to alleviate the influence of the occlusion. The plenty of experiments have been implemented on Flying Chairs, KITTI and MPI-Sintel benchmarks. The results show that our method is efficient and outperforms the peer unsupervised learning methods that are using the FlowNet-liked network.

Keywords: Patch Consistency, Optical Flow Estimation, Occlusion

*Corresponding author
Email addresses: sunshinezhe@sjtu.edu.cn (Zhe Ren), yanjunchi@sjtu.edu.cn (Junchi Yan), xkyang@sjtu.edu.cn (Xiaokang Yang), alan.l.yuille@gmail.com (Alan Yuille), zha@cc.gatech.edu (Hongyuan Zha)

1. Introduction

Optical flow, introduced by [2] in the 1950s, refers to a 2-D vector field caused by the relative motion between frames, which can provide motion-related information under an egocentric coordinate system. Optical flow estimation has been a building block for many computer vision problems, ranging from low level tasks such as object segmentation [3, 4], saliency detection [5], objection registration [6] to high level tasks like video action recognition [7], facial expression recognition [8] and object tracking [9]. Despite the tremendous progress that has been made over the years, optical flow estimation is still recognized as an open problem far from being solved [10], challenged by benchmarks including large motion and appearance variation e.g., KITTI [11, 12] and MPI-Sintel [13].

Traditionally, the seminal work [14] first proposes a variational energy optimization model under the assumption of brightness constancy and local smoothness constraint. Based on this model, a large number of works [15, 16, 17] are raised to make improvements in different aspects. [15] improve the performance by first detecting the edge and occlusion regions and then post-smooth these parts with different filters respectively. [16] propose to fuse optical flow estimation from different methods by the intensity error of corresponding matching patches. [17] improve the joint optical flow estimation and image restoration model by an edge-aware constraint for better edge preservation. However, this type of method usually requires small displacement as a prerequisite. Although large displacements can be solved by incorporating the coarse-to-fine warping technique [18, 19], the errors caused by missing details in the coarse level will be propagated and accumulated through the whole process.

For better solving the large displacement problem, [20] introduces the patch-based descriptor matching into the variational model so that sparse accurate matches can be obtained as a strong prior knowledge. Some works further resort to the PatchMatch [21, 22] method, which is an efficient approximate algorithm

for finding the nearest neighbors of image patches between two related images. All these works show that the image patch is more robust and reliable to be used to find correspondence than one single pixel point. Even though patch matching based approaches have no restriction for the displacement distance, they are usually time-consuming and can not be used in the real-time system.

As the triumph of deep learning in computer vision society, recent works [23, 24] start a trend to adopt convolutional neural networks (CNNs) to estimate the optical flow directly by learning from massive data. The state-of-the-art supervised method, PWCNet [25] has even outperformed nearly all former non deep learning based methods on popular public datasets KITTI and MPI-Sintel, which suggests a promising direction for solving optical flow problem.

Besides, thanks to the fast development of modern GPU, deep learning based methods in practice can be more efficient (typically 100 milliseconds or less per frame on a typical modern GPU) than non deep learning methods which usually include multi procedures and cost up to one or few minutes per frame.

However, the performance of supervised approaches largely depends on the plenty of labeled data, which is known very difficult to acquire in real-world scenes. Existing works often turn to the synthetic data for help, which significantly restricts their practical application. Hence unsupervised methods [1, 26] recently receive more and more attention. The main idea is to replace the regression loss on the ground truth with the reconstruction loss by warping, of which the origin can be found in traditional energy-based flow estimation algorithms. Although no ground truth is needed, the results of these methods still have a notable gap against their supervised counterparts.

In particular, we make two observations for current unsupervised optical flow methods: First, point-based image warping is used to compute the reconstruction loss, which means the correspondence found by networks only depends on the intensity of one pixel, which can be locally ambiguous. Second, occlusion is still a severe challenge for unsupervised optical flow learning methods, and no approach tries to estimate the occlusion mask by network directly.

In this paper, according to the patch consistency, we propose a patch-based

census constancy loss by patch-based warping to increase the matching accuracy. Instead of only considering one point, we make use of a local patch as the representation of the center point to locate its correspondence. Following the work EPPM [27], we compare the difference of patches with census transform, which is insensitive to the illumination variation. Besides, we also devise another branch to estimate a soft occlusion mask simultaneously in an unsupervised way. The estimated occlusion mask is supervised by the pixels with large patch constancy error and bad forward-backward flow consistency and is used in turn to reduce the effect of the occlusion region. In a nutshell, the main contributions are as follows:

1) A novel patch-based warping is proposed to construct a census constancy loss for measuring the patch consistency.

2) A soft occlusion mask is estimated and learned in an unsupervised way by devising a new decoder branch parallel with the optical flow estimation, to alleviate the influence of the occlusion.

3) Extensive experimental results show the efficiency of our techniques, which realizes the state-of-the-art results among the unsupervised learning methods that are using the FlowNet-liked structure.

The rest of the paper is organized as follows. Related work is reviewed in Section 2. In Section 3, the main techniques for the proposed approach are presented. Experimental results are reported in Section 4 and Section 5 concludes this paper in the end.

2. Related Work

Intensive research has been conducted in the field of optical flow estimation since the seminal work [14]. In their proposed classical variational formulation, the algorithm aims to minimize an energy function derived from the brightness constancy with an extra smoothness constraint. Readers can refer to [10] for a comprehensive study on these methods. In this part, we mainly focus on some milestone works, which are closely related to our approach. They will

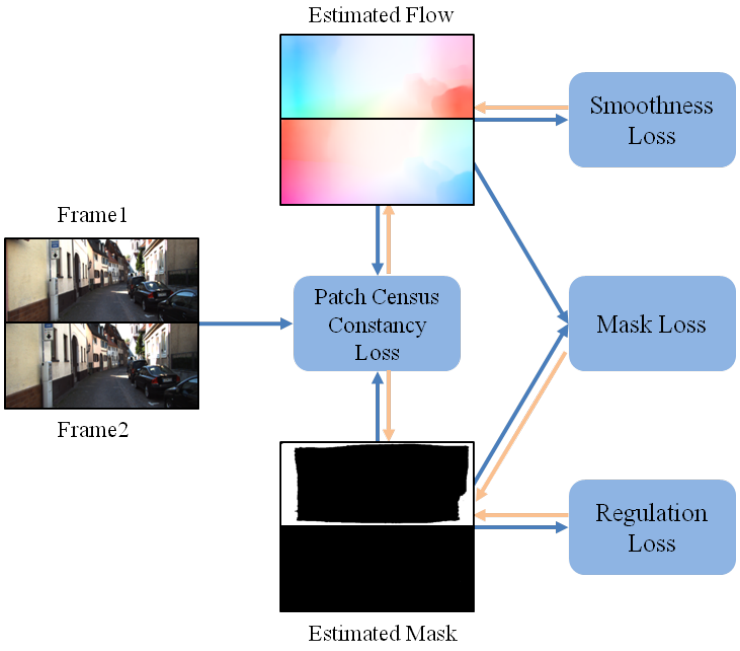


Figure 1: The learning pipeline of our method. The blue arrow denotes the forward information flow, and the orange arrow denotes the backpropagation from losses. The Estimated flows and masks are the output from the network by feeding a pair of images. Patch census constancy loss warps the images to compute patch consistency according to the estimated flows. It is further weighted by the estimated soft occlusion masks so that the effect of the loss over the occluded region is relatively lightened. The estimated flows are also constrained by the smoothness loss while the estimated masks are supervised by two other losses: the mask loss and the regulation loss. See Eq. 7 for the overall loss function.

be reviewed from three aspects: Patch-based methods, unsupervised learning methods, and occlusion handling for flow estimation.

