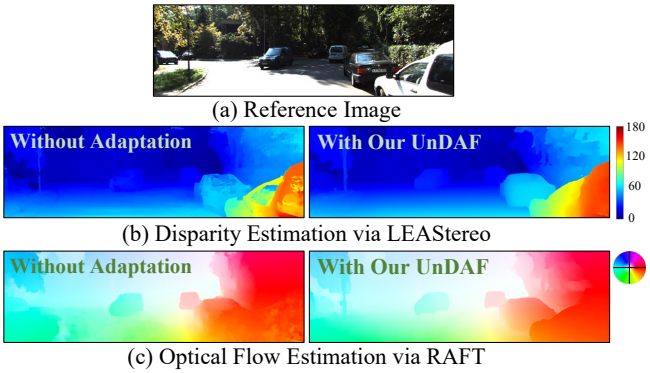
2022 IEEE International Conference on Robotics and Automation (ICRA)
May 23-27, 2022. Philadelphia, PA, USA

UnDAF: A General Unsupervised Domain Adaptation Framework for Disparity or Optical Flow Estimation

Hengli Wang, Rui Fan, Peide Cai, Ming Liu, and Lujia Wang

***Abstract*— Disparity and optical flow estimation are respectively 1D and 2D dense correspondence matching (DCM) tasks in nature. Unsupervised domain adaptation (UDA) is crucial for their success in new and unseen scenarios, enabling networks to draw inferences across different domains without manually-labeled ground truth. In this paper, we propose a general UDA framework (UnDAF) for disparity or optical flow estimation. Unlike existing approaches based on adversarial learning that suffers from pixel distortion and dense correspondence mismatch after domain alignment, our UnDAF adopts a straightforward but effective coarse-to-fine strategy, where a co-teaching strategy (two networks evolve by complementing each other) refines DCM estimations after Fourier transform initializes domain alignment. The simplicity of our approach makes it extremely easy to guide adaptation across different domains, or more practically, from synthetic to real-world domains. Extensive experiments carried out on the KITTI and MPI Sintel benchmarks demonstrate the accuracy and robustness of our UnDAF, advancing all other state-of-the-art UDA approaches for disparity or optical flow estimation. Our project page is available at https://sites.google.com/view/undaf.**



(a) Reference Image

(b) Disparity Estimation via LEAStereo

(c) Optical Flow Estimation via RAFT

Fig. 1: LEAStereo [3] for disparity estimation and RAFT [4] for optical flow estimation, which are both trained on a synthetic dataset and tested on a real-world dataset. It can be obviously seen that our UnDAF can significantly improve the DCM accuracy across different domains.

I. Introduction

Dense correspondence matching (DCM) is a fundamental task in computer vision. This technique has been prevalently applied in many robotics tasks, such as visual odometry [1] and object tracking [2]. The goal of DCM is to ascertain the relationship between each pair of pixels in two or more images of the same 3D scene [3], [4]. Specifically, disparity and optical flow estimation are respectively 1D and 2D DCM tasks in nature, which target at stereo images and consecutive video frames, separately.

With the evolution of artificial intelligence, deep learning has emerged as a highly practical and powerful technology for disparity [3], [5] and optical flow [4], [6] estimation, and the achieved results are very compelling. However, such data-driven approaches generally require a large amount of training data with human-labeled correspondence ground truth to learn the best model parameters for DCM tasks. This data labeling process is always very time-consuming and labor-intensive. Synthetic datasets with machine-labeled ground truth are easy-to-acquire, but in practice, it is still demanding to adapt a network from its learned scenario to a new and unseen one, especially from a synthetic domain to a real-world domain, as illustrated in Fig. 1. Therefore, domain adaptation has become a pressing need for various real-world DCM applications, especially for autonomous driving. Many supervised [3], [4] and unsupervised [7], [8] frameworks have been proposed in recent years. In this paper, we mainly explore the second direction and propose a general unsupervised domain adaptation framework (UnDAF) for DCM, as illustrated in Fig. 2, which requires no ground-truth labels in the target domain.

Existing unsupervised domain adaptation (UDA) approaches for DCM are typically developed under two different strategies: 1) designing an auxiliary loss for a specific DCM task so that the network can maintain the performance when adapting to the target domain [7], [8]; 2) training the network in the source domain, and then fine-tuning it in the target domain by minimizing unsupervised losses [9], [10]. However, the former typically requires complicated adversarial learning that suffers from pixel distortion and dense correspondence mismatch after domain alignment [8], while the latter always experiences efficiency degradation

This work was supported in part by the National Key R&D Program of China (Grant No. 2020AAA0108100). This work was also supported in part by Zhongshan Municipal Science and Technology Bureau Fund under Project ZSST21EG06, in part by Foshan-HKUST Industry-University-Research Cooperation Project under Project No. FSUST20-SHCIRI06C, and in part by Department of Science and Technology of Guangdong Province Fund under Project No. GDST20EG54, awarded to Prof. Ming Liu. *(Corresponding author: Lujia Wang.)*

Hengli Wang, Peide Cai, and Lujia Wang are with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong SAR, China (email: hwangdf@connect.ust.hk; pcaiaa@connect.ust.hk; eewanglj@ust.hk).

Ming Liu is with the Thrust of Robotics & Autonomous Systems, The Hong Kong University of Science and Technology (Guangzhou), Nansha, Guangzhou, 511400, Guangdong, China, and also with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong SAR, China (email: eelium@ust.hk).

