Learning by Analogy: Reliable Supervision from Transformations for Unsupervised Optical Flow Estimation

Liang Liu[1*] Jiangning Zhang[1] Ruifei He[1] Yong Liu[1†] Yabiao Wang[2]
Ying Tai[2] Donghao Luo[2] Chengjie Wang[2] Jilin Li[2] Feiyue Huang[2]
[1] Zhejiang University [2]Youtu Lab, Tencent
{leonliuz, 186368, rfhe}@zju.edu.cn, yongliu@iipc.zju.edu.cn
{casewang, yingtai, michaelluo, jasoncjwang, jerolinli, garyhuang}@tencent.com

Abstract

Unsupervised learning of optical flow, which leverages the supervision from view synthesis, has emerged as a promising alternative to supervised methods. However, the objective of unsupervised learning is likely to be unreliable in challenging scenes. In this work, we present a framework to use more reliable supervision from transformations. It simply twists the general unsupervised learning pipeline by running another forward pass with transformed data from augmentation, along with using transformed predictions of original data as the self-supervision signal. Besides, we further introduce a lightweight network with multiple frames by a highly-shared flow decoder. Our method consistently gets a leap of performance on several benchmarks with the best accuracy among deep unsupervised methods. Also, our method achieves competitive results to recent fully supervised methods while with much fewer parameters.

1. Introduction

Optical flow, as a motion description of images, has been widely used in high-level video tasks [47, 48, 52, 3, 2, 31]. Benefitting from the growth of deep learning, learning-based optical flow methods [39, 30] with considerable accuracy and efficient inference are gradually replacing the classical variational-based approaches [36, 25, 44]. However, it is tough to collect the ground truth of dense optical flow in reality, which makes most supervised methods heavily dependent on the large-scale synthetic datasets [7, 26], and the domain difference leads to an underlying degradation when the model is transferred to the real-world.

In another point of view, many works proposed to learn optical flow in an unsupervised way [37, 27, 42, 24], in which the ground truth is not necessary. These works aim to train networks with objective from view synthesis [51, 49], *i.e*. optimizing the difference between reference images and the flow warped target images. This objective is based on the assumption of brightness constancy, which will be violated for challenging scenes, *e.g*. with extreme brightness or partial occlusion. Hence, proper regularization such as occlusion handling [42, 17] or local smooth [27] is required. Recent studies have focused on more complicate regularizations such as 3D geometry constraints [34, 41, 22] and global epipolar constraints [50]. As shown in Fig. 1, there is still a large gap between these works and supervised methods. In this paper, we do not rely on the geometrical regularizations but rethink the task itself to improve accuracy.

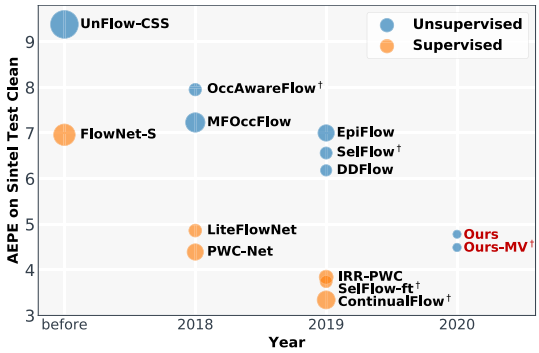

Figure 1. Timeline of average end-point error (AEPE) advances in deep optical flow. Marker size indicates network size, and oversized markers have been adjusted. Our method outperforms all of the previous unsupervised methods, also yields comparable accuracy to supervised methods while with fewer parameters. † indicates the model using more than two frames.

Interestingly, we notice that almost all of the unsupervised works, such as [42, 24, 41], avoid using a heavy combination of augmentations, even if it has been proven effective in supervised flow works [15, 38, 14]. The reason we conclude is two-fold: *(i)* Data augmentation is essentially a trade-off between diversity and validity. It can improve the model by increasing the diversity of data, while also leads to a shift of data distribution which decreases the accuracy.

*Work mainly done during an internship at Tencent Youtu Lab.
†Corresponding author.

2575-7075/20/$31.00 ©2020 IEEE
DOI 10.1109/CVPR42600.2020.00652

In unsupervised learning, the benefit of diversity is limited since the abundant training data is easy to access. *(ii)* Data augmentation will generate challenging samples, for which view synthesis is more likely to be unreliable, so the objective cannot guide networks for a correct solution.

More recently, there are some works based on knowledge distillation that alleviate the problem of unreliable objective in occluded regions [23, 24]. The training of these methods is split into two stages. In the first stage, a teacher model is trained to make predictions on original data, and offline creating occluded samples with random crop or mask out. In the second stage, these artificial samples from the teacher model are used to update a student model. However, these methods were designed for the case of partial occluded only. Hence we ask: *Can we generalize the distillation of occlusion to other transformation cases?* Moreover, the distillation method has a bottleneck due to the frozen teacher model. We thus ask: *Can we jointly optimize teacher model and student model, or just training a single network?*

In this work, we address the above two questions with a novel unsupervised learning framework of optical flow. Specifically, for the first question, diverse transformations are used to generate challenging scenes such as low-light, overexposed, with large displacement or partial occlusion. For the second question, instead of optimizing two models with distillation, we simply twist the training step in the regular learning framework by running an additional forward with the input of transformed images, and the transformed flow from the first forward pass is treated as reliable supervision. Since the self-supervision from transformations avoids the unsupervised objective to be ambiguous in challenging scenes, our framework allows the network to learn by analogy with the original samples, and gradually mastering the ability to handle challenging samples.

In summary, our contributions are: *(i)* We propose a novel way to make use of the self-supervision signal from abundant augmentations for unsupervised optical flow by only training a single network; *(ii)* We demonstrate the applicability of our method for various augmentation methods. In addition to occlusion, we develop a general form for more challenging transformations. *(iii)* Our method leads in a leap of performance among deep unsupervised methods. It also achieves a comparable performance w.r.t. previous supervised methods, but with much fewer parameters and excellent cross dataset generalization capability.

2. Related Work

Supervised Optical Flow. Starting from FlowNet [7], various networks for optical flow with supervised learning have been proposed, *e.g*. FlowNet2 [15], PWC-Net [38], IRR-PWC [14]. These methods are comparable in accuracy to well-designed variational methods [36, 25], and are more effective during inference. However, the success of supervised methods heavily dependent on the large scale synthetic datasets [26, 7], which leads to an underlying degradation when transferring to real-world applications. As an alternative, we dig into the unsupervised method to alleviate the need for ground truth of dense optical flow.