Patch-based methods Compared with the image point, there is a better choice to find correspondence by using the image patch, which provides more context information. In order to address classical aperture problem in optical flow, the famous Lucas-Kanade method [28] assumes the optical flow is constant

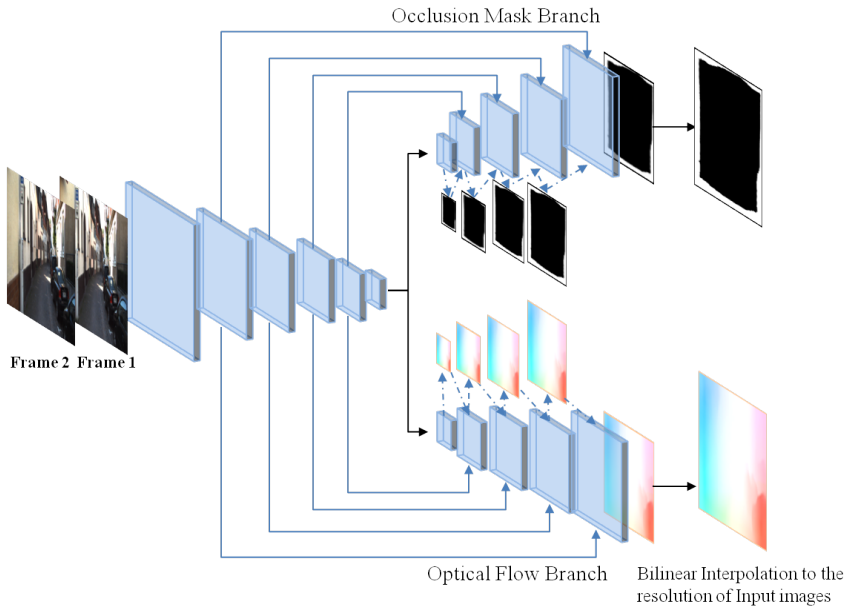


Figure 2: Network structure used in our method. Based on the FlowNetS network in [23], a parallel decoder branch is added for the occlusion mask estimation. The structure of the occlusion branch is nearly the same with the flow branch, which includes multiscale estimation and multiple skip connections from the lower layers to the higher layers. The estimate of each scale is also made as the input for the next scale prediction. The only difference between the two branches is that the channel dimension of each layer in the occlusion branch decrease to 512, 256, 128, 64 and 32 respectively. Both in training and testing, two branches share the same encoder features.

in a local neighborhood, which is actually using patches for matching implicitly. With more discriminative features [29, 30] extracted from image patches are proposed, some works [20, 31] implement sparse matching by the patch-based descriptor and make the result as the initial flow of the variational model. In [20], SIFT-like features are extracted from segmented patches and correspondences are computed by the nearest neighbor method. [31] adopts a hierarchical patch matching method [32], in which patch similarity is calculated by collecting the

correlation maps from its sub patches. Once the correspondence of a patch is determined, it will be used to trace back to the correspondence of its sub patches that make up of its optimal similarity. Inspired by the work PatchMatch [21, 22], many approaches [27, 33, 34] utilize this approximate nearest neighbor method to obtain dense correspondence efficiently. EPPM [27] proposes an edge-preserving PatchMatch so that more details are captured near the motion discontinuities. In order to deal with the drawback of the PatchMatch that the estimated correspondence is noisy, [33] proposes a kd-tree based initialization with a novel multi scales matching to reduce the initial outlier. A further improvement is realized in [34] that PatchMatch is implemented in a coarse-to-fine strategy with a limit range of random search to exclude outliers in the final estimation.

However, the performance of these methods heavily relies on the estimated correspondence, whose accuracy is fixed and can not be improved by making use of a large scale of data. In other words, these methods have no learning ability like deep learning based methods. Moreover, because of the post processing for outlier handling and refinement, patch-based methods are usually time-consuming and are difficult to be used in a real-time system. By introducing the patch consistency into the unsupervised deep learning framework, our approach not only enjoys the robustness of the patch matching but also have the learning ability to improve the performance from massive data. What's more, our method also inherits the speed advantage from the deep learning methods.

Unsupervised learning method Although supervised deep learning methods have dominated many tasks in the computer vision community, ravenous appetite for labeled data greatly hinders their applicability. For this reason, many works try to train the networks with unlabeled data in weakly supervised [35, 36], semi-supervised [37] or completely unsupervised ways [38]. In the optical flow field, the pioneer works [39, 1] first make use of the reconstruction loss from the variational model to train a CNN and adopt the differentiable bilinear interpolation for point-based warping such that the whole network can be trained end-to-end without ground truth. UnFlow [26] improves the results a lot

via stacking several networks with a novel bidirectional census loss. They instead compute the reconstruction loss with a differentiable ternary census transform, which is more robust for the illumination changes in the real scene. In [40], a new backward warping with a larger search space is proposed to overcome the problem that point-based warping easily gets stuck at the local optimal solution. By incorporating 3D scene geometry, GeoNet [41] decomposes the optical flow into static and dynamic parts separately by estimating flow along with depth and camera pose simultaneously. DF-Net [42] further improves the accuracy of both optical flow and depth estimation by adopting cross-task consistency. Even though performance is promoted progressively by these methods, it is still far from satisfaction. In fact, all these approaches are still based on point-based warping, which is significantly distinguished with our method using patch-based warping.