Rui Fan is with the Department of Control Science and Engineering, College of Electronics and Information Engineering, Tongji University, Shanghai 201804, P. R. China (e-mail: rui.fan@ieee.org).

2022 IEEE International Conference on Robotics and Automation (ICRA) | 978-1-7281-9681-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICRA46639.2022.9811811

Authorized licensed use limited to: Univ of Calif San Diego. Downloaded on July 19,2022 at 00:22:02 UTC from IEEE Xplore. Restrictions apply.

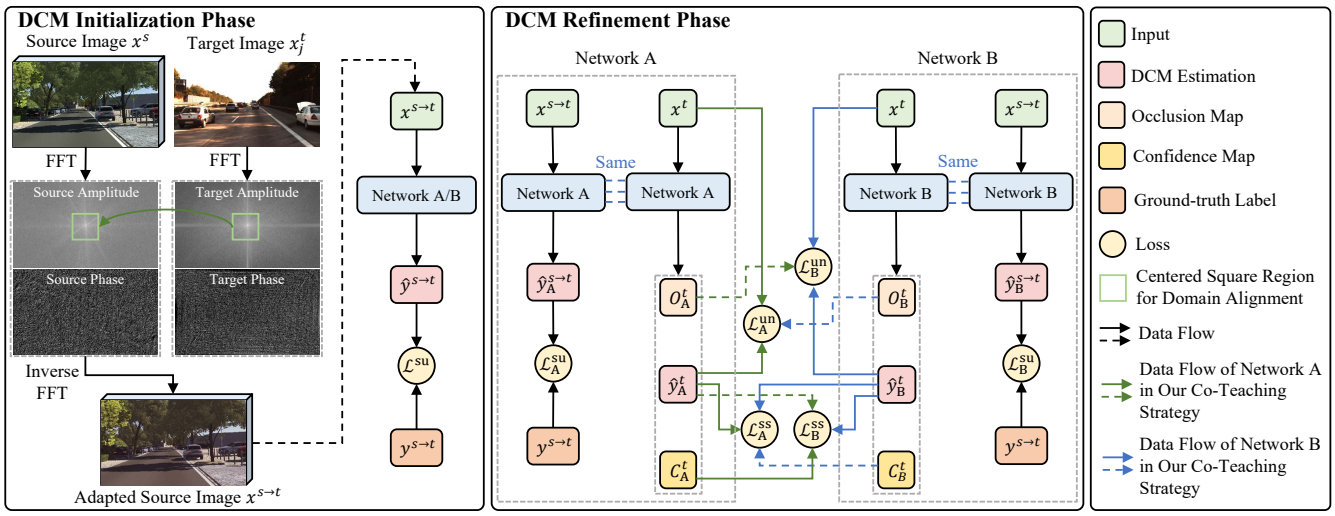


Fig. 2: An overview of our UnDAF, consisting of a DCM initialization phase and a DCM refinement phase. In the first phase, we employ Fourier transform for domain alignment and utilize the adapted source dataset $D^{s\to t}$ for coarse DCM initialization. In the second phase, we adopt a co-teaching strategy, which enables two networks to evolve by complementing each other for further DCM refinement.

in the fine-tuning phase, due to the catastrophic forgetting problem [11].

Excitedly, our UnDAF can address all these issues. Inspired by [12], we first employ Fourier transform for domain alignment and then adopt a co-teaching strategy for DCM refinement. Our UnDAF not only inherently preserves the DCM consistency between input images after domain alignment but also avoids the catastrophic forgetting problem. Moreover, our UnDAF avoids extra training beyond the primary DCM task, such as complicated adversarial learning. The simplicity of our approach makes it extremely easy to be combined with any existing supervised DCM network for unsupervised adaptation across different domains, or more practically, from synthetic to real-world domains.

To validate the effectiveness and robustness of our UnDAF, we conduct extensive experiments on the public benchmarks: 1) the KITTI Stereo 2012 [13] and 2015 [14] for disparity estimation; and 2) the KITTI Optical Flow 2012 [13] and 2015 [14], and the MPI Sintel [15] for optical flow estimation. Extensive experimental results demonstrate that our UnDAF outperforms all other state-of-the-art unsupervised domain adaptation approaches for disparity or optical flow estimation. The major contributions of this paper can be summarized as follows:

- We demonstrate that simply employing Fourier transform for domain alignment is effective for unsupervised domain adaptation in DCM.
- We develop a co-teaching strategy, which can effectively avoid error accumulation and improve the stability and accuracy of unsupervised domain adaptation for DCM.
- We present extensive experiments on the public benchmarks that demonstrate the state-of-the-art performance of our UnDAF.

II. Related Work

A. Dense Correspondence Matching

Traditional disparity estimation approaches generally employ local block matching or minimize a global energy function using Markov Random Field (MRF)-based optimization techniques [16]. Similarly, traditional optical flow estimation approaches typically minimize a global energy related to both brightness consistency and spatial smoothness [17].

