StereoFlowGAN: Co-training for Stereo and Flow with Unsupervised Domain Adaptation

Zhexiao Xiong[1]
x.zhexiao@wustl.edu

Feng Qiao[2]
feng.qiao.ad@gmail.com

Yu Zhang[3]
yuzh03@gmail.com

Nathan Jacobs[1]
jacobsn@wustl.edu

[1] Department of Computer Science & Engineering,
Washington University in St. Louis
St. Louis, MO, USA

[2] Institute for Automotive Engineering,
RWTH Aachen University
Templergraben 55, Aachen, Germany

[3] Department of Computer Science,
University of Kentucky
Lexington, KY, USA

Abstract

We introduce a novel training strategy for stereo matching and optical flow estimation that utilizes image-to-image translation between synthetic and real image domains. Our approach enables the training of models that excel in real image scenarios while relying solely on ground-truth information from synthetic images. To facilitate task-agnostic domain adaptation and the training of task-specific components, we introduce a bidirectional feature warping module that handles both left-right and forward-backward directions. Experimental results show competitive performance over previous domain translation-based methods, which substantiate the efficacy of our proposed framework, effectively leveraging the benefits of unsupervised domain adaptation, stereo matching, and optical flow estimation.

1 Introduction

Stereo matching and optical flow estimation are closely related computer vision tasks. Given an image pair, the aim of optical flow estimation is to predict a 2D vector field that reflects the pixel-wise displacement of temporally adjacent frames. In rectified stereo, the task is essentially the same. We simply use the known relative camera geometry to impose epipolar constraints, thereby reducing the correspondence-search problem from 2D to 1D. The result is a disparity map between the left image and right images which can be easily translated into a pixel-wise displacement field. We propose a co-training approach that exploits this inter-task similarity to simultaneously train networks for both tasks.

Acquiring high-quality, real training data for stereo matching and optical flow estimation is challenging and expensive, often requiring calibration between the depth sensor and stereo cameras. Therefore, optical flow estimation and stereo-matching methods highly rely on synthetic data for training. However, there is often a severe domain gap between synthetic

and real data, affecting the cross-domain generalization performance. To address this problem, unsupervised domain adaptation methods have been proposed to bridge the domain gap between synthetic and real data.

Inspired by StereoGAN [22], which first applied domain translation to stereo matching task, we propose a multi-task framework that concurrently executes optical flow estimation and stereo matching tasks with a shared domain translation module, in which we do not need the ground-truth disparity and optical flow of target domain real images. In the shared domain translation module, we use two ResNet-based generators of opposite directions to perform cross-domain image translation. Then two discriminators are constructed to minimize the discrepancy between translated and original images. Furthermore, we adopt perceptual loss to maintain the feature-level consistency and cosine similarity loss to regularize the cross-domain generation. Utilizing synthetic-to-real and real images, we predict corresponding disparity maps via a stereo-matching network using only ground-truth disparity of synthetic stereo data. Simultaneously we train an optical flow estimation network using the adjacent frames of synthetic-to-real images and real images, leveraging ground-truth optical flow data and occlusion masks from only synthetic data. The two task-agnostic networks are jointly optimized. To connect the three modules, we build a multi-scale left-right feature warping module and a forward-backward feature warping module, which not only provide supervision for image translation but also for the training of task-specific networks.

The key contributions of our paper are summarized below:

- We build an end-to-end joint learning framework to combine unsupervised domain translation with optical flow estimation and stereo matching in the absence of real ground truth optical flow and disparity, which facilitates the co-optimization of models, yielding superior performance compared with executing each task in isolation.

- We apply novel constraints on the cycle domain translation process to achieve cross-domain translation with global and local consistency, which significantly reduces the pixel distortion during the domain translation stage.

- We employ task-specific multi-scale feature warping loss and iterative feature warping loss during the training phase to regulate the training process of the shared domain translation module and task-specific module in both spatial and temporal dimensions.

- Experimental results demonstrate that our proposed model achieves top-tier results compared to other unsupervised domain adaptation methods for stereo matching and optical flow estimation.

2 Related Work

2.1 Stereo Matching

The aim of stereo matching is to generate disparity maps from left and right epipolar images. Traditionally, this involved a four-step process: matching cost computation, cost aggregation, disparity computation, and disparity refinement. Since DispNet [27], deep learning-based works [3, 17, 30, 31, 55] have become popular for more accurate, real-time stereo matching.

Inspired by RAFT [36], iterative 2D methods have been applied to this task. Notable models include AANet [46], which forgoes 3D convolutions for efficiency; RAFT-Stereo [18], an adaptation of previous optical flow work; CREStereo [16], a cascaded recurrent network

for practical stereo matching; and IGEV-Stereo [45], which uses a combined geometry encoding volume for iterative disparity map updates.

2.2 Optical Flow Estimation

Optical flow estimation aims to estimate per-pixel motion between video frames. Since FlowNet [7], a series of deep neural networks have been proposed, with some [4, 11, 40] using U-Net encoder-decoder structures that often lose detail in feature maps, while others [9, 10, 33] use spatial pyramid networks with feature warping to reduce feature-space distance and adaptively regulate flow. A classic method RAFT [36] extracts per-pixel features and iteratively updates a flow field through a recurrent unit using multi-scale 4D correlation volumes, which enables strong cross-dataset generalization. Recently, GMFlow [47] and its follow-up work Unimatch [48], both based on Transformers, reformulate optical flow as a global matching problem and compare feature similarities directly instead of applying extensive iterative refinements.