Occlusion handling for flow estimation Occlusion is one of major challenges in optical flow estimation. Because of occlusion, the fundamental assumption of brightness constancy is not valid any longer. So, it is necessary to involve occlusion handling for optical flow approaches. In early work [43], occlusion is explicitly reasoned by the forward-backward consistency and the range of opposite direction flow. The reasoned occlusion mask is then used to assist the optimization of the variational model. Forward-backward consistency is based on the fact that if the pixel is visible in both images, its forward flow should be the same as the backward flow of its corresponding pixel but in the opposite direction. In the PatchMatch based methods [27, 33, 34], they usually filter the outlier matches in the occlusion area by the forward-backward consistency and interpolate the missing flows by following the Epicflow [44]. Instead of solving occlusion by post processing, some works implement a joint optimization of optical flow and occlusion mask estimation by assigning a constant penalty for occlusion pixels [45, 46]. In the recent deep learning method, Unflow [26] is the first to introduce the forward-backward consistency into the unsupervised framework. Based on the idea that the pixels mapped by the backward flow should not be occluded, OccAwareFlow [40] models the non-occluded region

as the range map of the backward flow by a differentiable forward warping. Different from existing methods that are just reasoning out occlusion from the estimated optical flow, we devise another CNN-based branch to estimate a soft occlusion mask explicitly. Besides the supervision from the patch consistency error, a novel pseudo labels extracted from the forward-backward consistency check are proposed to learn the occlusion mask in an unsupervised way.

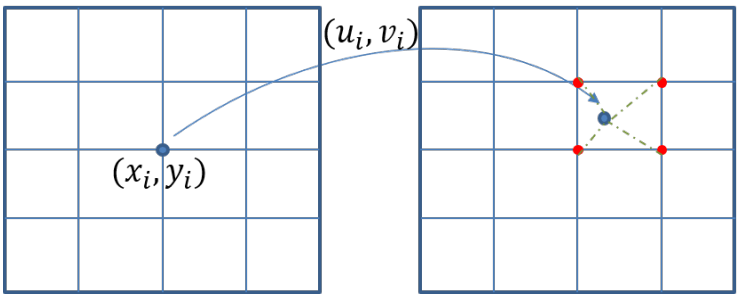
In this paper, we improve the vanilla unsupervised flow learning framework [1] via two novel techniques: patch-based census consistency and CNN-based occlusion learning branch, which can be easily integrated into optical flow network for end-to-end unsupervised learning. The learning pipeline and network structure used in the paper are shown in Fig. 1 and Fig. 2 respectively. More details of the method are illustrated in section 3 and plenty of experiments are exhibited in section 4.

3. Main Approach

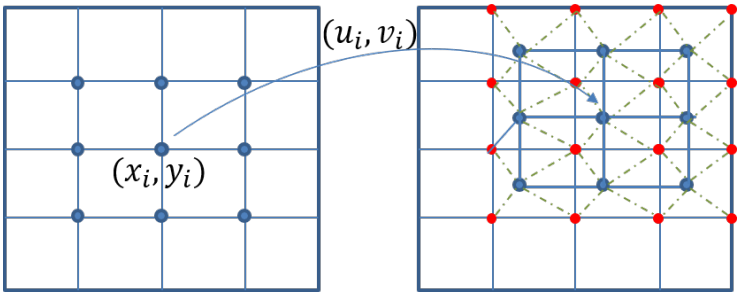
In this section, we first review the widely used point-based photometric constancy loss and propose our patch-based census constancy loss in section 3.1. We further introduce our mask loss for unsupervised occlusion mask learning in section 3.2 and edge-aware smoothness constraint in section 3.3. Finally, the overall loss function is described in section 3.4.

3.1. Patch-based census constancy

We believe patch consistency between two related images that not only the two corresponding pixel points but also the surrounding context should be as similar as possible. Concretely, given two images I_1 and I_2, optical flow estimation is aimed to estimate the pixel-wise correspondence. Assuming forward flow, which is from I_1 to I_2, at any point (x_i, y_i) in source image I_1 is $F(x_i, y_i) = (u_i^f, v_i^f)$, where u_i^f and v_i^f is the displacement at the width and height direction respectively. Sometimes, superscript "f" for flow is omitted for



(a) Point-based warping.



(b) Patch-based warping.

Figure 3: Illustration for point-based warping and patch-based warping. In (a), we project point (x_i, y_i) to its corresponding point by the estimated optical flow (u_i, v_i) at (x_i, y_i). The pixel value of the corresponding point is obtained by the bilinear interpolation from its four nearest neighbors. In (b), we instead warp a 3×3 patch centered at (x_i, y_i) to the patch centered at $(x_i + u_i, y_i + v_i)$ by the estimated flow (u_i, v_i). The pixel value of each patch point is acquired by the bilinear interpolation. In training, the consistency loss will backpropagate the gradient to the estimated optical flow (u_i, v_i) so that the flow can be updated to make the corresponding points or patches as similar as possible. We highlight the pixels used to interpolate corresponding pixel values by the red color. It is obvious that compared with the point-based warping, patch-based warping will consider more context information for finding each single optical flow estimate (u_i, v_i).

brevity. To better understand the idea of our approach, we first review the

point-based photometric constancy loss [1, 26, 40]:

$$\ell_{point} = \sum_{(x_i,y_i)} \Psi\left(|I_2(x_i+u_i, y_i+v_i) - I_1(x_i,y_i)|^2\right) \tag{1}$$

where (x_i, y_i) indexes all the points in I_1 and $\Psi(s) = (s+C)^\gamma$, which is the Charbonnier penalty widely adopted for its robustness [10].

For the guarantee of the gradient backpropagation, the image value $I_2(x_i + u_i, y_i + v_i)$ is usually calculated by a point-based warping with the differentiable bilinear interpolation technique [47]. Following the same reason, we also adopt this standard technique in our patch-based warping operation so that our method can be trained in an end-to-end manner. To better illustrate the difference between two warping methods, we make a straightforward comparison in Fig. 3.

In contrast to the point-based photometric constancy loss in Eq. 1, our patch-based census constancy loss is as follows:

$$\ell_{patch} = \sum_{(x_i,y_i)} \frac{1}{K^2} \sum_{(x_j,y_j)\in\mathcal{N}_i} \frac{|Cen[I_2(x_j+u_i, y_j+v_i)] - Cen[I_1(x_j,y_j)]|^2}{|Cen[I_2(x_j+u_i, y_j+v_i)] - Cen[I_1(x_j,y_j)]|^2 + 0.1} \tag{2}$$

where K is the patch size and $(x_j, y_j) \in \mathcal{N}_i$ denotes the point within the $K \times K$ square neighborhood of the point (x_i, y_i). $Cen[\cdot]$ represents the census transform of the patch, which is proved more robust for illumination changes [48]. In practice, we first sample the patches from two images via interpolation and then adopt a differentiable ternary census transform used in [26]. The similarity of patches is measured by the Hamming distance between the corresponding census features. Note that each step above is differentiable so that the loss can be directly used for the training of CNN network.