With recent advances in deep learning, supervised approaches based on convolutional neural networks (CNNs) have achieved promising results for DCM. Specifically, these networks first employ CNNs to extract visual features, and then adopt a correlation layer to compute matching costs. Finally, several convolution layers are utilized to generate specific DCM results, such as disparity [3], [5] and optical flow [4], [6]. However, as aforementioned, these supervised approaches generally require a large amount of training data with human-labeled correspondence ground truth, and this data labeling process is always very time-consuming and labor-intensive. Differently, unsupervised approaches learn DCM without using correspondence ground truth. This is achieved by minimizing joint losses, which typically include a photometric loss and a smoothness loss [18], [19]. However, these unsupervised approaches can only achieve limited performance due to the lack of ground truth. Since synthetic datasets with machine-labeled ground truth are easy-to-acquire, unsupervised domain adaptation approaches that adapt a network from its learned synthetic domain to a real-world domain has become a promising direction for various real-world DCM applications. Therefore, in this paper, we embed existing supervised approaches into our UnDAF to achieve effective and efficient unsupervised domain adaptation for disparity or optical flow estimation.

Authorized licensed use limited to: Univ of Calif San Diego. Downloaded on July 19,2022 at 00:22:02 UTC from IEEE Xplore. Restrictions apply.

B. Unsupervised Domain Adaptation

UDA aims at reducing the gap between source and target domains. Based on the concept of adversarial learning introduced in unsupervised image-to-image translation [20], many UDA approaches utilize a discriminator to distinguish between source and target samples, and thus mitigating domain gaps [21], [22]. To apply UDA in DCM, many researchers have proposed to design an auxiliary loss for each specific DCM task, such as disparity [8] and optical flow [7] estimation. Other mainstream UDA approaches for DCM advocate training networks on the source domain, and then fine-tuning them on the target domain by minimizing unsupervised losses [9], [10]. However, as discussed above, these two categories of UDA approaches either require complicated adversarial learning that suffers from pixel distortion and dense correspondence mismatch after domain alignment [8] or perform adaptation inefficiently due to the catastrophic forgetting problem [11]. In contrast, our UnDAF not only inherently preserves the DCM consistency between input images after domain alignment but also avoids the catastrophic forgetting problem. Moreover, our UnDAF avoids extra training beyond the primary DCM task, such as complicated adversarial learning. Please note that in [12], Yang *et al.* adopted Fourier transform for UDA in semantic segmentation. Inspired by their work [12], we also adopt a similar Fourier transform technique for UDA in DCM. The difference between these two works is that semantic segmentation typically takes one image as input, while DCM tasks generally take two images as input. Therefore, one additional advantage of adopting this Fourier transform technique in our DCM task is that it can inherently preserve the DCM consistency between input images after domain alignment.

III. Methodology

Given a source dataset $D^s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ and a target dataset $D^t = \{x_i^t\}_{i=1}^{N_t}$, where x and y respectively denote the input images and ground-truth labels of a specific DCM task (disparity or optical flow estimation), our goal is to train a network to generate the corresponding DCM estimations $\hat{y}^t$ for the target domain. Specifically, $x_i = (x_l, x_r)_i$ denotes a pair of stereo images for disparity estimation; and $x_i = (x_t, x_{t+1})_i$ denotes two consecutive video frames for optical flow estimation. Since our UnDAF is a general framework designed for these two DCM tasks, we will mainly use x_i and y_i instead of their detailed components in the following content for notation convenience.

Fig. 2 provides the pipeline of our UnDAF, which consists of a DCM initialization phase and a DCM refinement phase. In the first phase, we employ Fourier transform to align source and target domains, and then train two identical networks initialized with different parameters on the adapted source dataset $D^{s\to t}$. In the second phase, we adopt a co-teaching strategy to evolve the two networks simultaneously for further DCM refinement.

A. DCM Initialization Phase

As mentioned above, training a network on D^s and fine-tuning it on D^t can lead to significant efficiency degradation due to the catastrophic forgetting problem [11]. Therefore, it is more effective to perform domain alignment from D^s to D^t before fine-tuning the network. It is observed in [12] that the semantic information can be retained when low-level spectrum (amplitude) changes significantly, and the low-level information is speculated to be the key to adaptation across different domains [12]. Therefore, we follow [12] and simply perform domain alignment between source and target domains based on the low-level spectral signals computed by Fourier transform, which inherently preserves the DCM consistency between input images after domain alignment. Such a paradigm also enables our UnDAF to avoid additional training process, such as complicated adversarial learning.

Specifically, we use $\mathcal{F} : \mathbb{R}^{H\times W\times 3} \to \mathbb{C}^{H\times W\times 3}$ to denote the Fourier transform of an RGB image x, which can be simply yielded using fast Fourier transform (FFT) algorithm [23] as follows:

$$\mathcal{F}(x)(u,v,c) = \sum_{h,w} x(h,w,c) e^{-j2\pi\left(\frac{h}{H}u + \frac{w}{W}v\right)}. \quad (1)$$

We also denote the amplitude and phase components of $\mathcal{F}$ as $\mathcal{F}^A : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^{H\times W\times 3}$ and $\mathcal{F}^P : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^{H\times W\times 3}$, respectively. $\mathcal{F}^{-1} : \mathbb{C}^{H\times W\times 3} \to \mathbb{R}^{H\times W\times 3}$ denotes the inverse Fourier transform, which can transform spectral signals back to images. Furthermore, we follow [12] and define a mask M_α of size $H \times W$, where all values are 0 except a centered square region with the side length of $\alpha \cdot \min(H, W)$ and $\alpha \in (0, 1)$. The centered square region is filled with 1, as illustrated in Fig. 2.