Besides supervised methods, some unsupervised methods [14, 15, 20, 21, 24] have also been proposed, among which UPFlow [24] proposed a self-guided upsample module to tackle the interpolation blur problem in optical flow estimation, [20] proposed to use more reliable supervision from transformations, and [14, 15, 21] tried to utilize the relationships between stereo matching and optical flow estimation task. In our work, we suggest an efficient end-to-end co-training framework for improving performance on both tasks.

2.3 Unsupervised Domain Adaptation

Transfer learning has been widely used in many computer vision tasks, such as detection [1, 23, 49], segmentation [25, 34, 44], and stereo matching [19, 51]. Unsupervised domain adaptation is a special transfer learning technique, which uses labeled source data and unlabeled target data, with numerous methods [6, 8, 13, 39, 41, 43, 50] developed to bridge the domain gaps. Many works have applied unsupervised domain adaptation to stereo matching and optical flow estimation tasks. Key contributions include a self-adaptation method with graph Laplacian regularization [29], real-time online deep stereo adaptation [38], Information-Theoretic Shortcut Avoidance (ITSA) [5] for domain generalization, StereoGAN [22] employing an end-to-end training framework, and AdaStereo [32] utilizing a non-adversarial progressive color transfer algorithm. In optical flow, strategies like co-teaching [42] for domain alignment and meta-training [28, 37] have also been proposed.

3 Method

We first describe the problem of domain translation-based optical flow estimation and stereo-matching joint training. Then we introduce the overall framework of our proposed pipeline. After that, we introduce the main components of the pipeline in detail, including the domain translation module, the stereo matching and optical flow feature warping module, and the unsupervised joint optimization scheme. The overall pipeline of our proposed framework is shown in Figure 1.

3.1 Problem and Motivation

Given a set of N synthetic left-right-forward-disparity-flow tuples $\{(x_l, x_r, x_{l\_(t+1)}, x_d, x_f)_i\}_{i=1}^{N}$ of source domain A, and a set of M real image tuples $\{(y_l, y_r, y_{l\_(t+1)})_i\}_{i=1}^{M}$ of target domain
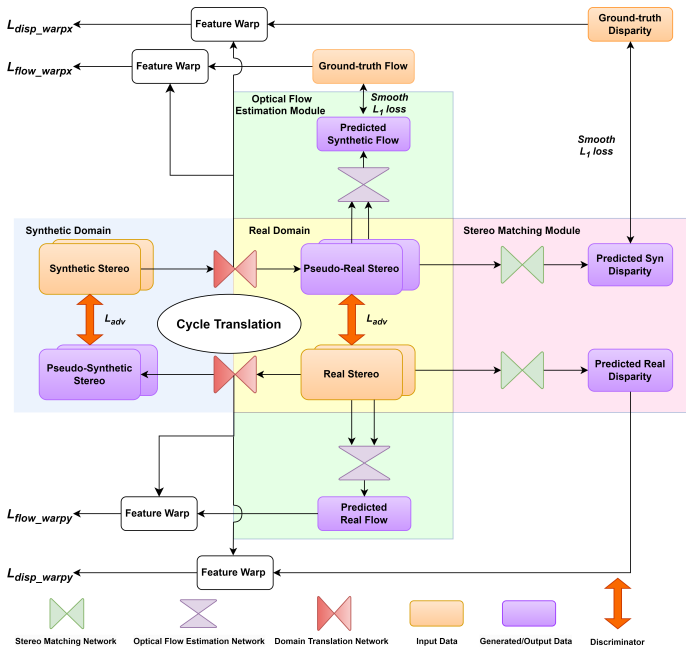
Figure 1: The framework of our proposed method. The blue block represents the synthetic domain and the yellow block represents the real domain. Domain translation is conducted between these two blocks. The pink and green blocks show stereo matching and optical flow estimation modules respectively. Please see Figure 2 for the detail of the cycle translation module and Figure 3 for the feature warp module.

B without ground truth, our goal is to conduct accurate domain translation, jointly optimize the disparity estimation network F_{disp} and optical flow estimation network F_{flow} for estimating the disparity $\hat{y}_d$ and optical flow $\hat{y}_f$ on the target domain. We propose to use left-right and forward-backward feature warping to jointly supervise the cross-domain translation and the task-specific framework in both spatial and temporal dimensions.