Compared with the loss of Eq. 1, our patch-based loss will be more robust because more context information is considered for estimating each flow (u_i, v_i) by an additional inner summation of each point within the neighborhood. In other words, we find the pixel-wise correspondence by using the surrounding patches $\mathcal{N}_i$ instead of the single pixel (x_i, y_i). In the training stage, losses propagate gradients back to the estimated optical flow, which can be viewed as

the searching direction for finding the correspondence. The gradient for point-based loss is only obtained from its four nearest neighbors while that for our patch-based loss comes from more pixels in the neighborhood.

3.2. Occlusion Mask Estimation

In contrast to the existing methods, we propose a CNN-based branch to estimate a soft occlusion mask explicitly. As shown in Fig. 2, by adding another decoder branch, which is parallel with the flow branch, a soft occlusion map with two channels is predicted. After conducting a softmax operation, the value in two channels is normalized into $[0, 1]$. Specifically, we let the first channel as the possibility of the pixels being occluded. The estimated soft mask is utilized for weighting the patch-based consistency loss so that the pixel that is more likely being occluded has less effect on the final result. Consequently, our patch-based loss is reformed as follows:

$$\ell_{patch} = \sum_{(x_i,y_i)} \frac{\bar{O}_1(x_i,y_i)}{K^2} \sum_{(x_j,y_j)\in\mathcal{N}_i} \frac{|Cen[I_2(x_j+u_i,y_j+v_i)] - Cen[I_1(x_j,y_j)]|^2}{|Cen[I_2(x_j+u_i,y_j+v_i)] - Cen[I_1(x_j,y_j)]|^2 + 0.1} \tag{3}$$

where $\bar{O}_i$ is the possibility of pixels in image I_i that are not occluded. In fact, $\bar{O}_i$ is the second channel in our mask estimation.

The supervision for training mask mainly comes from two sources: one is the patch-based consistency loss, and the other is the forward-backward consistency check. We argue that the pixel with more possibility being occluded will derive larger error from patch consistency loss. In Eq. 3, the loss is differentiable over $\bar{O}_i$ so that patch consistency errors can backpropagate into occlusion mask branch.

Another source we use is the idea that the optical flow of the occluded pixel will be inconsistent with that of the pixel it maps. During training, we estimate the forward and backward optical flow $F(x_i, y_i) = (u_i^f, v_i^f)$ and $B(x_i, y_i) = (u_i^b, v_i^b)$ respectively. Assuming that (x_2, y_2) in I_2 corresponds to (x_1, y_1) in I_1 by $(x_2, y_2) = (x_1, y_1) + F(x_1, y_1)$, the consistency will be validated by the

following condition:

$$\begin{cases} \frac{||(x_2,y_2)+B(x_2,y_2)-(x_1,y_1)||}{||F(x_1,y_1)||} < \epsilon_1\,, & if \;\; ||F(x_1,y_1)|| \neq 0\,, \\ \frac{||(x_2,y_2)+B(x_2,y_2)-(x_1,y_1)||}{0.5} < \epsilon_1\,, & if \;\; ||F(x_1,y_1)|| = 0\,, \end{cases} \tag{4}$$

where $||\cdot||$ is the ℓ_2 norm.

As a result, a map $\bar{O}_1^{psdo}$ for image I_1 is obtained where all the points that satisfy the conditions above are labeled as **1** and others as **0**. We regard this map as a pseudo label for the mask learning so that a mask loss is proposed as:

$$\ell_{mask} = \sum_{(x_i,y_i)} ||\bar{O}_1(x_i,y_i) - \bar{O}_1^{psdo}(x_i,y_i)|| \tag{5}$$

Similarly, the pseudo label $\bar{O}_2^{psdo}$ for image I_2 can be derived in the same way.

Following a similar technique in [49] used for depth estimation, a regularization term $\ell_{reg}(\bar{O}_i)$ is imposed on occlusion mask $\bar{O}_i$ to prevent trivial solution of all-zero mask by minimizing the cross-entropy loss with constant label **1** at each position, which means all points are assumed to be non-occluded initially.

3.3. Edge-aware Smoothness Constraint

Smoothness is an essential constraint for optical flow estimation, which assumes each flow should be similar to its neighbors in the local area. Different from existing methods that are using the numeric approximation of the first-order or second-order derivation for smoothness measurement, we directly minimize the difference of each flow estimate with its four neighbors. However, the smoothness assumption does not hold at the motion boundaries where flows from two sides are different. To preserve the discontinuity of flow at these areas, we use the edge of images as the cue for the motion boundaries and weight each pair of neighbors by their intensity difference in the Lab color space.

$$\ell_{smooth} = \sum_{(x_i,y_i)} \sum_{(x_j,y_j)\in\mathcal{N}_i} \omega(I_i^{Lab},(x_i,y_i),(x_j,y_j))\Psi\left((u_i-u_j)^2+(v_i-v_j)^2\right), \tag{6}$$

Table 1: Benchmark statistics. KITTI has two versions where single-view (s-view) samples are in pairs and are labeled with sparse ground truth (GT). The multi-view (m-view) one is an extension set without GT.

Statistics \ Dataset	MPI-Sintel	Flying Chairs	KITTI Benchmark			
			2012		2015	
			s-view	m-view	s-view	m-view
#Pairs	1593	22872	389	7736	400	8000
#Train Set	1041	22232	194	—	200	—
#Test Set	552	640	195	—	200	—
Has GT?	yes	yes	sparse	no	sparse	no

where $\mathcal{N}_i$ is the collection of 4-connected neighbors of point (x_i, y_i) in vertical and horizontal directions. I_i^{Lab} is the image by converting I_i in the *Lab* color space and $\omega(I_i^{Lab}, (x_i, y_i), (x_j, y_j)) = \exp(\frac{-|I_i^{Lab}(x_i,y_i) - I_i^{Lab}(x_j,y_j)|^2}{\sigma^2})$.

3.4. Overall loss function

The overall loss is the weighted sum of patch consistency term ℓ_{patch} (weighted by the mask $\bar{O}_1$), smoothness term ℓ_{smooth}, mask term ℓ_{mask} and the regularization term ℓ_{reg} for both forward and backward optical flows F and B:

$$\begin{aligned}\ell(F, B, \bar{O}_1, \bar{O}_2) = \ell_{patch}(F) + \ell_{patch}(B) + \alpha_1 \left(\ell_{smooth}(F) + \ell_{smooth}(B)\right) \\ + \alpha_2 \left(\ell_{mask}(\bar{O}_1) + \ell_{mask}(\bar{O}_2)\right) + \alpha_3 \left(\ell_{reg}(\bar{O}_1) + \ell_{reg}(\bar{O}_2)\right)\end{aligned} \tag{7}$$

where α_1, α_2 and α_3 are the weight parameters.

4. Experiments and Discussion

4.1. Datasets

Evaluation is performed on three popular benchmarks whose statistics are summarized in Table 1. Generally, we follow the protocol of DSTFlow [1] for the unsupervised flow method evaluation.