Then, for an arbitrary $x^s \in D^s$, we randomly sample an image $x_j^t \in x^t$ where $x^t \in D^t$, and conduct domain alignment for each component $x_k^s \in x^s$ as follows [12]:

$$x_k^{s\to t} = \mathcal{F}^{-1}([M_\alpha \odot \mathcal{F}^A(x_j^t) + (1 - M_\alpha) \odot \mathcal{F}^A(x_k^s), \mathcal{F}^P(x_k^s)]), \quad (2)$$

where $\odot$ is the element-wise multiplication operation, and j and k both belong to $\{l, r\}$ and $\{t, t+1\}$ for disparity and optical flow estimation, respectively. Eq. (2) shows that we first replace the low frequency part of the amplitude of the source image x_k^s with the same part of the target image x_j^t [12]. Then, the modified amplitude map combined with the original phase map of the source image x_k^s is transformed back to the image $x_k^{s\to t}$, which constitutes the adapted source dataset $D_k^{s\to t}$ [12]. Fig. 2 shows that the scenario of $x_k^{s\to t}$ remains the same as that of x_k^s, but the image style becomes similar to that of x_j^t. Please note that we only sample x_j^t once for each x^s to preserve the DCM consistency between the images of x^s after domain alignment, which can also ensure the consistency of the DCM ground-truth labels between D^s and $D^{s\to t}$.

After domain alignment, we have the adapted source dataset $D^{s\to t}$, which contains ground-truth labels for DCM tasks and presents a similar image style to the target domain. Subsequently, we train two identical networks with different

Authorized licensed use limited to: Univ of Calif San Diego. Downloaded on July 19,2022 at 00:22:02 UTC from IEEE Xplore. Restrictions apply.

Algorithm 1: Co-Teaching Strategy

Input: θ_{A} and θ_{B}, learning rate η, constant threshold τ_c and τ_o, epoch T_k and $T_{\max}$, iteration $N_{\max}$.
Output: θ_{A} and θ_{B}.

1 **for** $T = 1 \rightarrow T_{\max}$ **do**
2 **Update** $\mathcal{R}_c(T) = \tau_c \cdot \min\{\frac{T}{T_k}, 1\}$ ▷ Update the threshold to filter out pixels with low confidence in $\hat{y}^t$
3 **Update** $\mathcal{R}_o(T) = 1 - \tau_o \cdot \min\{\frac{T}{T_k}, 1\}$ ▷ Update the threshold to filter out pixels with high occlusion probability in O^t
4 **for** $N = 1 \rightarrow N_{\max}$ **do**
5 **Forward** individually to obtain $\hat{y}_i^{s \rightarrow t}$, $\hat{y}_i^t$, C_i^t and O_i^t, $i \in \{\mathrm{A}, \mathrm{B}\}$
6 **Set** $C_i^t \left(C_i^t < \mathcal{R}_c(T)\right) = 0$, $i \in \{\mathrm{A}, \mathrm{B}\}$ ▷ Filter out pixels with low confidence in $\hat{y}^t$
7 **Set** $O_i^t \left(O_i^t > \mathcal{R}_o(T)\right) = 1$, $i \in \{\mathrm{A}, \mathrm{B}\}$ ▷ Filter out pixels with high occlusion probability in O^t
8 **Compute** $\mathcal{L}_{\mathrm{A}} = \mathcal{L}_{\mathrm{A}}^{\mathrm{su}}(\hat{y}_{\mathrm{A}}^{s \rightarrow t}, y^{s \rightarrow t}) + \lambda_1 \cdot \mathcal{L}_{\mathrm{A}}^{\mathrm{ss}}(\hat{y}_{\mathrm{A}}^t, \hat{y}_{\mathrm{B}}^t, C_{\mathrm{B}}^t) + \lambda_2 \cdot \mathcal{L}_{\mathrm{A}}^{\mathrm{un}}(x^t, \hat{y}_{\mathrm{A}}^t, O_{\mathrm{B}}^t)$
9 **Compute** $\mathcal{L}_{\mathrm{B}} = \mathcal{L}_{\mathrm{B}}^{\mathrm{su}}(\hat{y}_{\mathrm{B}}^{s \rightarrow t}, y^{s \rightarrow t}) + \lambda_1 \cdot \mathcal{L}_{\mathrm{B}}^{\mathrm{ss}}(\hat{y}_{\mathrm{B}}^t, \hat{y}_{\mathrm{A}}^t, C_{\mathrm{A}}^t) + \lambda_2 \cdot \mathcal{L}_{\mathrm{B}}^{\mathrm{un}}(x^t, \hat{y}_{\mathrm{B}}^t, O_{\mathrm{A}}^t)$
10 **Update** $\theta_i = \theta_i - \eta \nabla \mathcal{L}_i$, $i \in \{\mathrm{A}, \mathrm{B}\}$
11 **end**
12 **end**

initialization parameters on $D^{s \rightarrow t}$ using commonly adopted supervised loss $\mathcal{L}^{\mathrm{su}}(\hat{y}^{s \rightarrow t}, y^{s \rightarrow t})$ for any DCM task, where $\hat{y}^{s \rightarrow t}$ denotes the corresponding DCM estimation. However, since $D^{s \rightarrow t}$ does not possess the same data distribution as D^t, networks trained only on $D^{s \rightarrow t}$ cannot adapt to D^t well. Therefore, this training process only lasts for several epochs for DCM initialization, and then we step into the next phase for DCM refinement by further utilizing D^t. The reasons why we employ two identical networks with different initialization parameters are explained in the next subsection.

