Promoting Single-Modal Optical Flow Network for Diverse Cross-Modal Flow Estimation

Shili Zhou, Weimin Tan, Bo Yan *

School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing
Fudan University, Shanghai, China
slzhou19@fudan.edu.cn, wmtan@fudan.edu.cn, byan@fudan.edu.cn

Abstract

In recent years, optical flow methods develop rapidly, achieving unprecedented high performance. Most of the methods only consider single-modal optical flow under the well-known brightness-constancy assumption. However, in many application systems, images of different modalities need to be aligned, which demands to estimate cross-modal flow between the cross-modal image pairs. A lot of cross-modal matching methods are designed for some specific cross-modal scenarios. We argue that the prior knowledge of the advanced optical flow models can be transferred to the cross-modal flow estimation, which may be a simple but unified solution for diverse cross-modal matching tasks. To verify our hypothesis, we design a self-supervised framework to promote the single-modal optical flow networks for diverse corss-modal flow estimation. Moreover, we add a Cross-Modal-Adapter block as a plugin to the state-of-the-art optical flow model RAFT for better performance in cross-modal scenarios. Our proposed Modality Promotion Framework and Cross-Modal Adapter have multiple advantages compared to the existing methods. The experiments demonstrate that our method is effective on multiple datasets of different cross-modal scenarios.

Introduction

With the emergence of a variety of sensors, collecting images of multiple modalities for comprehensive analysis has become the best solution for many application systems, including medical diagnosis, remote sensing, monitoring, etc. However, the images of different modalities are usually non-aligned due to the limitation of devices, making it difficult for the computer to analyze automatically. Therefore, finding the pixel-level correspondences of cross-modal images is a very meaningful task.

There have been many studies on cross-modal image matching. However, the current methods have several general shortcomings: (i) Some methods can only match sparse keypoint pairs. To align the whole images, we have to assume a definite transformation rule. (ii) Most of the cross-modal methods are only aimed at matching in a specific cross-modal scene, which cannot be generalized to other scenarios. (iii) Many traditional methods do not use high-performance deep learning architectures, making them computationally inefficient and not robust.

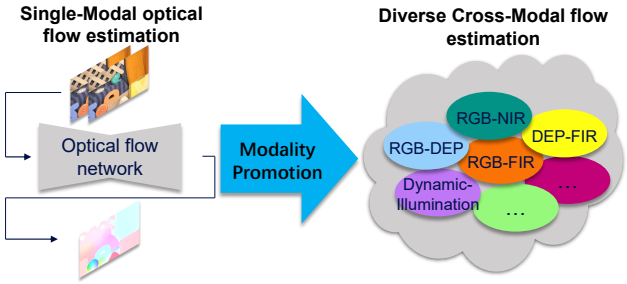


Figure 1: The illustration of the modality promotion process, which converts the existing deep-learning-based single-modal optical flow models to the cross-modality flow estimation models.

Optical flow estimation can be regarded as a subtask of image matching, which aims at estimating the dense pixel correspondences between two time-consecutive frames. In recent years, the rapid development of deep learning has aroused a huge wave in the field of optical flow estimation. Many deep models have been proposed, which greatly refresh the state-of-the-art performance of optical flow estimation. Thus, we come up with an idea: **can we make full use of the existing high-performance optical flow models in the task of cross-modal image matching?**

In this paper, we propose a Modality Promotion Framework (MPF), which can extend the existing single-modal optical flow estimation models for estimating the flow of diverse cross-modal scenarios, as shown in Figure 1. Following the self-supervised distillation process used in some unsupervised optical flow estimation methods (Liu et al. 2019a,b), we use a composite cross-modal augmentation generator to convert the ordinary RGB frame tuples to diverse cross-modal frame tuples, and construct a self-supervised framework based on the off-the-shelf optical flow estimation models. Our framework has the following

*Corresponding author: Bo Yan.

This work is supported by NSFC (Grant No.: U2001209, 61902076) and Natural Science Foundation of Shanghai (21ZR1406600).