Unsupervised Optical Flow. Yu *et al*. [18] first introduced a method for learning optical flow with brightness constancy and motion smoothness, which is similar to the energy minimization in conventional methods. Further researches improve accuracy through occlusion reasoning [42, 27], multi-frame extension [17, 11], epipolar constraint [50], 3D geometrical constraints with monocular depth [53, 49, 34] and stereo depth [41, 22]. Although these methods have become complicated, there is still a large gap with state-of-the-art supervised methods. Recent works improve the performance by learning the flow of occluded pixels in a knowledge distillation manner [23, 24], while the two-stage training in these works is trivial. Instead of studying the complicated geometrical constraints, our approach focuses on the basic training strategy. It generalizes the case of occlusion distillation to more kinds of challenging scenes with a straightforward single-stage learning framework.

Learning with Augmentation. Data augmentation is one of the easiest ways to improve training. Recently, there has been something new about integrating augmentation into the learning frameworks. Mounsaveng *et al*. [29] and Xiao *et al*. [45] suggested learning data augmentation with a spatial transformer network [16] to generate more complex samples. Xie *et al*. [46] proposed to use augmentation in the semi-supervised tasks by consistency training. Peng *et al*. [33] introduced to optimize data augmentation with the training of task-specific networks jointly. As a new trend in AutoML, several efforts to automatically search for the best policy of augmentations [5, 12, 21] are proposed. All these methods aimed at supervised or semi-supervised learning. In this work, we present a simple yet effective approach to integrate abundant augmentations with unsupervised optical flow. We propose to use reliable predictions of original samples as a self-supervision signal to guide the predictions of augmented samples.

3. Preliminaries

This work aims to learn optical flow from images without the need for ground truth. For completeness, we first briefly introduce the general framework for unsupervised optical flow methods, which is shown in the left part of Fig. 2.

Given a dataset of image sequences $\mathcal{I}$, our goal is to train a network $f(.)$ to predict dense optical flow $\mathbf{U}_{12}$ for two consecutive RGB frames $\{\mathbf{I}_1, \mathbf{I}_2\} \in \mathcal{I}$,

$$\mathbf{U}_{12} = f(\mathbf{I}_1, \mathbf{I}_2; \Theta), \tag{1}$$

where Θ is the set of learnable parameters in the network.

Despite the lack of direct supervision from ground truth, the network can be trained implicitly with view synthesis. Specifically, image $\mathbf{I}_2$ can be warped to synthesize the view of $\mathbf{I}_1$ with the prediction of optical flow $\mathbf{U}_{12}$,

$$\hat{\mathbf{I}}_1(\mathbf{p}) = \mathbf{I}_2\left(\mathbf{p} + \mathbf{U}_{12}(\mathbf{p})\right), \tag{2}$$

where $\mathbf{p}$ denotes pixel coordinates in the image, and bilinear sampling is used for the continuous coordinates. Then, the objective of view synthesis, also known as photometric loss $\mathcal{L}_{\text{ph}}$, can be formulated as:

$$\mathcal{L}_{\text{ph}} \sim \sum_{\mathbf{p}} \rho(\hat{\mathbf{I}}(\Theta), \mathbf{I}), \tag{3}$$

where $\rho(.)$ is a pixel-wise similarity measurement, *e.g.* ℓ_1 distance or structural similarities (SSIM).

Nevertheless, the photometric loss is violated when pixels are occluded or moved out of view so that there are no corresponding pixels in $\mathbf{I}_2$. As a common practice in [27, 40], we denote these pixels by a binary occlusion map $\mathbf{O}_{12}$. This map is obtained by the classical forward-backward checking method, where the backward flow is estimated by swapping the order of input images. The photometric loss in the occluded region will be discarded.

Furthermore, supervision solely based on the photometric loss is ambiguous for somewhere textureless or with repetitive patterns. One of the most common ways to reduce ambiguity is named smooth regularization,

$$\mathcal{L}_{\text{sm}} \sim \sum_{d \in x,y} \sum_{\mathbf{p}} \left\|\nabla_d \mathbf{U}_{12}\right\|_1 e^{-|\nabla_d \mathbf{I}|}, \tag{4}$$

which constrains the prediction similar to the neighbors in x and y directions when no significant image gradient exists.

4. Method

Since the general pipeline suffers from unreliable supervision for challenging cases, previous unsupervised works avoid using heavy augmentations. In this section, we introduce a novel framework to reuse existing heavy augmentations that have been proven effective in the supervised scenario, but with different forms. The pipeline is shown in Fig. 2, and we will explain in detail next.

4.1. Augmentation as a Regularization

Formally, we define an augmentation parameterized by a random vector θ as $\mathcal{T}_\theta^{\text{img}} : \mathbf{I}_t \mapsto \bar{\mathbf{I}}_t$, from which one can sample augmented images $\{\bar{\mathbf{I}}_1, \bar{\mathbf{I}}_2\}$ based on original images $\{\mathbf{I}_1, \mathbf{I}_2\}$ in the dataset. In the general pipeline, the network is trained with the data sampled from the augmented dataset. In contrast, we train the network on original data, but leverage augmented samples as a regularization.

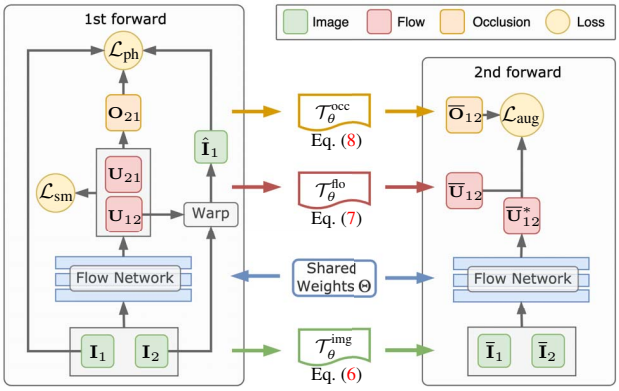

Figure 2. The pipeline of our proposed method. A complete training step includes two forwards: *(i)* The left side shows the first forward with original samples by the regular pipeline introduced in Section 3. Then, we perform transformations on images, predicted flow, and occlusion map respectively to construct an augmented sample. *(ii)* The right side shows an additional forward with the input of transformed images, and the output flow is supervised by the flow prediction of original samples.