B. DCM Refinement Phase

When using D^t for further DCM refinement, our UnDAF turns UDA for DCM into a semi-supervised learning (SSL) problem [12]. One popular technique for SSL problems is self-training [24]. Specifically, the predictions and the corresponding confidence maps in the target domain are first generated, and then the predictions with high confidence are regarded as pseudo ground-truth labels for supervision, which minimizes the following loss [9]:

$$\mathcal{L}^{\mathrm{ss}}(\hat{y}^t, \tilde{y}^t, C^t) = \frac{\sum_p \mathcal{L}^{\mathrm{diff}}(\hat{y}^t, \tilde{y}^t) \odot \mathcal{S}\left(C^t\right)}{\sum_p \mathcal{S}\left(C^t\right)}, \tag{3}$$

where $\tilde{y}^t$ and C^t respectively denote the pseudo ground-truth labels and the corresponding confidence maps in the target domain; p denotes all valid pixels; $\mathcal{L}^{\mathrm{diff}}(\cdot, \cdot)$ measures the difference between two inputs, which is similar to $\mathcal{L}^{\mathrm{su}}(\cdot, \cdot)$; and $\mathcal{S}(\cdot)$ is the stop-gradient operation. However, since the predicted confidence maps and the DCM estimations are highly correlated, the selected highly confident pseudo ground-truth labels can be highly noisy, which will lead to significant performance degradation for the target domain.

Another loss prevalently adopted in UDA for DCM is the unsupervised occlusion-aware photometric loss [25], [26], which has the following formulation:

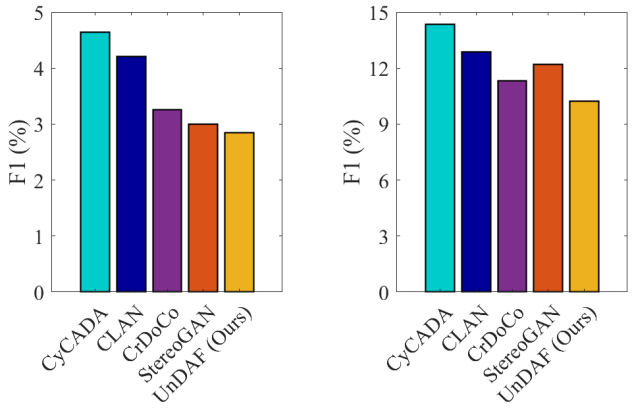
$$\mathcal{L}^{\mathrm{un}}(x^t, \hat{y}^t, O^t) = \frac{\sum_p \mathcal{L}^{\mathrm{ph}}(x^t, \hat{y}^t) \odot \left(1 - \mathcal{S}\left(O^t\right)\right)}{\sum_p \left(1 - \mathcal{S}\left(O^t\right)\right)}, \tag{4}$$

where $\mathcal{L}^{\mathrm{ph}}(x^t, \hat{y}^t)$ measures the photometric difference in x^t based on $\hat{y}^t$; and $O^t \in [0, 1]$ is the occlusion map, which measures the occlusion probability for each pixel. However, since the DCM estimations and the occlusion maps are highly correlated, simply using $\mathcal{L}^{\mathrm{un}}$ can also lead to the same performance degradation problem caused by overfitting to the noise as simply using $\mathcal{L}^{\mathrm{ss}}$.

Inspired by [27], we adopt a co-teaching strategy in the refinement phase to address these issues, as shown in Algorithm 1. We simultaneously train two identical networks A (with parameter θ_{A}) and B (with parameter θ_{B}), which have been initialized in the previous phase. In each epoch, we first update two thresholds $\mathcal{R}_c(T)$ and $\mathcal{R}_o(T)$ to filter out the pixels with low confidence in $\hat{y}^t$ and the pixels with high occlusion probability in O^t (Line 2 and 3), respectively. $\mathcal{R}_c(T)$ and $\mathcal{R}_o(T)$ change gradually with epoch increases to ensure that the confidence of the selected pixels in $\hat{y}$ becomes increasingly high and the occlusion probability of the selected pixels in O^t becomes increasingly low. Subsequently, we let two networks forward individually on $D^{s \rightarrow t}$ and D^t to generate several outputs (Line 5), respectively. After that, we filter out pixels with low confidence in $\hat{y}^t$ and pixels with high occlusion probability in O^t (Line 6 and 7). The key to our co-teaching strategy is that each network computes its own loss after two networks exchange several important variants including $\hat{y}^t$, C^t and O^t (Line 8 and 9), as illustrated in Fig. 2. Finally, we update the parameters of two networks separately (Line 10).

The reason why our co-teaching strategy can improve the accuracy of UDA for DCM is twofold. 1) Generally, networks first learn clear patterns, and then overfit to the noise, which further causes significant performance degradation [28]. The way of gradually changing the filtering thresholds $\mathcal{R}_c(T)$ and $\mathcal{R}_o(T)$ enables the networks to avoid overfitting to the possible outliers [27]. 2) In addition to the dynamic threshold scheme, we also let two networks exchange several important variants including $\hat{y}^t$, C^t and O^t, so that they can further evolve with better robustness and accuracy. Specifically, when simply using $\mathcal{L}^{\mathrm{ss}}$ and $\mathcal{L}^{\mathrm{un}}$, there always exist two highly correlated variants, which can induce a lot of noise and further cause performance degradation. By forcing two networks to exchange these variants, the correlation chain can be broken, which can

Authorized licensed use limited to: Univ of Calif San Diego. Downloaded on July 19,2022 at 00:22:02 UTC from IEEE Xplore. Restrictions apply.