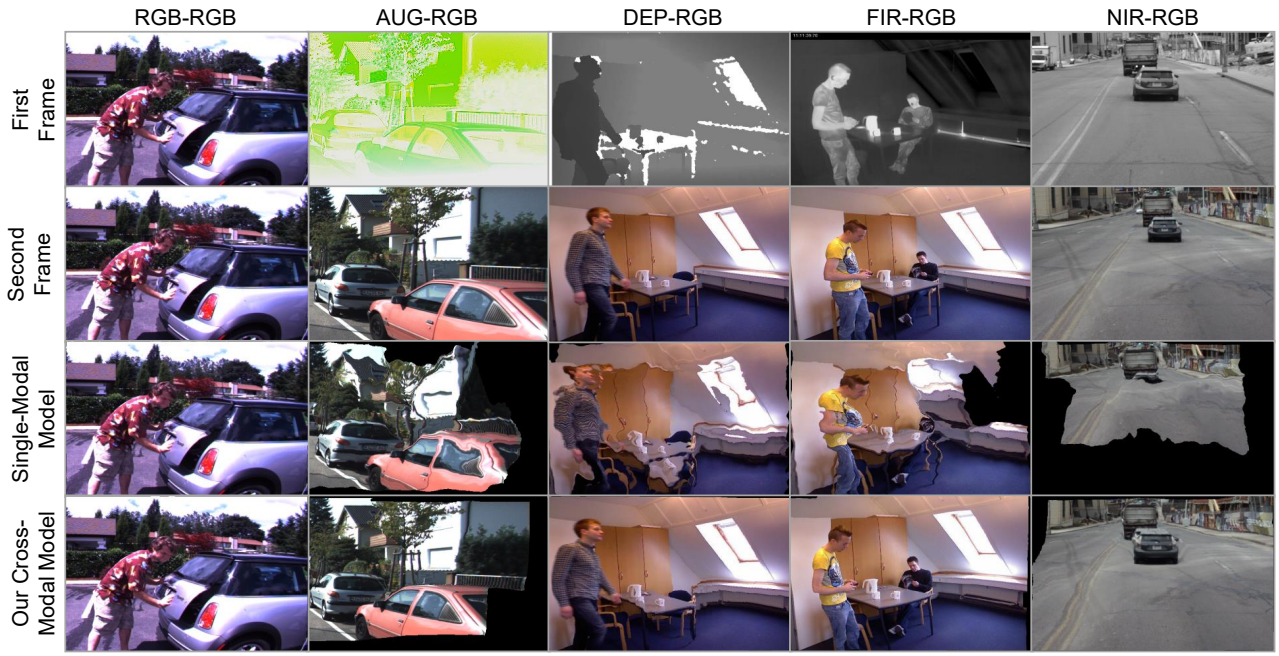


Figure 2: The single-modal optical flow estimation model versus our cross-modal flow estimation model. We show examples of inputs from different modalities in the first two rows, and show the frames warped with the flow estimated by the two models in the last two rows. The single-modal model shows good performance only in the RGB-RGB scenario, while our cross-modal modal can handle all scenarios. The AUG frame is generated by applying random color transforms on the original RGB frame. The frames of other modalities are collected with different devices.

advantages: **(i) Our framework does not rely on hard-to-obtain cross-modal datasets during training. Instead, only the ordinary RGB video clips are required. (ii) Our framework has impressive generalization ability. A single model trained in our framework can be applied to multiple cross-modal scenarios without retraining. (iii) Our framework makes full use of the prior knowledge of existing optical flow models, achieving competitive performance on multiple cross-modal datasets.** Examples to show the advantages of our cross-modal flow estimation model can be find in Figure 2.

Furthermore, we add a block named Cross-Modal Adapter (CMA) to RAFT (Teed and Deng 2020). The RAFT model extracts feature map from each frame respectively for calculating the all-pair correlations. Our CMA utilizes the cross-attention of the input image pair, and predicts a filter to generate modal-adaptive feature maps for better cross-modal flow estimation. In summary, our contributions are listed as follows:

- To exploit the existing optical flow models for cross-modal pixel-level correspondences, we propose a Modality Promotion Framework, which fine-tunes the single-modal optical flow models in a self-supervised scheme.
- To further improve the performance in cross-modal flow estimation, we propose a Cross-Modal-Adapter block as a plugin to RAFT to generate modal-adaptive feature maps for accurate flow estimation.
- We conduct extensive experiments on multiple datasets of different cross-modal scenarios, including RGBNIR-Stereo, TriModalHuman, and a dataset synthesized by ourselves. The experimental results demonstrate that our proposed MPF and CMA are able to promote the single-modal optical flow model to estimate diverse cross-modal flow.

Related Work

Cross-Modal Image Matching

Image matching is a classic task in the area of computer vision, and cross-modal image matching is a special sub-task of image matching, which aims at matching images of different modalities. Although cross-modal image matching has a wide range of applications and has been studied by many researchers, it is often studied as a dedicated task for the specific scenarios rather than a general task. As listed in a review (Jiang et al. 2021), there are dozens of cross-modal image matching methods for medical diagnosis and remote sensing, while we mainly focus on studies of daily-life scenarios.