3.2 Domain Translation Module

In the domain translation module, take A as the source domain and B as the target domain. In a data batch of dataloader, we load $I_{leftA}, I_{rightA}, I_{leftA\_(t+1)}, dispA, flowA, I_{leftB}, I_{rightB}, I_{leftB\_(t+1)}$. Inspired by pixel2pixel[□], we build a generator G_{A2B} to translate the synthetic image I_{leftA} into to real domain and get I_{fake_leftB}, and a discriminator D_B to help distinguish between the synthetic-to-real translated data I_{fake_leftB} and the real data I_{leftB}. Similarly, we build another generator G_{B2A} and discriminator D_A with the same structure to do real-to-synthetic image translation from I_{leftB} to I_{fake_leftA} in an adversarial manner. The adversarial loss is defined as:

$$\mathcal{L}_{adv}(G_{A2B}, D_B, \mathcal{X}, \mathcal{Y}) = \mathbb{E}_{y \sim \{\mathcal{Y}_L, \mathcal{Y}_R\}}[\log D_B(y)] + \mathbb{E}_{x \sim \{\mathcal{X}_L, \mathcal{X}_R\}}[\log(1 - D_B(G_{A2B}(x)))], \quad (1)$$

where $x \sim \{\mathcal{X}_L, \mathcal{X}_R\}$ represents the synthetic image pair and $y \sim \{\mathcal{Y}_L, \mathcal{Y}_R\}$ represents the real image pair. Similarly the inverse real-to-synthetic domain generation is represented as $\mathcal{L}_{adv}(G_{B2A}, D_A, \mathcal{Y}, \mathcal{X})$.
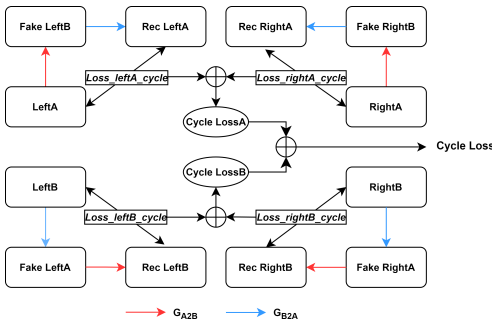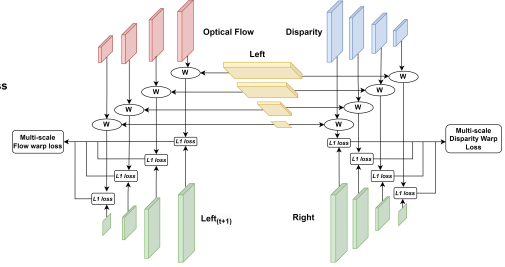
Figure 2: Cycle translation module.

Figure 3: Multi-scale feature warping module.

Furthermore, inspired by CycleGAN [52], we make cycle domain translation from the fake-real domain back to the synthetic domain through G_{B2A} and get I_{rec_leftA}, with the reversed process from the fake-synthetic domain back to the real domain processed by passing G_{A2B} and we get I_{rec_leftB}. The framework of cycle domain translation is shown in Figure 2. We name it cycle loss, which is formulated as below:

$$
\begin{aligned}
\mathcal{L}_{cyc}(G_{A2B}, G_{B2A}) &= \mathbb{E}_{y \sim \{\mathcal{Y}_L, \mathcal{Y}_R\}}[\|G_{A2B}(G_{B2A}(y)) - y\|_1 + (1 - SSIM(G_{A2B}(G_{B2A}(y) - y)))] \\
&+ \mathbb{E}_{x \sim \{\mathcal{X}_L, \mathcal{X}_R\}}[\|G_{B2A}(G_{A2B}(x)) - x\|_1 + (1 - SSIM(G_{A2B}(G_{B2A}(y) - y)))],
\end{aligned}
\tag{2}
$$

where $\mathbb{E}_{x \sim \{\mathcal{X}_L, \mathcal{X}_R\}}$ represents source domain image pairs and $\mathbb{E}_{y \sim \{\mathcal{Y}_L, \mathcal{Y}_R\}}$ represents target domain image pairs. *SSIM* means structural similarity index measure(SSIM).

Specifically, to generate photorealistic cross-domain images, we use a VGG-19-based encoder-decoder structure to maintain the global feature-level similarity between the cycle-synthesized image and the source image, and adopt a perceptual loss. We also use cosine similarity to measure the pixel-level distance between the source domain and target domain and develop a cosine-similarity loss to help the domain translation network maintain local similarity, which are defined as:

$$
\mathcal{L}_p(G_{A2B}, G_{B2A}) = \mathcal{L}_{perceptual}(G_{A2B}(G_{B2A}(y)), y) + \mathcal{L}_{perceptual}(G_{B2A}(G_{A2B}(x)), x)
\tag{3}
$$

$$
\mathcal{L}_{cos}(G_{A2B}, G_{B2A}) = [1 - \cos(G_{A2B}(G_{B2A}(y)), y)] + [1 - \cos(G_{B2A}(G_{A2B}(x)), x)]
\tag{4}
$$

The domain translation loss could be calculated by summarizing the losses together in the cycle-consistency component, which is defined as:

$$
\begin{aligned}
\mathcal{L}_{translation}(G_{A2B}, G_{B2A}, D_A, D_B) &= \mathcal{L}_{adv}(G_{A2B}, D_B, \mathcal{X}, \mathcal{Y}) + \mathcal{L}_{adv}(G_{B2A}, D_A, \mathcal{Y}, \mathcal{X}) \\
&+ \lambda_{cyc}\mathcal{L}_{cyc}(G_{A2B}, G_{B2A}) + \mathcal{L}_p(G_{A2B}, G_{B2A}) + \mathcal{L}_{cos}(G_{A2B}, G_{B2A})
\end{aligned}
\tag{5}
$$

3.3 Stereo Matching and Optical Flow Feature Warping Loss

Feature warping losses are widely used in stereo matching and optical flow estimation tasks. Direct domain translation may lack the precise location information between the left and right images. Therefore, to supervise the training of domain translation and help generate images with precise information, we extract features of the source image and conduct feature warping between both left-right images and forward-backward images. For models like DispNetC, as they naturally output multi-scale disparities from correlation features of network layers, we do multi-scale warp in the real-synthetic-real cycle translation. For recent models like IGEV-Stereo [45], and Unimatch [48], as they are trained in an iterative

refinement method like RAFT [36], we extract the predicted disparity map or optical flow of different refinement stages. Inspired by [13], we calculated the smooth $\mathcal{L}_1$ loss between the warped image and target image during different refinement stages, and calculate the weighted warping loss. The framework of warping loss is shown in Figure 3.