(a) Disparity Estimation (b) Optical Flow Estimation

Fig. 3: Performance comparison among CyCADA [21], CLAN [22], CrDoCo [7], StereoGAN [8] and our UnDAF on the KITTI 2015 training set [14].

Approach	ST	KITTI 2012		KITTI 2015	
		Noc	All	Noc	All
PSMNet [5]	✓	1.49	1.89	2.14	2.32
GwcNet-gc [32]	✓	1.32	1.70	1.92	2.11
AcfNet [33]	✓	1.17	1.54	1.72	1.89
LEAStereo [3]	✓	**1.13**	**1.45**	**1.51**	**1.65**
OASM-Net [34]	–	6.39	8.60	7.39	8.98
MC-CNN-WS [35]	–	3.02	4.45	4.11	4.97
MADNet [10]	–	–	–	4.27	4.66
SsSMnet [25]	–	2.30	3.00	3.06	3.40
UnDAF (Ours)	–	**1.79**	**2.25**	**2.33**	**2.56**

TABLE I: Disparity evaluation results (%) on the KITTI Stereo 2012 [13] and KITTI Stereo 2015 [14] benchmarks. "ST" denotes supervised training on the benchmarks. "Noc" and "All" represent the F1 for non-occluded pixels and all pixels, respectively [13], [14]. "UnDAF" is short for "UnDAF-LEAStereo". Best results for "ST" and "non-ST" approaches are both bolded.

effectively improve the DCM performance.

Please note that although our co-teaching strategy looks similar to several existing approaches [27], [29], [30], they have several major differences. 1) [27] focuses on image-level tasks (image classification), while our approach focuses on pixel-level tasks (DCM). 2) [27] is designed for supervised learning with noisy labels, and our previous works [29], [30] are designed for unsupervised stereo matching and optical flow estimation. Differently, our approach is designed for UDA, a completely new task. Therefore, the details of our approach and these existing approaches [27], [29], [30] are also different. Specifically, our approach performs domain alignment before adopting the co-teaching strategy. In addition, our co-teaching strategy takes confidence maps into consideration to improve the accuracy of UDA for DCM.

IV. Experiments

A. Datasets and Implementation Details

To validate the effectiveness of our UnDAF, we respectively adopt LEAStereo [3] and RAFT [4] as the backbone networks for disparity and optical flow estimation, since they have achieved impressive performance for the corresponding DCM tasks. Their combinations with our UnDAF are respectively referred to as UnDAF-LEAStereo and UnDAF-RAFT. The supervised loss $\mathcal{L}^{\text{su}}$, learning rate η, maximum number of epochs T_{max} and optimizer are the same as those used in the backbone network for each DCM task. For α selection, we follow [12] and select three values, $\alpha = 0.01$, $\alpha = 0.05$ and $\alpha = 0.09$. The final DCM estimation is the average value of these three models.

In our experiments, we first compare our UnDAF with other state-of-the-art UDA approaches. However, many existing approaches do not publish their results on the above-mentioned benchmarks. Also, we cannot test their performance on these benchmarks because of their submission policies. Therefore, we respectively set the Virtual KITTI 2 dataset [31] and the KITTI 2015 training set [14] as the source and target domains, and compare the DCM performance of our UnDAF with other state-of-the-art UDA approaches. The experimental results are presented in Section IV-B.

To further evaluate our UnDAF on the public benchmarks, the following adaptation schemes are adopted: 1) Scene Flow [36] $\rightarrow$ MPI Sintel [15], 2) Virtual KITTI 2 [31] $\rightarrow$ KITTI 2012 [13] and 3) Virtual KITTI 2 [31] $\rightarrow$ KITTI 2015 [14]. The experimental results are presented in Section IV-C. Furthermore, we conduct ablation studies to demonstrate the effectiveness of our co-teaching strategy and selected loss functions. The experimental results are shown in Section IV-D.

Two standard evaluation metrics are used for performance comparison, 1) the average end-point error (AEPE) that measures the average difference between the DCM estimations and ground-truth labels and 2) the percentage of erroneous pixels (F1) that measures the percentage of bad pixels whose error is larger than 3 pixels [13]–[15]. The AEPE and F1 can be computed over all pixels or only non-occluded pixels [13], [14]. Please note that the two metrics are both computed over all pixels, if not specified.

B. Comparison with Other UDA Approaches

We compare our UnDAF with several existing UDA approaches, including CyCADA [21], CLAN [22], CrDoCo [7] and StereoGAN [8]. All approaches employ the same backbone network as our UnDAF for each DCM task, and the performance comparison is shown in Fig. 3. It is clearly observed that our UnDAF outperforms the state-of-the-art UDA approaches for both disparity and optical flow estimation, which demonstrates the effectiveness of our UnDAF.

C. Evaluation Results on Public Benchmarks

Referring to the online leaderboards of the KITTI 2012 [13], KITTI 2015 [14] and MPI Sintel [15] benchmarks shown in Table I and II, our UnDAF greatly surpasses all

Authorized licensed use limited to: Univ of Calif San Diego. Downloaded on July 19,2022 at 00:22:02 UTC from IEEE Xplore. Restrictions apply.