Flying Chairs is a synthetic benchmark [23] that consists of segmented background images from Flickr overlayed by random images of chairs. In line

Table 2: Comparison with peer methods on Flying Chairs and MPI-Sintel datasets. Average EPE (End-Point Errors) metric is used for the measurement. Parentheses mean training and testing are on the same dataset. Note that the result of OccAware-Flow on the Flying Chairs is tested based on a different training-test split with other methods.

Method	Chairs	MPI-Sintel Clean		MPI-Sintel Final	
	Test	Train	Test	Train	Test
LDOF [20]	3.47	4.29	7.56	6.42	9.12
EpicFlow [44]	2.94	2.40	4.12	3.70	6.29
EPPM [27]	-	-	6.49	-	8.38
FlowField [33]	-	-	3.75	-	5.81
FlowNetS [23]	**2.71**	4.50	7.42	5.45	8.43
FlowNetS+ft [23]	3.04	(3.66)	6.96	(4.44)	7.76
FlowNet2 [24]	-	2.02	3.96	3.14	6.02
FlowNet2+ft [24]	-	(**1.45**)	4.16	(**2.01**)	5.74
PWCNet [25]	-	2.55	-	3.93	-
PWCNet+ft [25]	-	(1.70)	3.86	(2.21)	**5.13**
DSTFlow [1]	5.11	6.93	10.40	7.82	11.11
Unflow-C [26]	-	-	-	8.64	-
OccAwareFlow [40]	**3.30**	(4.03)	7.95	(5.95)	9.15
PatchFlow	**3.65**	(4.45)	**7.7**	(4.99)	**7.98**

with [23], we split the data into 22232 and 640 image pairs for training and testing respectively.

KITTI dataset [11] collects photos in city streets captured by a driving platform. Optical flow methods have been challenged by its large displacements, various lighting conditions, and severe occlusion. 'KITTI2012' [11] consists of

Table 3: Comparison with peer methods on the KITTI dataset. Besides the average EPE (End-Point Error) metric, Fl-all means the ratio of the outliers over all the region. A pixel is considered to be correctly estimated if its end-point error is < 3 or $< 5\%$ of the ground truth.

Method	KITTI2015			KITTI2012	
	Train		Test	Train	Test
	EPE	Fl-all	Fl-all	EPE	EPE
LDOF [20]	18.23	37%	-	13.73	12.4
EpicFlow [44]	9.57	28%	27%	3.47	3.80
EEPM [27]	-	-	-	-	9.2
FlowField [33]	8.33	24.43%	-	3.33	-
FlowNetS [23]	-	-	-	8.26	-
FlowNetS+ft [23]	-	-	-	7.52	9.10
FlowNet2 [24]	10.06	30.37%	-	4.09	-
FlowNet2+ft [24]	(2.30)	(**8.61**%)	10.41%	(**1.28**)	1.80
PWCNet [25]	10.35	33.67%	-	-	-
PWCNet+ft [25]	(**2.16**)	(9.80%)	**9.60**%	(1.45)	**1.70**
DSTFlow [1]	16.79	36.00%	39.00%	10.43	12.40
Unflow-C [26]	8.80	28.94%	29.46%	3.78	4.5
OccAwareFlow [40]	8.88	-	31.20%	3.55	4.20
DF-Net [42]	8.98	26.01%	25.70%	3.54	4.40
PatchFlow	**6.91**	**21.82**%	**23.46**%	**3.34**	**4**

194 training pairs and 195 test pairs while 'KITTI2015' [12] consists of 200 training pairs and 200 test pairs. In line with [1], we combine the multi-view extended image data (20 frames per scene without ground truth) from the two datasets as

the training set with 13372 image pairs[1], and use the data with ground truth as the validation set (194 pairs for 'KITTI2012' and 200 for 'KITTI2015'). Evaluation of the test set is performed via KITTI's online protocol[2].

MPI-Sintel dataset [13] is obtained from an animated movie which contains large and non-rigid motions. It has the 'Clean' and 'Final' version data. Compared with 'Clean' version data that is rendered with the real illumination effect, 'Final' version data is more challenging with extra atmospheric effects and motion blur.

4.2. Experimental settings

In this paper, patch size K is set as 9 and batch size is chosen as 4. $\sigma = 10$, $\gamma = 0.45$ and $C = 10^{-6}$. Adam optimization method is used with parameter $\beta_1 = 0.9$ and $\beta_2 = 0.999$. In overall loss (Eq. 7), we set α_1, α_2 and α_3 as the 0.0045, 0.178 and 0.31 for the KITTI and MPI-Sintel dataset and make them as 0.009, 0.267 and 0.467 for the Flying Chairs dataset. The start learning rate λ is set by 10^{-4} and decreases by half after a number of iterations. In general, we train 300K iterations for Flying Chairs and 240k iteration for KITTI. Because too few data in MPI-Sintel, we use the model trained on Flying Chairs as initialization and train on MPI-Sintel for 120k iterations with 10^{-5} as the starting learning rate. We first train the network with only optical flow estimation and then make the occlusion mask learning simultaneously. On KITTI and MPI-Sintel datasets, images are resized to the 384×896 for both training and testing. The evaluation results are measured by the widely used end-point error (EPE), which is the average ℓ_2 norm distance between the estimated flow and the ground truth for all pixels. Another metric 'Fl', which is used especially in the KITTI2015 benchmarks, is the percentage of the outliers, which are defined as the pixel with end-point error larger than 3 and 5% of its ground truth at

[1] We have excluded the pairs with ground truth (GT) and their two neighboring frames in multi-view datasets for unsupervised training to avoid the mixture of training and testing samples.

[2] http://www.cvlibs.net/datasets/kitti/user_login.php

Table 4: Ablation study of each component in our method. 'Point' is the model trained only by the point-based photometric loss (Eq. 1) and our edge-aware smoothness loss (Eq. 6). 'Patch Consis' represents using our patch-based census constancy loss (Eq. 2). 'Patch+Mask' is the full model of our method, which includes patch-based loss with occlusion mask branch.

Methods	KITTI2012			KITTI2015			
	ALL	NOC	OCC	ALL	NOC	OCC	Fl-all
Point	6.85	2.65	30.26	13.07	5.94	45.16	32.96%
Patch Consis	3.54	1.44	15.04	7.18	3.25	24.62	23.34%
Patch+Mask	**3.34**	**1.31**	**14.51**	**6.91**	**3.04**	**23.76**	**21.82%**

the same time. All experiments are conducted on a server installed with Titan XP GPU. Note that we have tested the speed of our method in the inference phase and the running time of our approach for single 370×1226 KITTI image is 0.035s, which is much faster than traditional patch-based methods.