Early on, researchers adopt the general matching methods directly to match the cross-modality images. For example, some use descriptors like SIFT (Lowe 2004), BRIEF (Butler et al. 2012), and DAISY (Tola, Lepetit, and Fua 2009) along with the post-processing matching optimization like Sift-flow (Liu, Yuen, and Torralba 2010) to solve the problem. Due to the huge differences between the modalities of the input images, these attempts perform poorly. Then, some

cross-modal descriptors are proposed, such as LSS (Shechtman and Irani 2007), DASC (Kim et al. 2015) and DSC (Kim et al. 2021), which utilize the self-similarity of images. These solutions improve the accuracy of cross-modal image matching. However, they are complicated and computationally inefficient because they are not combined with the convenient and efficient deep-learning frameworks, and the flow results generated by them are not accurate enough due to the limitation of the less-fine post-processing optimizations.

Recently, deep learning methods attract attention of many researchers. Thus, some deep-learning-based cross-modal matching methods are proposed.

Methodologically, most of the deep-learning cross-modal methods follow a Spatial-Transfer-Net (STN) and Modality-Transfer-Net (MTN) joint-learning scheme. Due to the lack of direct annotations of the correspondences in cross-modal images, researchers tacitly use an additional MTN to construct the unsupervised/semi-supervised loss function for training. The MTN transfers images of one modality to another modality without pixel-shift in the space, while the STN predicts the spatial offsets of the two input images. The core issue is how to keep the two networks functionally separated. For this purpose, (Zhi et al. 2018) relays on auxiliary material annotations, while (Liang et al. 2019) and (Jeong et al. 2019) use the networks similar to CycleGAN (Zhu et al. 2017) as their MTNs. The later one further contains a feature triplet loss as used in contrastive learning. (Arar et al. 2020) adopts different orders of the STN and MTN in the process of forward propagation for training, which decouples the two modules.

In terms of scope of application, due to the data-driven characteristic of deep-learning, the existing deep-learning cross-modal methods often aim at specific scenes. For example, (Zhi et al. 2018) and (Liang et al. 2019) are designed for VIS-NIR stereo matching in driving, while (He et al. 2019) and (Duan et al. 2020) are designed for face recognition with multi-spectral surveillance cameras. As they are trained on data of the specific scenarios, these models cannot generalize to other cross-modal datasets without retraining.

Optical Flow

Optical flow estimation is a basic technology of many image/video processing algorithms. The traditional optical flow estimation methods are based on the brightness constancy assumption, and various search and optimization algorithms are adopted to find the solution that minimizes the brightness errors of the matched pixels. Meanwhile, the deep-learning optical flow estimation methods are divided into the supervised methods and the unsupervised methods.

Supervised optical flow estimation FlowNet (Dosovitskiy et al. 2015) is the first CNN-based optical flow estimation model, and the first large-scale dataset for supervised training is proposed. This paper proves the feasibility of deep learning optical flow estimation. On the basis of FlowNet, FlowNet2 (Ilg et al. 2017) uses more complex models and more complex datasets, demonstrating the superiority of deep learning in optical flow estimation. Subsequent PWC-Net (Sun et al. 2018) and LiteFlowNet (Hui, Tang, and Loy 2018) introduce multi-level pyramids, wrapping, and local cost volume into CNN models, which greatly enhance the model performance and reduce the computational cost. VCN (Yang and Ramanan 2019) uses separate 4D convolution to take advantage of the additional spatial dimension information of cost volume. RAFT (Teed and Deng 2020) uses a pre-processing scheme that can calculate the global cost volume efficiently.

Unsupervised optical flow estimation As optical flow annotations are difficult to obtain, the existing large-scale datasets are all synthesized, which contain huge domain gaps with real-world images. The unsupervised methods rely on the brightness constancy constraint and establish a loss function on the original image pixels. In addition to the basic brightness constancy loss, there are many other techniques used for unsupervised optical flow learning, such as smooth constraint (Ren et al. 2017; Jason, Harley, and Derpanis 2016), census transformation (Meister, Hur, and Roth 2018). One difficulty of the unsupervised methods is that the pixels in the occluded area do not satisfy the brightness constancy assumption. Therefore, a variety of occlusion detection and handling schemes are proposed (Wang et al. 2018; Janai et al. 2018). Among them, the most enlightening solution is to use the unsupervised optical flow model for self-supervised distillation training (Liu et al. 2019a,b, 2020), which can effectively supervise the difficult cases such as occlusion without relying on annotations. Our method also uses a similar self-supervised framework to learn cross-modal flow.