More specifically, after a regular forward pass for original images, we additionally run another forward for transformed images to predict the optical flow $\overline{\mathbf{U}}_{12}^*$. Meanwhile, the prediction of optical flow in the first forward is transformed consistently by $\mathcal{T}_\theta^{\text{flo}} : \mathbf{U}_{12} \mapsto \overline{\mathbf{U}}_{12}$.

The basic assumption of our method is that augmentation brings challenging scenes in which the unsupervised loss will be unreliable, while the transformed predictions of original data can provide reliable self-supervision. Therefore, we optimize the consistency for the transformed samples instead of the objective of view synthesis. We follow the generalized Charbonnier function that commonly used in the supervised learning of optical flow as:

$$\mathcal{L}_{\text{aug}} \sim \sum_{\mathbf{p}} \left(\left|\mathcal{S}\left(\overline{\mathbf{U}}_{12}(\mathbf{p})\right) - \overline{\mathbf{U}}_{12}^*(\mathbf{p})\right| + \epsilon\right)^q, \tag{5}$$

where $\mathcal{S}(.)$ stands for stop-gradient, and the same setting as supervised work [38] with $q = 0.4$ and $\epsilon = 0.01$ gives less penalty to outliers. For stability, we stop the gradients of $\mathcal{L}_{\text{aug}}$ propagating to the transformed original flow $\overline{\mathbf{U}}_{12}$. Also, only the loss in the non-occluded region is considered. After twice forwarding, the photometric loss Eq. (3), the smooth regularization Eq. (4), and the augmentation regularization Eq. (5) are backward at once to update the model.

Our learning framework can be integrated with almost all types of augmentation methods. In the following, we summarize three kinds of transformations, which compose the common augmentations for the optical flow task. Some examples are shown in Fig. 3.

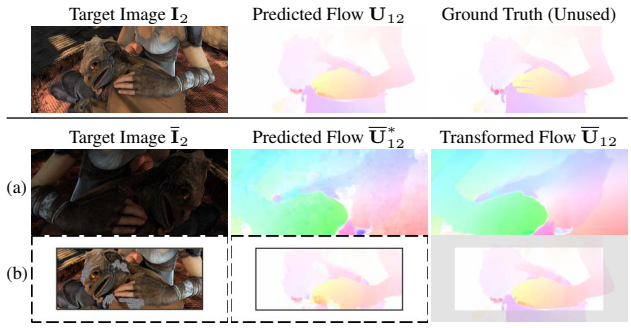

Figure 3. Some examples of the main idea. The same network is used to predict the optical flow of original images and transformed images, respectively. (a) Spatial transformation and appearance transformation generate a scene with large displacement and low brightness. (b) Occlusion transformation introduces additional occlusions. The pseudo label $\overline{\mathbf{U}}_{12}$ that transformed from the original predictions $\mathbf{U}_{12}$ can provide reliable supervision.

Spatial Transformation. We assume the transformation that results in a change in the location of pixels is called spatial transformation, which includes random crop, flip, zoom, affine transform, or more complicated transformations such as thin-plate-spline or CPAB transformations [8].

Here we show a general form for these transformations. Let τ_θ be a transformation of pixel coordinates. The transformation of image $\mathcal{T}^{\text{img}}_\theta : \mathbf{I}_t \mapsto \bar{\mathbf{I}}_t$ can be formulated as:

$$\bar{\mathbf{I}}_t(\mathbf{p}) = \mathbf{I}_t\left(\tau_\theta(\mathbf{p})\right), \tag{6}$$

which can be implemented by a differentiable warping process, same as the one used in Eq. (2).

Since changing pixel locations will lead to a change in optical flow, we should warp on an intermediate flow field $\widetilde{\mathbf{U}}_{12}$ instead of the original flow. The transformation of optical flow is $\mathcal{T}^{\text{flo}}_\theta : \mathbf{U}_{12} \mapsto \overline{\mathbf{U}}_{12}$ can be formulated as:

$$\begin{cases} \widetilde{\mathbf{U}}_{12}(\mathbf{p}) = \tau_\theta\left(\mathbf{p} + \mathbf{U}_{12}(\mathbf{p})\right) - \tau_\theta\left(\mathbf{p}\right), \\ \overline{\mathbf{U}}_{12}(\mathbf{p}) = \widetilde{\mathbf{U}}\left(\tau_\theta\left(\mathbf{p}\right)\right). \end{cases} \tag{7}$$

Additionally, the spatial transformation brings new occlusions. As we mentioned above, we explicitly reasoning occlusion from the predictions of bi-directional optical flow. Since predictions of transformed samples are noisy, we infer the transformed occlusion map from original predictions instead. The transformation $\mathcal{T}^{\text{occ}}_\theta : \mathbf{O}_{12} \mapsto \overline{\mathbf{O}}_{12}$ consists of two parts: the old occlusion $\overline{\mathbf{O}}^{\text{old}}_{12}(\mathbf{p})$ in the new view and the new occlusion $\overline{\mathbf{O}}^{\text{new}}_{12}(\mathbf{p})$ for pixels whose correspondences are out of the boundary Ω. The former can be obtained by the same warping process as $\mathcal{T}^{\text{img}}_\theta$ but with nearest-neighbor interpolation, and the latter can be explicitly estimated from the flow $\overline{\mathbf{U}}_{12}$ by checking the boundary:

$$\overline{\mathbf{O}}^{\text{new}}_{12}(\mathbf{p}) = \left(\mathbf{p} + \overline{\mathbf{U}}_{12}(\mathbf{p})\right) \notin \Omega. \tag{8}$$

The final transformed occlusion $\overline{\mathbf{O}}_{12}$ is a union of these two parts. Note that, the non-occluded pixels in $\overline{\mathbf{O}}^{\text{old}}_{12}$ might be occluded in $\overline{\mathbf{O}}^{\text{new}}_{12}$. It provides an effective way to learn the optical flow in occluded regions. For stability, only the non-occluded pixels in $\overline{\mathbf{O}}^{\text{old}}_{12}$ contribute to the loss $\mathcal{L}_{\text{aug}}$.