Approach	ST	MPI Sintel		KITTI 2012		KITTI 2015	
		Clean (px)	Final (px)	Noc (%)	All (%)	Noc (%)	All (%)
PWC-Net [6]	✓	4.39	5.04	4.22	8.10	6.12	9.60
LiteFlowNet [37]	✓	4.54	5.38	3.27	7.27	5.49	9.38
LiteFlowNet3 [38]	✓	2.99	4.45	**2.51**	**5.90**	4.29	7.34
RAFT [4]	✓	**1.61**	**2.86**	–	–	**3.07**	**5.10**
UnFlow [26]	–	9.38	10.22	4.28	8.42	7.46	11.11
DDFlow [19]	–	6.18	7.40	4.57	8.86	9.55	14.29
SelFlow† [39]	–	6.56	6.57	4.31	7.68	9.65	14.19
UFlow [40]	–	5.21	6.50	4.26	7.91	8.41	11.13
UnDAF-RAFT (Ours)	–	**3.91**	**5.08**	**3.36**	**7.32**	**5.95**	**9.56**

TABLE II: Optical flow evaluation results on the MPI Sintel [15], KITTI Optical Flow 2012 [13] and KITTI Optical Flow 2015 [14] benchmarks. "ST" denotes supervised training on the benchmarks. For the MPI Sintel Clean and Final benchmarks [15], the AEPE for all pixels is presented. For the KITTI Optical Flow 2012 and KITTI Optical Flow 2015 benchmarks, "Noc" and "All" represent the F1 for non-occluded pixels and all pixels, respectively [13], [14]. † indicates the model using more than two frames. Best results for "ST" and "non-ST" approaches are both bolded.

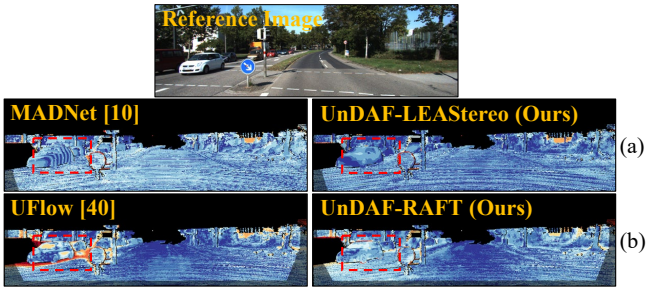


Fig. 4: Error maps for (a) disparity and (b) optical flow estimation on the KITTI 2015 benchmark [14]. Significantly improved regions are highlighted with red boxes.

No.	$\mathcal{L}^{ss}$	$\mathcal{L}^{un}$	CoT	F1 (%)	
				Disparity	Optical Flow
(a)	–	–	–	6.95	16.75
(b)	✓	–	–	4.59	13.70
(c)	✓	✓	–	3.36	11.67
(d)	✓	✓	✓	**2.85**	**10.23**

TABLE III: Experimental results of our UnDAF with some of the loss functions or our co-teaching (CoT) technique disabled. Best results are bolded.

other unsupervised approaches for both disparity and optical flow estimation. Since these unsupervised approaches and our UnDAF all do not require the ground-truth labels of the target dataset (real-world dataset), these experimental results have demonstrated the effectiveness of our UnDAF for constructing a bridge between synthetic and real-world DCM applications. Excitedly, it is observed that our UnDAF performs competitively, even compared with some supervised DCM approaches, which further verifies the effectiveness of our UnDAF. Some examples of the experimental results on the KITTI 2015 benchmark are shown in Fig. 4, where it is evident that our UnDAF can achieve significant improvements compared to the existing approaches.

D. Ablation Study

In our ablation studies, we respectively set the Virtual KITTI 2 dataset [31] and the KITTI 2015 training set [14] as the source and target domains. The experimental results of our UnDAF with some of the loss functions or our co-teaching strategy disabled are presented in Table III. We can clearly observe that our co-teaching strategy can significantly improve the performance of UDA in DCM, and the setup with the combination of $\mathcal{L}^{ss}$, $\mathcal{L}^{un}$ and our co-teaching strategy achieves the best results for both disparity and optical flow estimation, as shown in (d) of Table III. We analyze that compared with existing approaches that transfer errors back to themselves directly, our co-teaching strategy enables two networks to adaptively correct the inaccurate estimations, which can effectively avoid error accumulation and further improve the stability and accuracy of UDA for DCM.

V. Conclusion

In this paper, we proposed a general unsupervised domain adaptation framework (UnDAF) for dense correspondence matching (DCM) tasks, including disparity and optical flow estimation. Specifically, we demonstrated that simply employing Fourier transform for domain alignment is effective for unsupervised domain adaptation (UDA) in DCM, which not only inherently preserves the DCM consistency between input images after domain alignment but also avoids the catastrophic forgetting problem. It also enables our UnDAF to avoid additional training process beyond the primary DCM task, such as complicated adversarial learning. In addition, we developed a co-teaching strategy, which can effectively avoid error accumulation and improve both the stability and accuracy of UDA for DCM. Extensive experiments on the public benchmarks demonstrated the accuracy and robustness of our UnDAF, advancing all other state-of-the-art UDA approaches for disparity or optical flow estimation. Moreover, the DCM networks trained on the synthetic dataset can successfully adapt to real-world driving scenarios via our UnDAF, which constructed a bridge connecting synthetic and real-world DCM applications.

Authorized licensed use limited to: Univ of Calif San Diego. Downloaded on July 19,2022 at 00:22:02 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] Z. Min, Y. Yang, and E. Dunn, "VOLDOR: Visual odometry from log-logistic dense optical flow residuals," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4898–4909.

[2] T. Yang, C. Cappelle, Y. Ruichek, and M. El Bagdouri, "Online multi-object tracking combining optical flow and compressive tracking in markov decision process," *Journal of Visual Communication and Image Representation*, vol. 58, pp. 178–186, 2019.