4.3. Ablation Study

In order to show the efficiency of each component in our method, we conduct comparison experiments with different function modules on the KITTI dataset in Tab 4. The 'Point' model is the method trained only by point-based photometric constancy loss (Eq. 1) and edge-aware smoothness constraint (Eq. 6). We further replace the point-based loss with the proposed patch-based census constancy loss (Eq. 2) as the 'Patch Consis' model. The full model of our method is denoted as 'Patch+Mask', which consists of patch-based consistency and extra mask estimation and is trained by the overall loss (Eq. 7). All the experiments are made in the same setting.

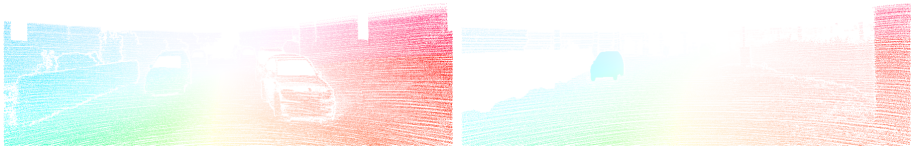
In Table 4, it can be found that both components are beneficial to the flow estimation performance. By comparing the first row with the second row, the proposed patch-based census consistency improves the results by a large margin on all the metrics, which demonstrates it is helpful not only on the non-occluded

(a) Input Images

(b) Point

(c) Patch Consistency

(d) Patch+Mask



(e) GT

Figure 4: Qualitative results of the models in the ablation study on the KITTI dataset. The Left column is from the KITTI2012 and the right one comes from KITTI2015.

Table 5: EPE results on KITTI2015 training set for patch-based census consistency with different patch sizes.

Patch Size	3x3	5x5	7x7	9x9	11x11	13x13
EPE	7.44	7.37	7.27	7.18	7.26	7.32

region, but also on the occluded region. From the second row and the third row, the occlusion branch further decreases the EPE error over the occluded region.

For a better presentation, we also visualize the estimated result of each model in Fig. 4. Estimates from the patch consistency model are obviously better than those of point-based model, especially at the bottom part with large displacement and on the occluded region. 'Patch+Mask' model further corrects the wrong estimation over the occluded region.

4.4. Influence of different patch sizes.

We did the comparative experiments for patch-based census consistency with different patch sizes on the KITTI dataset. Other settings are made the same.

As is shown in the Table 5, our method is robust when the patch size changes. With the patch size increasing, the result is getting better. The optimal patch size is 9×9. Too large patch size makes the performance begin to drop. We think this may because larger patch induces larger errors near the motion boundary and over the occlusion region.

4.5. Results and discussion

The comprehensive comparison with peer methods are conduted on the Flying Chairs and MPI-Sintel datasets in Table 2 and on the KITTI dataset in Table 3. We term our method as PatchFlow and compare with many peer methods including non deep learning methods: EpicFlow [44], EPPM [27], FlowField [33], LDOF [20]; Supervised learning methods: FlowNet [23], FlowNet2 [24], PWC-Net [25]; And unsupervised methods: DSTFlow [1], Unflow [26], OccAwareFlow

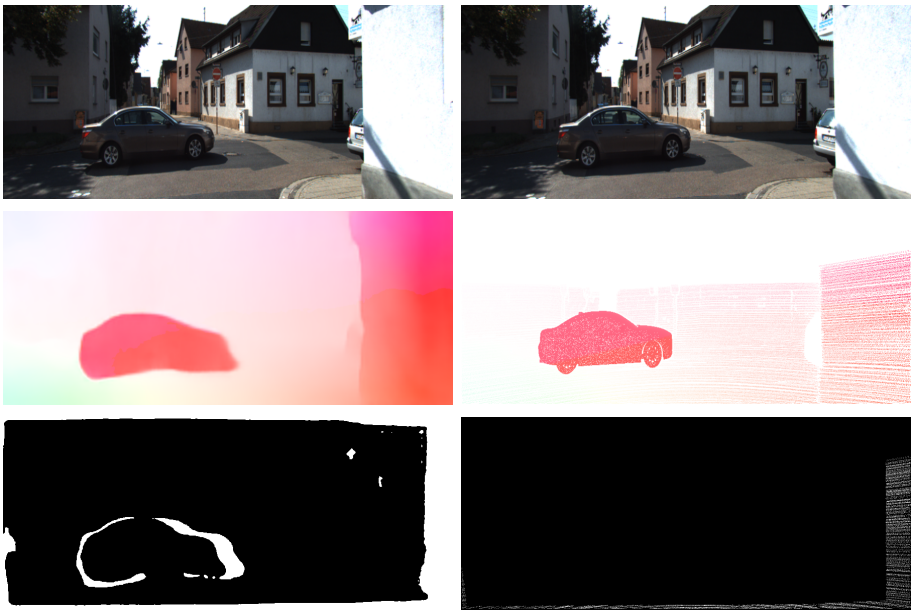


Figure 5: Qualitative results of our full model method on the KITTI2015. From the second row to the third row, the left images are the estimated optical flow and occlusion mask respectively. The right images are the corresponding ground truth. The mask is demonstrated by the threshold 0.5.

[40], and DF-Net [42]. Note that all the unsupervised methods considered here are using the FlowNet-like network structure.

In general, the proposed PatchFlow consistently outperforms other unsupervised methods. In Table 2, because OccAwareFlow [40] did not adopt the same training-test split with other methods on the Flying Chairs dataset, it is meaningless to compare its result with other methods directly. On the MPI-Sintel, our method performs better than other unsupervised methods on the 'Final' version data. Although PatchFlow is interior to the OccAwareFlow on the training set of 'Clean' version, it still achieves a better result on the test set, which shows that OccAwareFlow is prone to be overfitting than ours. Compared with supervised methods and non deep learning based methods, due to the lack of large scales of data on MPI-Sintel, unsupervised methods still lag behind other methods except that our method performs better than FlowNetS