Synthetic Image Loss   During the training of the domain translation network, if the generators are well-trained, based on the synthetic ground-truth disparity and optical flow, the warped features should match the features of the target image exactly. Therefore, to supervise the training of the generators, we warp the feature maps of G_{A2B} and G_{B2A} in the synthetic-real-synthetic cycle translation process. The left-right warping loss for synthetic images is formulated as Eq 6.

$$\mathcal{L}_{disp\_warpx}(G_{A2B}, G_{B2A}) = \mathbb{E}_{(x_l, x_r, x_d) \sim \mathcal{X}} \frac{1}{T_1} \sum_{i=1}^{T_1} \left[\left\| W\left(G_{A2B}^{(i)}(x_l), x_d\right) - G_{A2B}^{(i)}(x_r) \right\|_1 \right.$$
$$\left. + \left\| W\left(G_{B2A}^{(i)}(G_{A2B}(x_l)), x_d\right) - G_{B2A}^{(i)}(G_{A2B}(x_r)) \right\|_1 \right], \tag{6}$$

in which T_1 is the number of extracted feature layers of the domain translation generator for the stereo matching task. $G^{(i)}(x)$ represents the feature of image x at ith-layer in the domain translation network G, the warping function $W(G^{(i)}(x_l), x_d)$ warps the left feature map $G^{(i)}(x_l)$ with the ground truth disparity x_d. The inverse warp from I_{rightA} to I_{fake_leftB} is conducted in the meantime. Similarly, we use the forward-backward warping loss to provide temporal supervision. Based on the predicted flow, we use G_{A2B} generator to warp the image from t time synthetic domain I_{leftA_t} to $(t+1)$ time real domain $I_{fake\_leftB\_(t+1)}$, and conduct an inverse warp from $(t+1)$ time synthetic domain $I_{leftA\_(t+1)}$ to t time real domain $I_{fake\_leftB\_t}$ in the meantime. The process is formulated as:

$$\mathcal{L}_{flow\_warpx}(G_{A2B}, G_{B2A}) = \mathbb{E}_{(x_l, x_{l\_(t+1)}, x_f) \sim \mathcal{X}} \frac{1}{T_2} \sum_{i=1}^{T_2} \left[\left\| W\left(G_{A2B}^{(i)}(x_l), x_f\right) - G_{A2B}^{(i)}\left(x_{l\_(t+1)}\right) \right\|_1 \right.$$
$$\left. + \left\| W\left(G_{B2A}^{(i)}(G_{A2B}(x_l)), x_f\right) - G_{B2A}^{(i)}\left(G_{A2B}\left(x_{l\_(t+1)}\right)\right) \right\|_1 \right], \tag{7}$$

in which T_2 is the number of extracted feature layers of the domain translation generator for the optical flow estimation task and x_f is the ground truth flow of t time left image. The feature warping loss of synthetic images serves as a bond between the shared domain translation module and the task-specific module, which supervises the generator to maintain feature-level consistency in the domain translation process.

Real Image Loss   During the training of the task-specific modules, we further use multi-scale disparity maps and flow to warp the feature maps of G_{B2A}. If the task-specific networks are well-trained, based on the estimated disparity and flow of real data, the warped feature maps should match the feature maps of the target images. Therefore, we conduct left-right and forward-backward feature warping to supervise the training of stereo matching and optical flow estimation modules respectively. For the stereo matching task, based on the predicted disparity, we extract multi-scale features of the image and use the G_{B2A} generator to warp the image from I_{fake_rightA} to I_{leftB}. In the meantime, we also conduct inverse warp from I_{fake_leftA} to I_{rightB}. The disparity warping loss of real images is defined as:

$$\mathcal{L}_{disp\_warpy}(G_{B2A}) = \mathbb{E}_{(y_l, y_r) \sim (\mathcal{Y}_L, \mathcal{Y}_R)} \frac{1}{T_1} \sum_{i=1}^{T_1} \left[\left\| W\left(G_{B2A}^{(i)}(y_l), \hat{y}_d\right) - G_{B2A}^{(i)}(y_r) \right\|_1 \right.$$
$$\left. + \left\| W\left(G_{B2A}^{(i)}(y_r), -\hat{y}_d\right) - G_{B2A}^{(i)}(y_l) \right\|_1 \right], \tag{8}$$

where $\hat{y}_d$ is the estimated disparity of real stereo image pairs by $F_{disp}(y_l, y_r)$.