Besides, since we formulate the spatial transformation as a warping process, there might be pixels out of boundary after transformation. The common solution, such as padding with zero or the value of boundary pixels, will lead to severe artifacts. Therefore, we repeat sampling the transformations until all transformed pixels are in the region of the original view. On the other hand, this strategy increases the displacement of the pixel in general.

Occlusion Transformation. The spatial transformation provides reliable supervision for the flow with large displacement or occlusion around the boundary. As a complementary, recent work [23, 24] proposed to learn optical flow in arbitrary occluded regions with knowledge distillation. The general learning process of these methods consists of training a teacher model, offline creating occluded samples, and distilling to a student model. We argue that the way of model distillation is too trivial, and there is a performance bottleneck due to the frozen teacher model.

We integrate the occlusion hallucination into our one-stage training framework and named as occlusion transformation. Specifically, there are two steps: *(i)* Random crop. Actually, random crop is a kind of spatial transformation, but it efficiently creates new occlusion in the boundary. We crop the pair of images as a preprocess of occlusion transformation. *(ii)* Random mask out. We randomly mask out some superpixels in the target images with Gaussian noise, which will introduce new occlusion for the source image.

Note that, we adopt a strategy consistent with the spatial transformation that only the pixels not occluded in $\overline{\mathbf{O}}^{\text{old}}_{12}$ contribute to $\mathcal{L}_{\text{aug}}$. It is different from the previous distillation works, in which they reasoning a new occlusion map from the noisy prediction of transform images. Besides, in order to avoid creating transformed samples offline, we adopt a fast method of superpixel segmentation similar to [35]. The occlusion transformation in our framework simplifies the way of model distillation by optimizing a single model in one-stage with end-to-end learning.

Appearance Transformation. More transformations only change the appearance of images, such as random color jitter, random brightness, random blur, random noise. As a relatively simple case, appearance transformation does not change the location of pixels, nor introduce new occlusion. Still, the transformations lead to a risk for general methods, *e.g.* the photometric loss is meaningless when the image is overexposed, blurred, or in extremely low light. Instead, our method can exploit these transformations since the pre-

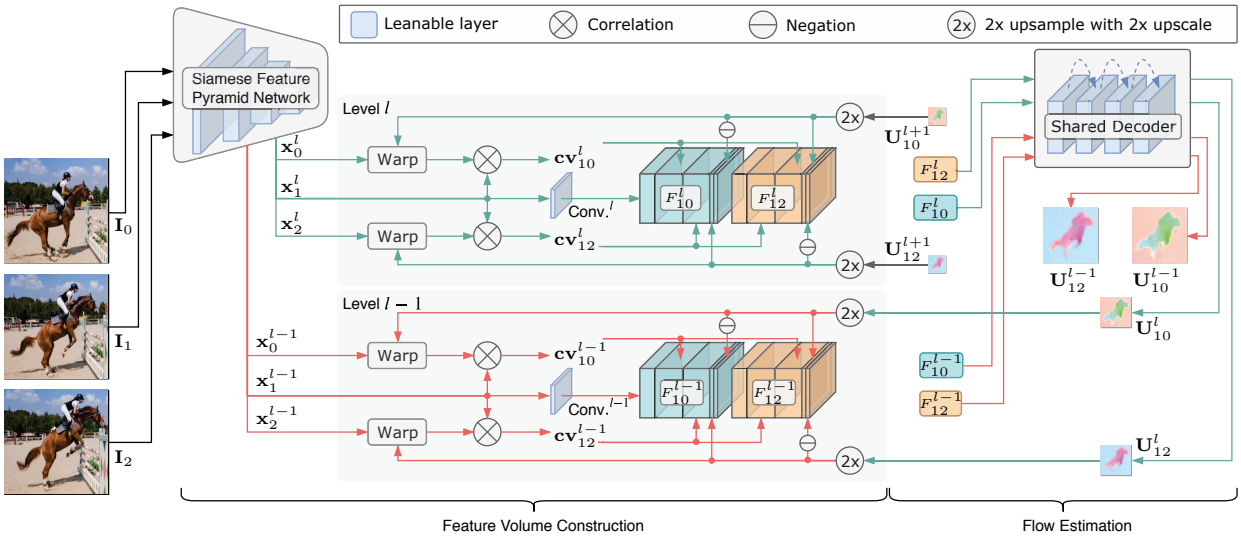

Figure 4. Network architecture for our lightweight multi-frame extension of PWC-Net [38]. It shares a semi-dense flow decoder for all of the levels across the pyramid with both forward flow and backward flow. For simplicity and completeness, the pipeline of two levels in the feature pyramid is displayed. Different line colors represent different levels of the process.

diction of the original sample provides a way to learn the optical flow in challenging transformed scenes.

4.2. Overall Objective and Convergence Analysis

Our framework assumes that transformed predictions are generally more accurate than the predictions of transformed samples, but what if samples are in the opposite case? In fact, we ensure convergence with the scope of each loss, *i.e.*, which pixels affect each loss.

As shown in Fig. 2, the overall objective for a training step consists of three loss terms in twice forwarding,

$$\mathcal{L}_{\text{all}} = \underbrace{\mathcal{L}_{\text{ph}}\left(\mathbf{U}_{12}\right) + \lambda_1 \mathcal{L}_{\text{sm}}\left(\mathbf{U}_{12}\right)}_{\text{1st forward}} + \underbrace{\lambda_2 \mathcal{L}_{\text{aug}}(\mathcal{S}(\overline{\mathbf{U}}_{12}), \overline{\mathbf{U}}^{*}_{12})}_{\text{2nd forward}}, \quad (9)$$

in which the first two terms propagate gradients for the original sample, and the last term is for the transformed sample. The original data and the augmented data are treated differently. By setting a minor weight of λ_2, we can ensure that the original data is always dominant, so the effects of bad cases are limited. Moreover, the scope of the photometric loss $\mathcal{L}_{\text{ph}}$ is the non-occluded pixels in $\mathbf{O}_{12}$. Thus the augmentation consistency loss becomes dominant for the new occluded pixels, which leads the network to learn the optical flow with occlusion effectively. Besides, the scope of augmentation loss $\mathcal{L}_{\text{aug}}$ avoids the network to be misguided from the original occluded predictions.

4.3. Lightweight Network Architecture

The learning framework we proposed can be applied to any flow networks. However, optical flow often plays a role as a sub-module in high-level video tasks [47, 48, 31] where the model size should be concerned. Hence, we introduce a lightweight architecture and extend it to multiple frames.