[3] X. Cheng, Y. Zhong, M. Harandi, Y. Dai, X. Chang, H. Li, T. Drummond, and Z. Ge, "Hierarchical neural architecture search for deep stereo matching," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[4] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *European Conference on Computer Vision*. Springer, 2020, pp. 402–419.

[5] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5410–5418.

[6] D. Sun, X. Yang, M.-Y. Liu, and 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.

[7] Y.-C. Chen, Y.-Y. Lin, M.-H. Yang, and J.-B. Huang, "CrDoCo: Pixel-level domain transfer with cross-domain consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1791–1800.

[8] R. Liu, C. Yang, W. Sun, X. Wang, and H. Li, "StereoGAN: Bridging synthetic-to-real domain gap by joint optimization of domain translation and stereo matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 757–12 766.

[9] A. Tonioni, M. Poggi, S. Mattoccia, and L. Di Stefano, "Unsupervised adaptation for deep stereo," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1605–1613.

[10] A. Tonioni, F. Tosi, M. Poggi, S. Mattoccia, and L. D. Stefano, "Real-time self-adaptive deep stereo," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 195–204.

[11] X. Tao, X. Chang, X. Hong, X. Wei, and Y. Gong, "Topology-preserving class-incremental learning," in *European Conference on Computer Vision*. Springer, 2020, pp. 254–270.

[12] Y. Yang and S. Soatto, "FDA: Fourier domain adaptation for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4085–4095.

[13] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3354–3361.

[14] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3061–3070.

[15] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *European Conference on Computer Vision*, ser. Part IV, LNCS 7577, A. Fitzgibbon et al. (Eds.), Ed. Springer-Verlag, Oct. 2012, pp. 611–625.

[16] R. Fan, X. Ai, and N. Dahnoun, "Road surface 3d reconstruction based on dense subpixel disparity map estimation," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 3025–3035, 2018.

[17] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial intelligence*, vol. 17, no. 1-3, pp. 185–203, 1981.

[18] C. Zhou, H. Zhang, X. Shen, and J. Jia, "Unsupervised learning of stereo matching," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1567–1575.

[19] P. Liu, I. King, M. R. Lyu, and J. Xu, "DDFlow: Learning optical flow with unlabeled data distillation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8770–8777.

[20] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Advances in Neural Information Processing Systems*, 2017, pp. 700–708.

[21] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "CyCADA: Cycle-consistent adversarial domain adaptation," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1989–1998.

[22] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2507–2516.

[23] M. Frigo and S. G. Johnson, "FFTW: An adaptive software architecture for the FFT," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98*, vol. 3. IEEE, 1998, pp. 1381–1384.

[24] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2517–2526.

[25] Y. Zhong, Y. Dai, and H. Li, "Self-supervised learning for stereo matching with self-improving ability," *CoRR*, 2017.

[26] S. Meister, J. Hur, and S. Roth, "UnFlow: Unsupervised learning of optical flow with a bidirectional census loss," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[27] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Advances in Neural Information Processing Systems*, 2018, pp. 8527–8537.

[28] D. Arpit, S. Jastrzębski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio *et al.*, "A closer look at memorization in deep networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 233–242.

[29] H. Wang, R. Fan, and M. Liu, "CoT-AMFlow: Adaptive modulation network with co-teaching strategy for unsupervised optical flow estimation," in *Proceedings of the 2020 Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, J. Kober, F. Ramos, and C. Tomlin, Eds., vol. 155. PMLR, 16–18 Nov 2021, pp. 143–155.

[30] H. Wang, R. Fan, and M. Liu, "Co-teaching: An ark to unsupervised stereo matching," in *2021 IEEE International Conference on Image Processing*. IEEE, 2021, pp. 3328–3332.

[31] Y. Cabon, N. Murray, and M. Humenberger, "Virtual KITTI 2," *CoRR*, 2020.

[32] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3273–3282.

[33] Y. Zhang, Y. Chen, X. Bai, S. Yu, K. Yu, Z. Li, and K. Yang, "Adaptive unimodal cost volume filtering for deep stereo matching." in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 12 926–12 934.

[34] A. Li and Z. Yuan, "Occlusion aware stereo matching via cooperative unsupervised learning," in *Proceedings of the Asian Conference on Computer Vision*. Springer, 2018, pp. 197–213.

[35] S. Tulyakov, A. Ivanov, and F. Fleuret, "Weakly supervised learning of deep metrics for stereo reconstruction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1339–1348.

[36] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4040–4048.

[37] T.-W. Hui, X. Tang, and C. Change Loy, "LiteFlowNet: A lightweight convolutional neural network for optical flow estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8981–8989.

[38] T.-W. Hui and C. C. Loy, "LiteFlowNet3: Resolving correspondence ambiguity for more accurate optical flow estimation," in *European Conference on Computer Vision*. Springer, 2020, pp. 169–184.

[39] P. Liu, M. Lyu, I. King, and J. Xu, "SelFlow: Self-supervised learning of optical flow," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4571–4580.

[40] R. Jonschkowski, A. Stone, J. T. Barron, A. Gordon, K. Konolige, and A. Angelova, "What matters in unsupervised optical flow," in *European Conference on Computer Vision*. Springer, 2020, pp. 557–572.

Authorized licensed use limited to: Univ of Calif San Diego. Downloaded on July 19,2022 at 00:22:02 UTC from IEEE Xplore. Restrictions apply.