Similarly, for the optical flow estimation task, we use the forward-backward warping loss to provide supervision and help maintain temporal consistency. Based on the predicted flow, we use G_{B2A} generator to warp the image from t time real domain I_{leftB_t} to $(t+1)$ time synthetic domain $I_{fake\_leftA\_(t+1)}$, and similarly do an inverse warp from $(t+1)$ time real domain $I_{leftB\_(t+1)}$ to t time synthetic domain $I_{fake\_leftA\_t}$. The flow warping loss of real images is formulated as:

$$\mathcal{L}_{flow\_warpy}(G_{B2A}) = \mathbb{E}_{(y_t, y_{t+1}) \sim (\mathcal{Y}_t, \mathcal{Y}_{t+1})} \frac{1}{T_2} \sum_{i=1}^{T_2} \left[\left\| W\left(G_{B2A}^{(i)}(y_t), \hat{y}_f\right) - G_{B2A}^{(i)}(y_{t+1}) \right\|_1 \right. $$
$$\left. + \left\| W\left(G_{B2A}^{(i)}(y_{t+1}), -\hat{y}_f\right) - G_{B2A}^{(i)}(y_t) \right\|_1 \right], \quad (9)$$

where $\hat{y}_f$ is the estimated optical flow of t time real stereo image pairs by $F_{flow}(y_t, y_{t+1})$.

3.4 Stereo Matching and Optical Flow joint training

Based on the cross-domain synthesized images, we jointly train stereo matching and optical flow estimation networks. We calculate the smooth L_1 loss between the predicted disparity/flow and estimated disparity/flow, during which features in the different refinement stages are all used under the supervision of the refinement stage. The loss functions are summarized below:

$$\mathcal{L}_{disp}(F_{disp}) = \mathbb{E}_{(x_l, x_r, x_d) \sim \mathcal{X}} \left[\left\| F_{disp}(G_{A2B}(x_l), G_{A2B}(x_r)) - x_d \right\|_1 \right], \quad (10)$$

$$\mathcal{L}_{flow}(F_{flow}) = \mathbb{E}_{(x_t, x_{t+1}, x_f) \sim \mathcal{X}} \left[\left\| F_{flow}(G_{A2B}(x_t), G_{A2B}(x_{t+1})) - x_f \right\|_1 \right], \quad (11)$$

where F_{disp} is the stereo matching network for estimating disparity and F_{flow} is the optical flow estimation network for estimating optical flow from real domain stereo images of left-right views and forward-backward views. We try different stereo-matching and optical flow estimation networks to evaluate the effectiveness of our proposed framework.

3.5 Joint Optimization

In the training process, we train the domain translation module, the stereo matching module, and the optical flow estimation module in an iterative way. For every k iteration, we update the gradient of the domain translation module while freezing the weights of stereo matching and optical flow estimation networks. During the nk to $(n+1)k-1$ iterations, the gradients of stereo matching and optical flow estimation modules are updated in the meantime while the parameters of the domain translation module are frozen.

Besides the losses we introduced above, we borrow correlation consistency loss $\mathcal{L}_{corr}$ and mode-seeking loss $\mathcal{L}_{ms}$ from StereoGAN [22], and follow the same loss setting as this work. The total loss for the domain translation network is the weighted sum of the individual loss functions:

$$\mathcal{L}_T(G_{A2B}, G_{B2A}, D_A, D_B) = \mathcal{L}_{translation}(G_{A2B}, G_{B2A}, D_A, D_B) + \lambda_{f_{disp\_warpx}} \mathcal{L}_{f_{disp\_warpx}}(G_{A2B}, G_{B2A})$$
$$+ \lambda_{f_{flow\_warpx}} \mathcal{L}_{f_{flow\_warpx}}(G_{A2B}, G_{B2A}) + \lambda_{corr} \mathcal{L}_{corr}(G_{A2B}, G_{B2A}) + \lambda_{ms} \mathcal{L}_{ms}(G_{A2B})$$
$$(12)$$

For the stereo matching network, the loss is formulated as:

$$\mathcal{L}_d(F_{disp}, G_{B2A}) = \lambda_{disp} \mathcal{L}_{disp}(F_{disp}) + \lambda_{f_{disp\_warpy}} \mathcal{L}_{f_{disp\_warpy}}(G_{B2A}) \quad (13)$$

For the optical-flow estimation task, the loss is formulated as:

$$\mathcal{L}_f(F_{flow}, G_{B2A}) = \lambda_{flow} \mathcal{L}_{flow}(F_{flow}) + \lambda_{f_{flow\_warpy}} \mathcal{L}_{f_{flow\_warpy}}(G_{B2A}) \quad (14)$$

Table 1: Results on datasets from Driving to KITTI2015. We take IGEV-Stereo [45] as the stereo matching network and Unimatch-flow [48] as the optical-flow estimation network. Source only means training on Driving and directly fine-tuning on KITTI2015.

Method	EPE	D1-all(%)	>2px(%)	>4px(%)	>5px(%)	EPE(flow)	F1-all(%)
IGEV-Stereo source only	2.48	16.40	28.23	11.08	8.06	–	–
Stereo GAN [22]	1.65	10.55	18.59	7.57	5.90	–	–
Unimatch-flow source only	–	–	–	–	–	14.72	42.20
proposed	**1.56**	**9.16**	**16.29**	**6.48**	**4.94**	**7.20**	**29.48**

Table 2: Results on datasets from Driving to KITTI2015. We take DispNetC [27] as the stereo-matching network and Unimatch-flow as the optical-flow estimation network.