We start from a well-known network for the optical flow task named PWC-Net [38]. The original network shares a feature encoder with a siamese feature pyramid network for the images. For the level l in the pyramid, the feature maps of target image $\mathbf{x}_2^l$ are aligned by warping operation with the flow prediction $\mathbf{U}_{12}^{l+1}$ from the higher level. Then the cost volume $\mathbf{cv}_{12}^l$ is constructed with correlation operation. The input for flow decoder F_{12}^l is organized by concatenating the feature maps of source image $\mathbf{x}_1^l$, the upsampled flow from the higher level $\mathbf{U}_{12}^{l+1}$, and the cost volume $\mathbf{cv}_{12}^l$. Finally, the specific flow decoder of level l predicts the optical flow $\mathbf{U}_{12}^l$. By iterating over the pyramid, the network predicts optical flow at different scales.

Our method follows the main pipeline of the original PWC-Net but with some modifications. The flowchart of our multi-frame extension is shown in Fig. 4. We notice that the majority of learnable parameters of PWC-Net is in the flow decoder of each feature level, so we take several steps to reduce the parameters: *(i)* The original implementation adopts a fully dense connection in each decoder, while we reduce the connections that only connections in the nearest two layers are retained. *(ii)* We share the flow decoder for all of the levels across the pyramid, with an additional convolution layer for each level to align the feature maps. *(iii)* We extend the model to multiple frames by repeating the warping and correlation to the backward features. The flow decoder is shared for both forward flow and backward flow in the multi-frame extension by changing the sign of optical flow and the order in feature concatenation.

5. Experimental Results

5.1. Implementation Details

We implement our end-to-end approach in PyTorch [32]. All models are trained by Adam optimizer [19] with $\beta_1 = 0.9$, $\beta_2 = 0.99$, batch size of 4. The learning rate is 10^{-4} without adjustment during training. The loss weights for regularizations are set to $\lambda_1 = 60$ and $\lambda_2 = 0.01$ for all datasets. In addition, an optional pre-training can be used for better results, which is under almost the same setting above, but with $\lambda_2 = 0$, *i.e.* a regular training step without the transformed pass in forward [1].

Only random flip and random time order switch are performed as the regular data augmentation. The heavy combination of augmentations in supervised works [15, 38, 13] are used as the appearance transformation and spatial transformation in our framework, including random rotate, translate, zoom in, as well as additive Gaussian noise, Gaussian blur and random jitter in brightness, color, and contrast.

5.2. Datasets

We first evaluate our method on three well-established optical flow benchmarks, MPI Sintel [1], KITTI 2012 [10], and KITTI 2015 [28]. Then, we conduct a cross dataset experiment with another optical flow dataset FlyingChairs [7] and a segmentation dataset CityScapes [4].

We follow a similar data setting in previous unsupervised works [23, 24]. For the MPI Sintel benchmark, we extract all frames from the raw movie and manually group frames by shots for pre-training, which consists of 14,570 image pairs. Then, the model is fine-tuned on the standard training set, which provides 1,041 image pairs with two different rendering passes ("Clean" and "Final"). For the KITTI 2012 and KITTI 2015, we pre-train the model on the KITTI raw dataset [9], but discard scenes that contain images appeared in the optical flow benchmarks. The pre-training set consists of 28,058 image pairs. Then the model is fine-tuned on the multi-view extension data, but discards samples containing frames related to validation, *i.e.* numbers 9-12. The final training set consists of 6,000 samples for our basic model and 3,600 samples for the multi-frame model.

5.3. Comparison with State-of-the-art

We compare our method with both supervised and unsupervised methods on optical flow benchmarks. Standard metrics for optical flow are used, including average endpoint error (AEPE), and percentage of erroneous pixels (Fl).

Table 1 reports the results on MPI Sintel benchmark. Our basic two-frame model "ARFlow" outperforms all previous unsupervised works with the least parameters. Furthermore, our multi-frame model "ARFlow-MV" reduces the previous best AEPE from 6.18 [23] to 4.49 on the clean pass, with 27.3% improvement, and from 6.57 [24] to 5.67 on the final pass, with 13.7% improvement.

[1]Code available at https://github.com/lliuz/ARFlow.

	Method	Sintel Training		Sintel Test		# Param.
		Clean	Final	Clean	Final	
Supervised	FlowNetS-ft [7]	(3.66)	(4.44)	6.96	7.76	32.07 M
	LiteFlowNet-ft[13]	**(1.64)**	(2.23)	4.86	6.09	5.37 M
	PWC-Net-ft[38]	(2.02)	(2.08)	4.39	5.04	8.75 M
	IRR-PWC-ft [14]	(1.92)	(2.51)	3.84	4.58	6.36 M
	SelFlow-ft† [24]	(1.68)	**(1.77)**	**3.74**	**4.26**	**4.79 M**
Unsupervised	UnFlow-CSS [27]	-	(7.91)	9.38	10.22	116.58 M
	OccAwareFlow [42]	(4.03)	(5.95)	7.95	9.15	5.12 M
	MFOccFlow† [17]	(3.89)	(5.52)	7.23	8.81	12.21 M
	EpiFlow train-ft [50]	(3.54)	(4.99)	7.00	8.51	8.75 M
	DDFlow [23]	(2.92)	(3.98)	6.18	7.40	4.27 M
	SelFlow† [24]	(2.88)	(3.87)	6.56	6.57	4.79 M
	Ours (ARFlow)	(2.79)	(3.73)	4.78	5.89	**2.24 M**
	Ours (ARFlow-MV†)	**(2.73)**	**(3.69)**	**4.49**	**5.67**	2.37 M

Table 1. **MPI Sintel Flow**: AEPE and the number of CNN parameters are reported. Missing entry (-) means that the results are not reported for the respective method, and † indicates the model using more than two frames.