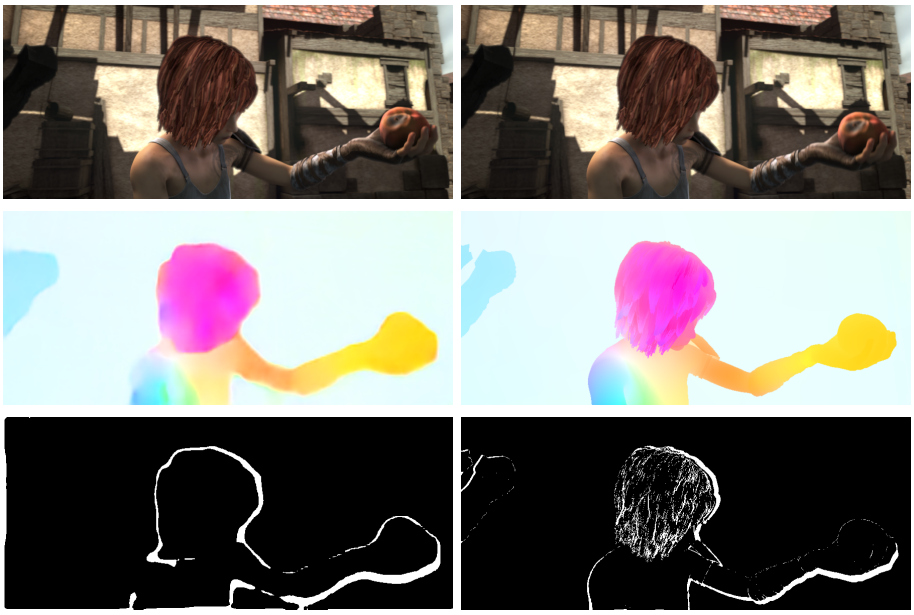


Figure 6: Qualitative results of our full model method on the MPI-Sintel-final. From the second row to the third row, the left images are the estimated optical flow and occlusion mask respectively. The right images are the corresponding ground truth. The mask is demonstrated by the threshold 0.5.

and LDOF methods on the 'Final' version data. On the KITTI dataset, our method realizes the best results among the unsupervised methods. Moreover, on the KITTI2015 dataset, which is more challenging with large displacement and severe occlusion, our method exceeds all the non deep learning methods and even performs better than the supervised method FlowNet2 and PWCNet without finetuning. Both of these two methods also use more powerful networks than ours. It fully indicates that unsupervised methods have an advantage of estimating optical flow in the real scene, where supervised methods are greatly limited if no ground truth is available.

We also demonstrate the estimated results of optical flow and occlusion mask in Fig. 5 and Fig. 6. On KITTI2015 dataset, compared with the sparse ground truth, our method estimates the optical flow and occlusion mask successfully. Note that our estimate also includes the shadow movement of the car while

ground truth is missing. Similarly, the occlusion region caused by car moving is also estimated by our method while it is lost in the ground truth. In the MPI-Sintel dataset, although there is a complex non-rigid motion, our method also makes a reasonable estimation for both optical flow and occlusion mask.

5. Conclusion

In this paper, we have presented two novel techniques to further improve the performance of the vanilla framework [1] for unsupervised optical flow estimation: 1) Patch-based consistency, which locates correspondence by the patches with more robust census transform; 2) Occlusion mask estimation, that extra occlusion branch is devised to estimate soft mask for occlusion handling explicitly. With more context information considered, our patch-based loss derives a more accurate flow estimation than the widely used point-based photometric loss. The census transforms also strengthen the robustness for the illumination changes and occlusion by comparing the relative ordering of intensities. By providing the pseudo label from the patch-based census consistency error and the forward-backward consistency validation for the unsupervised mask learning, our method succeeds in estimating the reasonable occlusion mask, which is used in turn to weight the patch-based constancy loss to alleviate the influence of the occlusion problem. As a result, our techniques efficiently improve the performance of the vanilla method [1] by a large margin and achieves the state-of-the-art results among the unsupervised approaches by using FlowNet-like network.

Limitations: Although a patch with a larger radius makes it more discriminative for matching, it also increases the burden of computation so that training will become more time-consuming. The proper patch size should be a compromise between the accuracy of the estimation and the training speed. Besides, patch consistency will be violated near the motion boundary so that the estimate near the boundary is blurred. So, it is an interesting direction to study how to maintain the motion boundary while using the larger patch. The

hyperparameter ϵ_1 used to validate the forward-backward consistency is important to set for the occlusion mask learning. Too small ϵ_1 at the beginning will make it sensitive for the noisy estimation while too large at the end of training will miss some true occlusion region. It is better to design an adaptive method to decide the ϵ_1 automatically. For optical flow estimation over the occlusion region, there is still no strong supervised signal for learning.

Future work: Also, we will leave these limitations for future study. For maintaining the motion boundary, it may be helpful to introduce the image edge information into the patch. To prevent too much computation burden by large patches, we may try to enlarge the size of patches by sampling each point at intervals of several pixels. As for the value of ϵ_1, a possible alternative is to decrease the ϵ_1 gradually during the training process. For occlusion handling, it is possible to use interpolation to provide strong supervision for estimation learning over the occlusion region.

Outlook: To make deep learning based flow methods applicable in the real scene, unsupervised learning methods will play an important role in the process. We believe our work can benefit not only the practical application but also future research. Our techniques proposed in the paper can serve as the building blocks to construct a practical real-time flow system. The subsequent works may develop better approaches based on our techniques. The model trained by our method is able to be used as a good initialization of training for the supervised learning method. It is also possible to utilize the estimation of our method as the input to provide motion information for solving other tasks.

Acknowledgement

The work is partially supported by NSFC (U19B2035,61972250), National Key R&D Program of China (2018AAA0100704), NSF (1763705) and STCSM (18DZ1112300).

References

[1] Z. Ren, J. Yan, B. Ni, B. Liu, X. Yang, H. Zha, Unsupervised deep learning for optical flow estimation, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 2017, pp. 1495–1501.

[2] J. Gibson, The ecological approach to visual perception: classic edition, Psychology Press, 2014.

[3] W. Wang, J. Shen, F. Porikli, Saliency-aware geodesic video object segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3395–3402.

[4] W. Wang, J. Shen, F. Porikli, R. Yang, Semi-supervised video object segmentation with super-trajectories, IEEE Trans. Pattern Anal. Mach. Intell. 41 (4) (2019) 985–998.

[5] W. Wang, J. Shen, L. Shao, Consistent video saliency using local gradient flow optimization and global refinement, IEEE Transactions on Image Processing 24 (11) (2015) 4185–4196.

[6] S. Yang, L. An, Y. Lei, M. Li, N. Thakoor, B. Bhanu, Y. Liu, A dense flow-based framework for real-time object registration under compound motion, Pattern Recognition 63 (2017) 279–290.

[7] L. Wang, Y. Wang, T. Jiang, D. Zhao, W. Gao, Learning discriminative features for fast frame-based action recognition, Pattern Recognition 46 (7) (2013) 1832–1840.

[8] X. Fan, T. Tjahjadi, A spatial-temporal framework based on histogram of gradients and optical flow for facial expression recognition in video sequences, Pattern Recognition 48 (11) (2015) 3407–3416.

[9] X. Lu, C. Ma, B. Ni, X. Yang, I. Reid, M.-H. Yang, Deep regression tracking with shrinkage loss, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 353–369.

[10] D. Sun, S. Roth, M. J. Black, A quantitative analysis of current practices in optical flow estimation and the principles behind them, International Journal of Computer Vision 106 (2) (2014) 115–137.

[11] A. Geiger, P. Lenz, C. Stiller, R. Urtasun, Vision meets robotics: The KITTI dataset, The International Journal of Robotics Research 32 (11) (2013) 1231–1237.