Method	EPE	D1-all(%)	>2px(%)	>4px(%)	>5px(%)	EPE(flow)	F1-all(%)
DispNetC source only	7.56	53.84	65.91	45.05	38.36	–	–
Stereo GAN [22]	3.65	36.36	51.31	27.24	20.79	–	–
Unimatch-flow source only	–	–	–	–	–	14.72	42.20
proposed	**2.98**	**29.62**	**44.31**	**21.52**	**16.13**	**8.30**	**28.79**

In the equations above, $\lambda_s, s \in \{translation, f_{disp\_warpx}, f_{flow\_warpx}, corr, ms, disp, f_{disp\_warpy}, flow, f_{flow\_warpy}\}$ represent the weights of the different losses respectively. In the training stage, we jointly optimize $\mathcal{L}_T$, $\mathcal{L}_d$ and $\mathcal{L}_f$ together.

4 Experiments

In this section, we validate the effectiveness of our proposed method for unsupervised learning of stereo matching and optical flow estimation on several standard benchmark datasets.

4.1 Datasets

We implement our experiments on three autonomous driving datasets. The first one is Driving, which is a subset of a commonly used synthetic dataset, Sceneflow [26]. The sum of this subset is 4400 in total, with both ground-truth disparity and optical flow provided. The image size is 540×960 and the disparity value range from 0 to 300. The second one is Virtual-KITTI2 (VKITTI2) [2] dataset, which is a large-scale virtual autonomous driving dataset with rich weather conditions. The resolution of the images is 1920×1080. The third one is the widely used KITTI2015 dataset, including 200 training images collected in real scenarios and the image size is 375×1242. We split the training set into 160 images for training and 40 images for validation. During the training stage, we use Driving and VKITTI2 datasets as synthetic data and the 160-image split from KITTI2015 as real data, and report the performance on the 40-image validation split.

4.2 Evaluation Metrics

For stereo matching task, we use the standard End-Point Error (EPE) and D1-all metrics to evaluate the performance of the model, among which EPE is the mean average disparity error in pixels, and D1-all means the percentage of pixels whose absolute disparity error is larger than 3 pixels or 5% of ground-truth. Also, we report the percentages of erroneous pixels larger than 2, 4, and 5. For the optical-flow estimation task, besides using EPE, we also use percentage or erroneous pixels (F1-all) as evaluation metrics, which share the same definition as D1-all.

Table 3: Results on datasets from VKITTI2 to KITTI2015. We take IGEV-Stereo [45] as the stereo matching network and Unimatch-flow [48] as the optical-flow estimation network. Source only means training on VKITTI2 and directly fine-tuning on KITTI2015.

Method	EPE	D1-all(%)	>2px(%)	>4px(%)	>5px(%)	EPE(flow)	F1-all(%)
IGEV-Stereo source only	1.01	3.80	7.91	2.78	2.23	–	–
Stereo GAN [22]	0.98	3.59	7.52	2.67	2.13	–	–
Unimatch-flow source only	–	–	–	–	–	5.79	21.81
proposed	**0.93**	**3.18**	**7.04**	**2.37**	**1.90**	**5.19**	**18.32**

Table 4: Results on datasets from VKITTI2 to KITTI2015. We take DispNetC [27] as the stereo-matching network and Unimatch-flow as the optical-flow estimation network.

Method	EPE	D1-all(%)	>2px(%)	>4px(%)	>5px(%)	EPE(flow)	F1-all(%)
DispNetC source only	1.30	7.14	14.27	4.75	3.45	–	–
Stereo GAN [22]	1.27	6.78	13.02	4.70	3.50	–	–
Unimatch-flow source only	–	–	–	–	–	5.79	21.81
proposed	**1.18**	**5.98**	**11.83**	**4.12**	**3.03**	**4.98**	**19.30**

4.3 Experimental Details

We implement our algorithm using Pytorch with Adam optimizer and AdamW optimizer for stereo matching and optical flow estimation networks respectively. We scale the images to the resolution of 512×256 during the training stage. For a fair comparison with the previous GAN-based stereo matching method StereoGAN [22], we do not use data augmentation in our training stage. We empirically set the weight factors of the losses as $\lambda_{translation} = 10$, $\lambda_{f_{disp\_warpx}} = 5$, $\lambda_{f_{flow\_warpx}} = 5$, $\lambda_{corr} = 1$, $\lambda_{ms} = 0.1$, $\lambda_{disp} = 1$, $\lambda_{f_{disp\_warpy}} = 5$, $\lambda_{flow} = 1$, $\lambda_{f_{flow\_warpy}} = 5$.

4.4 Results compared with other methods

We compare our method with other methods on both domain adaptation based unsupervised stereo matching and unsupervised optical flow estimation tasks. We use Driving & KITTI2015 and VKITTI2 & KITTI2015 as our datasets, which are shown in Table 1 & 2 and Table 3 & 4 respectively. For a fair comparison, for the stereo matching task, we compare our method with StereoGAN, which is the only previous work that applies image-to-image domain translation into stereo matching, and we do not use data augmentation in the training process. Notice that for the Driving dataset, we use frames_finalpass data, which is harder than frames_cleanpass used by StereoGAN. For IGEV-Stereo and Unimatch-flow joint training, we compare the results with the source-only result on these tasks, which are trained only on the source dataset and directly tested on the KITTI2015 validation set.