	Method	KITTI 2012		KITTI 2015	
		training	test	training	test (F1)
Supervised	FlowNet2-ft [15]	(1.28)	1.8	(2.30)	11.48%
	LiteFlowNet-ft [13]	(1.26)	1.7	(2.16)	11.48%
	PWC-Net-ft [38]	(1.45)	1.7	(2.16)	9.60%
	SelFlow-ft† [24]	**(0.76)**	**1.5**	**(1.18)**	**8.42%**
Unsupervised	BridgeDepthFlow§ [20]	2.56	–	7.02	–
	CCFlow§ [34]	–	–	5.66	25.27%
	UnOS-stereo§ [41]	1.64	1.8	5.58	18.00%
	EpiFlow-train-ft§ [50]	(2.51)	3.4	(5.55)	16.95%
	DDFlow [23]	2.35	3.0	5.72	14.29%
	SelFlow† [24]	1.69	2.2	4.84	14.19%
	Ours (ARFlow)	1.44	1.8	**2.85**	11.80%
	Ours (ARFlow-MV†)	**1.26**	**1.5**	3.46	**11.79%**

Table 2. **KITTI Optical Flow 2012 and 2015**: AEPE and Fl are reported. For unsupervised methods, only the works published in 2019 are shown. Missing entry (-) means that the results are not reported for the respective method. † indicates the model using more than two frames. § indicates training with geometrical constraints.

As for KITTI benchmarks, Table 2 shows a significant improvement. On the training set, we achieve AEPE=1.26 with 25.4% relative improvement on KITTI 2012 and AEPE=2.85 with 41.2% improvement on KITTI 2015 w.r.t. the previous best unsupervised method [24]. On the test set, our method reaches the best AEPE=1.5 and F1-all=11.79% among unsupervised methods, respectively.

Several representative supervised methods are also reported as a reference. As a result, our unsupervised models firstly reach or approach some powerful fully supervised methods such as LiteFlowNet [13], PWC-Net [38], even with 27.1% parameters of PWC-Net.

Samples on MPI Sintel and KITTI are shown in Fig. 5. Compared with the state-of-the-art competitor [24], for the low light and large displacement scenes in MPI Sintel, our method maintains better performance in general and is more

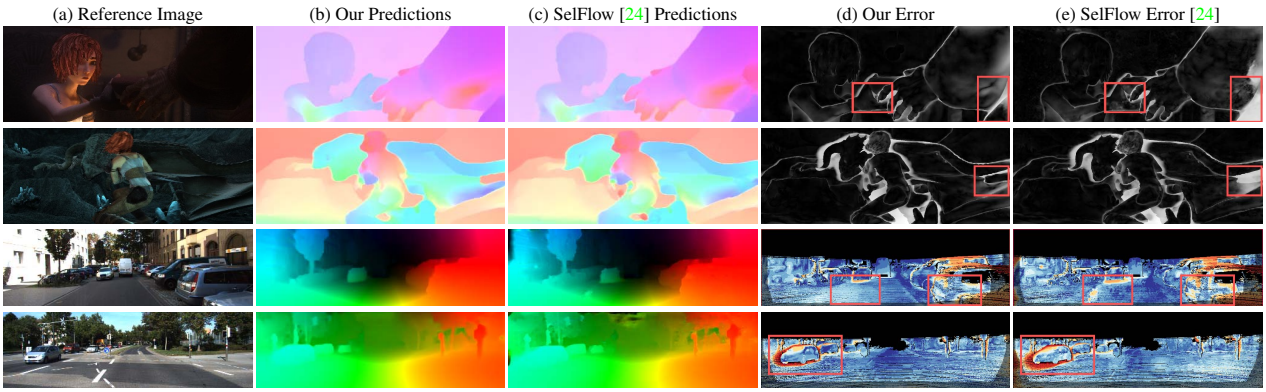

Figure 5. Qualitative visualization comparing with unsupervised SelFlow [24]. The first two rows are from the Sintel Final pass, where the errors are visualized in gray. The last two rows are from KITTI 2015, in which the correct predictions are depicted in blue and the wrongs in red for the error visualization. More samples will be available on the website of corresponding benchmarks.

Model Architecture	AR	Sintel Clean			Sintel Final			# Param.
		ALL	NOC	OCC	ALL	NOC	OCC	
PWC-Net [38]		2.48	1.19	21.71	3.47	1.98	25.19	8.75 M
PWC-Net-small [38]		2.76	1.28	23.92	3.62	2.16	28.15	4.05 M
+ Reduce Dense		2.53	1.23	21.36	3.47	2.03	24.20	5.32 M
	✓	2.04	0.90	18.47	2.97	1.72	21.05	
+ Share Decoder		2.30	1.08	20.00	3.19	1.84	22.77	2.24 M
	✓	1.95	0.85	17.85	2.86	1.66	20.25	
+ Multipe Frames		2.24	1.04	19.60	3.18	1.86	22.36	2.37 M
	✓	1.89	0.86	16.79	2.85	1.66	20.02	

Table 3. **Ablation study of our learning framework with multiple model architectures.** AEPE in specific regions of the scene and the number of CNN parameters are reported. **AR**: Training with augmentation as a regularization framework.

	ST	AT	OT	ALL	NOC	OCC	s0-10	s10-40	s40+
Sintel Clean				2.53	1.23	21.36	0.61	2.74	24.30
			✓	2.39	1.20	19.61	0.61	2.56	23.14
		✓		2.40	1.13	20.67	0.61	2.81	21.95
	✓			2.14	1.00	18.63	0.62	2.83	17.56
	✓	✓		2.09	0.95	18.90	**0.59**	2.65	18.03
	✓	✓	✓	**2.04**	**0.90**	**18.47**	0.61	**2.55**	**17.05**
Sintel Final				3.47	2.03	24.20	0.82	3.77	33.48
			✓	3.23	1.93	21.98	0.82	3.48	30.78
		✓		3.36	1.94	23.95	0.81	3.70	32.17
	✓			3.04	1.78	21.25	0.78	3.55	27.80
	✓	✓		3.01	1.76	21.40	**0.75**	3.48	28.48
	✓	✓	✓	**2.97**	**1.72**	**21.05**	0.77	**3.40**	**27.25**

Table 4. **Comparison of combinations of transformations.** AEPE in specific regions are reported. **ST**: Spatial transformation, **AT**: Appearance transformation, **OT**: Occlusion transformation.

accurate around the boundaries. For KITTI results, the shapes in our optical flow are more structured for objects and more accurate in texture-less regions.