[12] M. Menze, A. Geiger, Object scene flow for autonomous vehicles, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3061–3070.

[13] D. J. Butler, J. Wulff, G. B. Stanley, M. J. Black, A naturalistic open source movie for optical flow evaluation, in: Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, 2012, Proceedings, Part VI, 2012, pp. 611–625.

[14] B. Horn, B. Schunck, Determining optical flow, Artificial Intelligence 17 (1981) 185–203.

[15] Z. Tu, N. Van Der Aa, C. Van Gemeren, R. C. Veltkamp, A combined post-filtering method to improve accuracy of variational optical flow estimation, Pattern Recognition 47 (5) (2014) 1926–1940.

[16] Z. Tu, R. Poppe, R. C. Veltkamp, Weighted local intensity fusion method for variational optical flow estimation, Pattern Recognition 50 (2016) 223–232.

[17] Z. Tu, W. Xie, J. Cao, C. Van Gemeren, R. Poppe, R. C. Veltkamp, Variational method for joint optical flow estimation and edge-aware image restoration, Pattern Recognition 65 (2017) 11–25.

[18] T. Brox, A. Bruhn, N. Papenberg, J. Weickert, High accuracy optical flow estimation based on a theory for warping, in: Computer Vision - ECCV 2004, 8th European Conference on Computer Vision, Prague, 2004. Proceedings, Part IV, 2004, pp. 25–36.

[19] T. Amiaz, E. Lubetzky, N. Kiryati, Coarse to over-fine optical flow estimation, Pattern recognition 40 (9) (2007) 2496–2503.

[20] T. Brox, J. Malik, Large displacement optical flow: descriptor matching in variational motion estimation, IEEE transactions on pattern analysis and machine intelligence 33 (3) (2010) 500–513.

[21] C. Barnes, E. Shechtman, A. Finkelstein, D. B. Goldman, Patchmatch: A randomized correspondence algorithm for structural image editing, ACM Transactions on Graphics (ToG).

[22] C. Barnes, E. Shechtman, D. B. Goldman, A. Finkelstein, The generalized patchmatch correspondence algorithm, in: Computer Vision - ECCV 2010, 11th European Conference on Computer Vision, Greece, 2010, Proceedings, Part III, 2010, pp. 29–43.

[23] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazirbas, V. Golkov, P. van der Smagt, D. Cremers, T. Brox, Flownet: Learning optical flow with convolutional networks, in: 2015 IEEE International Conference on Computer Vision, 2015, pp. 2758–2766.

[24] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, T. Brox, Flownet 2.0: Evolution of optical flow estimation with deep networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1647–1655.

[25] D. Sun, X. Yang, M.-Y. Liu, J. Kautz, Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8934–8943.

[26] S. Meister, J. Hur, S. Roth, Unflow: Unsupervised learning of optical flow with a bidirectional census loss, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[27] L. Bao, Q. Yang, H. Jin, Fast edge-preserving patchmatch for large displacement optical flow, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 3534–3541.

[28] B. D. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, in: Proceedings of the 7th International Joint Conference on Artificial Intelligence, 1981, pp. 674–679.

[29] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International journal of computer vision 60 (2) (2004) 91–110.

[30] E. Rublee, V. Rabaud, K. Konolige, G. R. Bradski, Orb: An efficient alternative to sift or surf., in: ICCV, Vol. 11, Citeseer, 2011, p. 2.

[31] P. Weinzaepfel, J. Revaud, Z. Harchaoui, C. Schmid, Deepflow: Large displacement optical flow with deep matching, in: Proceedings of the IEEE international conference on computer vision, 2013, pp. 1385–1392.

[32] J. Revaud, P. Weinzaepfel, Z. Harchaoui, C. Schmid, Deepmatching: Hierarchical deformable dense matching, International Journal of Computer Vision 120 (3) (2016) 300–323.

[33] C. Bailer, B. Taetz, D. Stricker, Flow fields: Dense correspondence fields for highly accurate large displacement optical flow estimation, in: 2015 IEEE International Conference on Computer Vision, 2015, pp. 4015–4023.

[34] Y. Hu, R. Song, Y. Li, Efficient coarse-to-fine patchmatch for large displacement optical flow, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5704–5712.

[35] Z. Li, J. Tang, Weakly supervised deep metric learning for community-contributed image retrieval, IEEE Transactions on Multimedia 17 (11) (2015) 1989–1999.

[36] Z. Li, J. Tang, T. Mei, Deep collaborative embedding for social image understanding, IEEE transactions on pattern analysis and machine intelligence.

[37] Z. Li, J. Tang, X. He, Robust structured nonnegative matrix factorization for image representation, IEEE transactions on neural networks and learning systems 29 (5) (2017) 1947–1960.

[38] X. Lu, W. Wang, C. Ma, J. Shen, L. Shao, F. Porikli, See more, know more: Unsupervised video object segmentation with co-attention siamese networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3623–3632.

[39] J. Y. Jason, A. W. Harley, K. G. Derpanis, Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness, in: ECCV Workshops, 2016.

[40] Y. Wang, Y. Yang, Z. Yang, L. Zhao, P. Wang, W. Xu, Occlusion aware unsupervised learning of optical flow, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4884–4893.

[41] Z. Yin, J. Shi, Geonet: Unsupervised learning of dense depth, optical flow and camera pose, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1983–1992.

[42] Y. Zou, Z. Luo, J.-B. Huang, Df-net: Unsupervised joint learning of depth and flow using cross-task consistency, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 36–53.

[43] L. Alvarez, R. Deriche, T. Papadopoulo, J. Sánchez, Symmetrical dense optical flow estimation with occlusions detection, International Journal of Computer Vision 75 (3) (2007) 371–385.

[44] J. Revaud, P. Weinzaepfel, Z. Harchaoui, C. Schmid, Epicflow: Edge-preserving interpolation of correspondences for optical flow, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1164–1172.

[45] M. Unger, M. Werlberger, T. Pock, H. Bischof, Joint motion estimation and segmentation of complex scenes with label costs and occlusion modeling,

in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 1878–1885.

[46] D. Sun, C. Liu, H. Pfister, Local layering for joint motion estimation and occlusion detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1098–1105.

[47] M. Jaderberg, K. Simonyan, A. Zisserman, K. Kavukcuoglu, Spatial transformer networks, in: Advances in Neural Information Processing Systems, 2015, pp. 2017–2025.

[48] R. Zabih, J. Woodfill, Non-parametric local transforms for computing visual correspondence, in: European conference on computer vision, Springer, 1994, pp. 151–158.

[49] T. Zhou, M. Brown, N. Snavely, D. G. Lowe, Unsupervised learning of depth and ego-motion from video, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1851–1858.