As VKITTI2 dataset contains complex autonomous driving scenes, like different weather conditions (fog, clouds, and rain) and times of day (morning and sunset), there is a larger domain gap between the source and target datasets. Therefore, the improvement from VKITTI2 to KITTI2015 is not as significant as from Driving to KITTI2015. Please see the visualization in the supplementary material.

Table 5: Ablation study on datasets from Driving to KITTI2015 with different objectives. Lower value means better performance.

Model	Method	EPE	D1-all(%)	EPE(flow)	F1-all(%)
DispNet+ Unimatch-flow	baseline [22]	3.65	36.36	–	–
	w/ perceptual loss	3.46	33.16	–	–
	w/ cosine similarity loss	3.48	32.45	–	–
	w/o flow warp	3.06	31.06	11.40	37.56
	full obj.	**2.98**	**29.62**	**8.30**	**28.79**
IGEV-Stereo+Unimatch-flow	baseline	1.65	10.55	–	–
	w/ perceptual loss	1.61	10.03	–	–
	w/ cosine similarity loss	1.62	9.97	–	–
	w/o flow warp	1.60	9.57	12.22	37.96
	full obj.	**1.56**	**9.16**	**7.20**	**29.48**

Table 6: Evaluation of Real-Synthetic-Real Cycle Translation

	CycleGAN	w/disp_warpx	w/L_{per}	w/L_{cos}	w/flow_warpx	**Ours(full)**
PSNR↑	20.42	23.09	22.96	23.77	23.97	**24.50**
SSIM↑	0.8710	0.8725	0.9076	0.8802	0.9148	**0.9355**
LPIPS↓	0.2850	0.2545	0.1588	0.2053	0.1404	**0.1018**

4.5 Ablation Study

We conduct experiments to evaluate the efficiency of the loss functions to improve the effect of domain translation and the multi-scale warping loss of optical flow estimation and stereo matching, which is shown in Table 5. The perceptual loss and cosine similarity loss help the domain translation network generate images with both global and local consistency, which contribute to the training of stereo matching and optical flow estimation networks. In the training stage, we find that the feature warping loss serves as strong supervision, which not only contributes to better performance in evaluation but also contributes to the convergence of the optical flow estimation networks, which demonstrates the effectiveness of our proposed framework.

We also conduct further experiments to get the quantitative results of domain translation and compare it with CycleGAN [52] and StereoGAN [22]. We calculate the PSNR, SSIM and LPIPS between the real-synthetic-real translated image and the ground truth real image of our validation set, which reflects the domain translation ability of our model on both translation directions, shown in Table 6. It shows that our method improves the quality of domain translation.

5 Conclusion

We proposed a novel co-training framework that combines domain translation, stereo matching, and optical flow estimation. We demonstrated that models trained using our framework, which incorporates a multi-scale feature warping and a cycle-consistency loss, achieve better performance on both stereo matching and optical flow estimation tasks. The strong performance of our models on real images, all without any ground-truth labels for real images, demonstrates the effectiveness of our proposed framework in bridging the domain gap between the synthetic and real data domains.

References

[1] Xingyuan Bu, Junran Peng, Junjie Yan, Tieniu Tan, and Zhaoxiang Zhang. Gaia: A transfer learning system of object detection that fits your needs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

[2] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020.

[3] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[4] Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang. Segflow: Joint learning for video object segmentation and optical flow. In *Proceedings of the IEEE international conference on computer vision*, 2017.

[5] WeiQin Chuah, Ruwan Tennakoon, Reza Hoseinnezhad, Alireza Bab-Hadiashar, and David Suter. Itsa: An information-theoretic approach to automatic shortcut avoidance and domain generalization in stereo matching networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[6] Ning Ding, Yixing Xu, Yehui Tang, Chao Xu, Yunhe Wang, and Dacheng Tao. Source-free domain adaptation via distribution estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

[7] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Häusser, 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.

[8] Jiaxing Huang, Dayan Guan, Aoran Xiao, Shijian Lu, and Ling Shao. Category contrast for unsupervised domain adaptation in visual tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

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

[10] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. A lightweight optical flow cnn—revisiting data fidelity and regularization. *IEEE transactions on pattern analysis and machine intelligence*, 43(8):2555–2569, 2020.

[11] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[13] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G. Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[14] Taewoo Kim, Kwonyoung Ryu, Kyeongseob Song, and Kuk-Jin Yoon. Loop-net: Joint unsupervised disparity and optical flow estimation of stereo videos with spatiotemporal loop consistency. *IEEE Robotics and Automation Letters*, 5(4), 2020. doi: 10.1109/LRA.2020.3009065.

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

[16] Jiankun Li, Peisen Wang, Pengfei Xiong, Tao Cai, Ziwei Yan, Lei Yang, Jiangyu Liu, Haoqiang Fan, and Shuaicheng Liu. Practical stereo matching via cascaded recurrent network with adaptive correlation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[17] Zhengfa Liang, Yiliu Feng, Yulan Guo, Hengzhu Liu, Wei Chen, Linbo Qiao, Li Zhou, and Jianfeng Zhang. Learning for disparity estimation through feature constancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[18] Lahav Lipson, Zachary Teed, and Jia Deng. Raft-stereo: Multilevel recurrent field transforms for stereo matching. In *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021.