5.4. Ablation Study

To further analyze the capability of each component, we conduct four groups of ablation studies. We randomly re-split the Sintel training set into a new training set and a validation set by the scene. We evaluate AEPE in different regions over all pixels (ALL), non-occluded pixels (NOC), occluded pixels (OCC), and according to speed (s0-10, s10-40 and s40+ are pixels that move less than 10 pixels, between 10 and 40, and more than 40, respectively)

Main Ablation. Table 3 assesses the overall improvement of our augmentation as a regularization learning framework under multiple model architectures. Our framework consistently improves the accuracy of optical flow over 10% for all architectures, whether for occluded or non-occluded pixels.

For the consideration of the number of model parameters, we start from the original PWC-Net and a variant named PWC-Net-small without dense connections in the flow decoders [38]. Although removing dense connections can reduce half parameters, it leads to severe performance degradation. In contrast, our reduced dense variant maintains the performance while reducing 39.2% parameters. Sharing decoder across feature pyramid yields an improvement on flow with only 25.6% parameters of the original model. The multi-frame extension reaches the best performance with the minimal extra overhead of parameters.

Combination of Transformations. Furthermore, we delve into the type of transformations in our framework. Table 4 shows the performance of the model trained with several combinations of the three kinds of transformations. There are some critical observations: *(i)* Each transformation can improve the performance individually. *(ii)* Spatial transformation is the most helpful to all measurements, especially for large displacement estimation. *(iii)* The accuracy in the occluded region can be significantly improved by occlusion transformation or spatial transformation. All these observations are consistent with our assumption that the transformation will introduce new challenging scenes, and our approach can provide reliable supervision.

Usage of Augmentation. As we mentioned above, almost all of the unsupervised learning approaches avoid using a heavy combination of augmentations. As a reference, we

Method	Sintel Clean				Sintel Final			
	ALL	s0-10	s10-40	s40+	ALL	s0-10	s10-40	s40+
Without Aug.	2.53	**0.61**	2.74	24.30	3.47	0.82	3.77	33.48
Aug. Directly	2.71	0.69	3.11	27.13	3.80	0.95	4.03	35.90
Aug. Distillation	2.36	0.64	2.61	19.90	3.31	0.86	3.50	30.18
Ours(aug. as reg.)	**2.04**	**0.61**	**2.55**	**17.05**	**2.97**	**0.77**	**3.40**	**27.25**

Table 5. Comparison of our learning framework with direct data augmentation and the data distillation framework used in [23, 24].

Method	Sintel Clean				Sintel Final			
	ALL	s0-10	s10-40	s40+	ALL	s0-10	s10-40	s40+
Without Aug.	2.53	0.61	2.74	24.30	3.47	0.82	3.77	33.48
CPAB [8] + AT	2.38	0.61	2.78	21.60	3.32	0.81	3.59	31.09
AutoAugment [5]	2.30	0.62	2.59	21.18	3.29	0.81	3.53	30.11
Ours(ST + AT)	**2.09**	**0.59**	**2.65**	**18.03**	**3.01**	**0.75**	**3.48**	**28.48**

Table 6. Comparison of different augmentation transformations integrated with our framework. AT: appearance transformation, ST: spatial transformation.

evaluate the same transformations with different usages. Table 5 reports the results of *(i)* training without heavy augmentation, *(ii)* using transformation as a regular data augmentation and training directly, *(iii)* training with data distillation that similar in [23, 24], *(iv)* training with the learning framework we proposed. The results show that directly augmentation makes all metrics worse. Instead of applying transformations directly, distillation alleviates the problem of unreliable supervision. However, the frozen teacher model is still a bottleneck for the student model. Also, the tedious multi-stage training process of knowledge distillation is undesired. Our framework avoids the unreliable photometric loss for the transformed samples. It achieves the best results with a single-stage optimization.

Integrate Complicated Augmentation. By implementing the corresponding transformation of optical flow and occlusion map, our framework can be integrated with almost all types of augmentation. We assess a complicated spatial transformation called CPAB [8] and a recent work in AutoML on searching for the best augmentation policy called AutoAugment [5]. Note that random zoom in is applied first to avoid invalid coordinate values of transformations. Table 6 shows that both strategies integrated with our framework can improve accuracy. Note that AutoAugment is too time consuming for our task, therefore we adopt the final policy searched from ImageNet [6] classification task. It is promising that our framework with AutoAugment will be further improved with policy fine-tuning.

5.5. Cross Dataset Generalization

Although deep optical flow methods have been far ahead of the most popular classical variational method TV-L1 [43] on optical flow benchmarks, the latter has not gone away. One possible reason is that supervised learning methods are prone to overfitting, which results in poor generalization when transferring to high-level video tasks.

Method	Training Set	Chairs Full	Sintel Clean	Sintel Final	KIITI 2012	KITTI 2015
PWC-Net [38]	Sintel	3.69	**(1.86)**	**(2.31)**	3.68	10.52
Ours(ARFlow)	Sintel	**3.50**	(2.79)	(3.73)	3.06	9.04
	CityScapes	5.10	5.22	6.01	**2.11**	**5.33**

Table 7. Generalization performance of cross datasets evaluation. The numbers indicate AEPE on each dataset. For KITTI and Sintel, the results are evaluated on the training set. () indicates the results of a dataset that the method has been trained on.

Hence, we report the cross dataset accuracy in Table 7, in which our unsupervised method is compared with a fully supervised method PWC-Net [38]. The supervised PWC-Net consistently outperforms for the dataset that the model is trained on, while our unsupervised method works much better when transferring to other datasets. In addition, we train a model on an urban street dataset named CityScapes [4], in which 50,625 image pairs are used for training without the ground truth. This model performs best on the KITTI 2012 and KITTI 2015 than any other model trained on the synthetic dataset. Our method makes it possible to fit the domain of high-level video tasks by training a model on the unlabeled videos from that domain.

Remarkably, despite the lack of cross dataset results from other unsupervised methods, the accuracy of our model trained on CityScapes is even better than most of the previous works trained on KITTI (*c.f*. Table 2), which shows the superiority of our method. The results demonstrate a significant improvement of our method for unsupervised optical flow task with an excellent generalization.

6. Conclusion

We proposed a novel framework that learns optical flow from unlabeled image sequences with the self-supervision from augmentations. To avoid the objective of view synthesis being unreliable on transformed data, we twist the basic learning framework by adding another forward pass for transformed images, where the supervision is from the transformed prediction of original images. Besides, a lightweight network and its multi-frame extension were presented. Extensive experiments have shown that our methods significantly improve accuracy, with high compatibility and generalization ability. We believe that our learning framework can be further combined with other geometrical constraints or transferred to other visual geometry tasks, such as depth or scene flow estimation.

Acknowledgment We thank anonymous reviewers for their constructive comments, and LL would like to thank Pengpeng Liu for helpful suggestions. This work is partially supported by the National Natural Science Foundation of China (NSFC) under Grant No. 61836015 and Key R&D Program Project of Zhejiang Province (2019C01004).

References

[1] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *European Conference on Computer Vision (ECCV)*, 2012. 6

[2] Dongdong Chen, Jing Liao, Lu Yuan, Nenghai Yu, and Gang Hua. Coherent online video style transfer. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[3] Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang. Segflow: Joint learning for video object segmentation and optical flow. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6, 8

[5] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 8

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 8

[7] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 1, 2, 6

[8] Oren Freifeld, Søren Hauberg, Kayhan Batmanghelich, and Jonn W Fisher. Transformations based on continuous piecewise-affine velocity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2017. 4, 8

[9] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 6

[10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 6

[11] Shuosen Guan, Haoxin Li, and Wei-Shi Zheng. Unsupervised learning for optical flow estimation using pyramid convolution lstm. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2019. 2

[12] Daniel Ho, Eric Liang, Ion Stoica, Pieter Abbeel, and Xi Chen. Population based augmentation: Efficient learning of augmentation policy schedules. In *International Conference on Machine Learning (ICML)*, 2019. 2

[13] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 6

[14] Junhwa Hur and Stefan Roth. Iterative residual refinement for joint optical flow and occlusion estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 6

[15] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2017. 1, 2, 6

[16] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015. 2

[17] Joel Janai, Fatma Guney, Anurag Ranjan, Michael Black, and Andreas Geiger. Unsupervised learning of multi-frame optical flow with occlusions. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 2, 6

[18] J Yu Jason, Adam W Harley, and Konstantinos G Derpanis. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In *European Conference on Computer Vision (ECCV)*, 2016. 2

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 6

[20] Hsueh-Ying Lai, Yi-Hsuan Tsai, and Wei-Chen Chiu. Bridging stereo matching and optical flow via spatiotemporal correspondence. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[21] Sungbin Lim, Ildoo Kim, Taesup Kim, Chiheon Kim, and Sungwoong Kim. Fast autoaugment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2

[22] Liang Liu, Guangyao Zhai, Wenlong Ye, and Yong Liu. Unsupervised learning of scene flow estimation fusing with local rigidity. In *International Joint Conference on Artificial Intelligence, (IJCAI)*, 2019. 1, 2

[23] Pengpeng Liu, Irwin King, Michael R Lyu, and Jia Xu. Ddflow: Learning optical flow with unlabeled data distillation. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019. 2, 4, 6, 8

[24] Pengpeng Liu, Michael Lyu, Irwin King, and Jia Xu. Selflow: Self-supervised learning of optical flow. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 4, 6, 7, 8

[25] Daniel Maurer and Andrés Bruhn. Proflow: Learning to predict optical flow. *British Machine Vision Conference (BMVC)*, 2018. 1, 2

[26] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2

[27] Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018. 1, 2, 3, 6

[28] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 6

[29] Saypraseuth Mounsaveng, David Vazquez, Ismail Ben Ayed, and Marco Pedersoli. Adversarial learning of general transformations for data augmentation. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[30] Michal Neoral, Jan Šochman, and Jiří Matas. Continual occlusion and optical flow estimation. In *Asian Conference on Computer Vision (ACCV)*, 2018. 1

[31] David Nilsson and Cristian Sminchisescu. Semantic video segmentation by gated recurrent flow propagation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 5

[32] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 6

[33] Xi Peng, Zhiqiang Tang, Fei Yang, Rogerio S Feris, and Dimitris Metaxas. Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[34] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 6

[35] C. Y Ren, V. A. Prisacariu, and I. D Reid. gSLICr: SLIC superpixels at over 250Hz. *ArXiv pre-prints:1509.04232*, 2015. 4

[36] Zhile Ren, Orazio Gallo, Deqing Sun, Ming-Hsuan Yang, Erik Sudderth, and Jan Kautz. A fusion approach for multi-frame optical flow estimation. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2019. 1, 2

[37] Zhe Ren, Junchi Yan, Bingbing Ni, Bin Liu, Xiaokang Yang, and Hongyuan Zha. Unsupervised deep learning for optical flow estimation. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2017. 1

[38] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2018. 1, 2, 3, 5, 6, 7, 8

[39] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Models matter, so does training: An empirical study of cnns for optical flow estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2019. 1

[40] Narayanan Sundaram, Thomas Brox, and Kurt Keutzer. Dense point trajectories by gpu-accelerated large displacement optical flow. In *European Conference on Computer Vision (ECCV)*, 2010. 3

[41] Yang Wang, Peng Wang, Zhenheng Yang, Chenxu Luo, Yi Yang, and Wei Xu. Unos: Unified unsupervised optical-flow and stereo-depth estimation by watching videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 6

[42] Yang Wang, Yi Yang, Zhenheng Yang, Liang Zhao, Peng Wang, and Wei Xu. Occlusion aware unsupervised learning of optical flow. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 6

[43] Andreas Wedel, Thomas Pock, Christopher Zach, Horst Bischof, and Daniel Cremers. An improved algorithm for tv-l 1 optical flow. In *Dagstuhl Motion Workshop*, 2009. 8

[44] Jonas Wulff, Laura Sevilla-Lara, and Michael J Black. Optical flow in mostly rigid scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[45] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[46] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. Unsupervised data augmentation for consistency training. *ArXiv pre-prints:1904.12848*, 2019. 2

[47] Rui Xu, Xiaoxiao Li, Bolei Zhou, and Chen Change Loy. Deep flow-guided video inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 5

[48] Yanchao Yang, Antonio Loquercio, Davide Scaramuzza, and Stefano Soatto. Unsupervised moving object detection via contextual information separation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 5

[49] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2

[50] Yiran Zhong, Pan Ji, Jianyuan Wang, Yuchao Dai, and Hongdong Li. Unsupervised deep epipolar flow for stationary or dynamic scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 6

[51] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[52] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. Deep feature flow for video recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[53] Yuliang Zou, Zelun Luo, and Jia-Bin Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *European Conference on Computer Vision (ECCV)*, 2018. 2