[19] Biyang Liu, Huimin Yu, and Guodong Qi. Graftnet: Towards domain generalized stereo matching with a broad-spectrum and task-oriented feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[20] Liang Liu, Jiangning Zhang, Ruifei He, Yong Liu, Yabiao Wang, Ying Tai, Donghao Luo, Chengjie Wang, Jilin Li, and Feiyue Huang. Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[21] Pengpeng Liu, Irwin King, Michael R Lyu, and Jia Xu. Flow2stereo: Effective self-supervised learning of optical flow and stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[22] Rui Liu, Chengxi Yang, Wenxiu Sun, Xiaogang Wang, and Hongsheng Li. Stereogan: Bridging synthetic-to-real domain gap by joint optimization of domain translation and stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[23] Wei Liu, Karoll Quijano, and Melba M Crawford. Yolov5-tassel: Detecting tassels in rgb uav imagery with improved yolov5 based on transfer learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:8085–8094, 2022.

[24] Kunming Luo, Chuan Wang, Shuaicheng Liu, Haoqiang Fan, Jue Wang, and Jian Sun. Upflow: Upsampling pyramid for unsupervised optical flow learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[25] Michael Majurski, Petru Manescu, Sarala Padi, Nicholas Schaub, Nathan Hotaling, Carl Simon Jr, and Peter Bajcsy. Cell image segmentation using generative adversarial networks, transfer learning, and augmentations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019.

[26] N. Mayer, E. Ilg, P. Häusser, 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. arXiv:1512.02134.

[27] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[28] Chaerin Min, Taehyun Kim, and Jongwoo Lim. Meta-learning for adaptation of deep optical flow networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023.

[29] Jiahao Pang, Wenxiu Sun, Chengxi Yang, Jimmy Ren, Ruichao Xiao, Jin Zeng, and Liang Lin. Zoom and learn: Generalizing deep stereo matching to novel domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[30] Faranak Shamsafar, Samuel Woerz, Rafia Rahim, and Andreas Zell. Mobilestereonet: Towards lightweight deep networks for stereo matching. In *Proceedings of the ieee/cvf winter conference on applications of computer vision*, 2022.

[31] Zhelun Shen, Yuchao Dai, and Zhibo Rao. Cfnet: Cascade and fused cost volume for robust stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[32] Xiao Song, Guorun Yang, Xinge Zhu, Hui Zhou, Zhe Wang, and Jianping Shi. Adastereo: A simple and efficient approach for adaptive stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[33] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[34] Ruoqi Sun, Xinge Zhu, Chongruo Wu, Chen Huang, Jianping Shi, and Lizhuang Ma. Not all areas are equal: Transfer learning for semantic segmentation via hierarchical region selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[35] Vladimir Tankovich, Christian Hane, Yinda Zhang, Adarsh Kowdle, Sean Fanello, and Sofien Bouaziz. Hitnet: Hierarchical iterative tile refinement network for real-time stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[36] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, 2020.

[37] Alessio Tonioni, Oscar Rahnama, Thomas Joy, Luigi Di Stefano, Thalaiyasingam Ajanthan, and Philip HS Torr. Learning to adapt for stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[38] Alessio Tonioni, Fabio Tosi, Matteo Poggi, Stefano Mattoccia, and Luigi Di Stefano. Real-time self-adaptive deep stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[39] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[40] Victor Vaquero, German Ros, Francese Moreno-Noguer, Antonio M Lopez, and Alberto Sanfeliu. Joint coarse-and-fine reasoning for deep optical flow. In *2017 IEEE International Conference on Image Processing (ICIP)*, 2017.

[41] Fan Wang, Zhongyi Han, Yongshun Gong, and Yilong Yin. Exploring domain-invariant parameters for source free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

[42] Hengli Wang, Rui Fan, Peide Cai, Ming Liu, and Lujia Wang. Undaf: A general unsupervised domain adaptation framework for disparity or optical flow estimation. In *2022 International Conference on Robotics and Automation (ICRA)*, 2022.

[43] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

[44] Aoran Xiao, Jiaxing Huang, Dayan Guan, Fangneng Zhan, and Shijian Lu. Transfer learning from synthetic to real lidar point cloud for semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2022.

[45] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. Iterative geometry encoding volume for stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[46] Haofei Xu and Juyong Zhang. Aanet: Adaptive aggregation network for efficient stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[47] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[48] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–18, 2023. doi: 10.1109/TPAMI. 2023.3298645.

[49] Keren Ye, Adriana Kovashka, Mark Sandler, Menglong Zhu, Andrew Howard, and Marco Fornoni. Spotpatch: Parameter-efficient transfer learning for mobile object detection. In *Proceedings of the Asian Conference on Computer Vision*.

[50] Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I. Jordan. Universal domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[51] Jure Zbontar, Yann LeCun, et al. Stereo matching by training a convolutional neural network to compare image patches. *J. Mach. Learn. Res.*, 17(1):2287–2318, 2016.

[52] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*, 2017.