Proposed Method

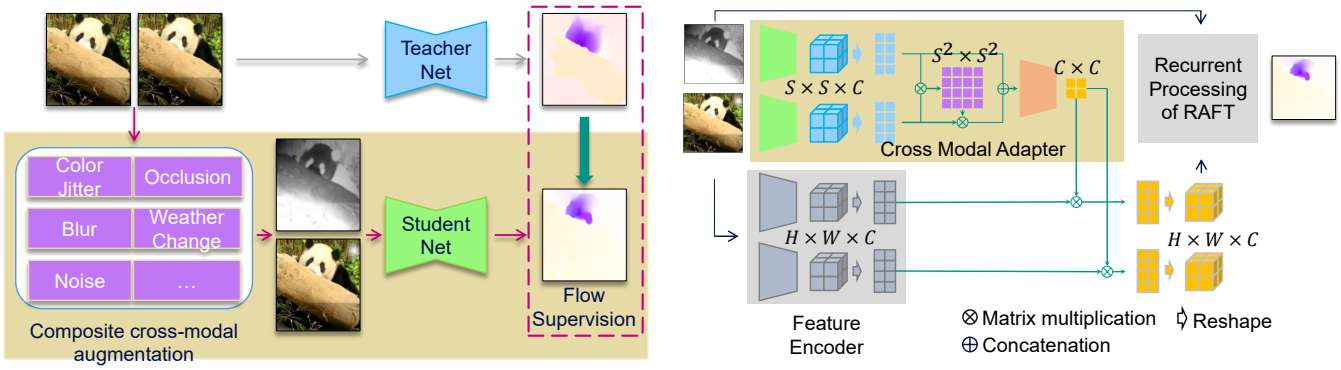
Our proposed method aims at solving the cross-modal matching problem by exploiting the existing high-performance optical flow models. In this section, we first describe our proposed Modality Promotion Framework. Then, we introduce our CrossRAFT model which adds a Cross Modal Adapter to RAFT (Teed and Deng 2020) to further increase the accuracy of cross-modal flow estimation.

Modality Promotion Framework

As shown in Figure 3 (a), the modality promotion framework needs a pre-trained optical flow network as the teacher model. For a pair of related images, we first use the teacher model to generate a pseudo-ground-truth optical flow annotation. Next, to train the student model for cross-modal flow estimation, we convert the input frames to a random cross-modal pair by a modality augmentation process. Finally, we compute the error between the output of the student model and the pseudo-ground-truth annotation, and update the model parameters with the gradients obtained by back-propagation. We use the following loss function during training:

$$\mathcal{L} = \frac{1}{HW} \sum_{x,y} (\| f^{stu}(x,y) - f^{tea}(x,y) \|_1 + \epsilon)^q, \quad (1)$$

where f^{tea} and f^{stu} are the flow maps estimated by the teacher model and the student model respectively. H, W



(a) The Modality Promotion Framework

(b) The CrossRAFT with Cross-Modal Adapter

Figure 3: (a) The illustration of our proposed Modality Promotion Framework. We use the off-the-shelf single-modal optical flow estimation model as the teacher model to generate the pseudo-ground-truth of flow. Meanwhile, the composite cross-modal augmentation convert the RGB-RGB pair to a random cross-modal pair, which is used as the input for the student model. (b) The structure of the CrossRAFT. The blocks filled with gray are structures from RAFT. The green trapezoids represent a sub-network with five CNN layers and an adaptive pooling layer. The orange trapezoid represents a sub-network with two CNN layers and a fully-connection layer.

are height and width of the flow maps, and (x, y) is the coordinate of each enumerated pixel. We empirically set the hyper-parameters $\epsilon = 0.01$ and $q = 0.4$ to reduce the influence of outlier pixels. When the student model is RAFT or our CrossRAFT, we use a sequence loss function with $\gamma = 0.8$ for the result sequence of N flow maps, as shown in formula:

$$\mathcal{L} = \sum_{i=1}^{N} \gamma^{i-N} L(f^{tea}, f_i^{stu}) \tag{2}$$

Composite Cross-modal Augmentation Existing deep-learning cross-modal matching methods often rely on huge amount of images of specific cross-modal scenarios to train an image translation model. On the contrary, we come up with the composite cross-modal augmentation process, which gives up generating data similar to the target cross-modal scenario, but generates data as diverse as possible. Therefore, the trained models not only have good performance in the specific cross-modal scenarios, but also perform well in a wide range of different scenarios. In each process of our composite cross-modal augmentation, multiple random transformations are combined haphazardly to construct inputs of a brand new cross-modal scenario.

We experiment the effect of different augmentation settings, and the results are shown in Table 1. **We can summarize a law that more diverse cross-modal scenarios in the training data bring better flow estimation performance in an unseen new cross-modal situation**. Among the tested augmentation settings, the occlusion is the most important transformation, which brings totally unreliable areas in the second frame, forcing the model to learn to estimate flow without direct pixel correspondences. Noise, sharpening, solarizing, and glass blur are the second most important transforms, which change the texture of frames to prevent the model from relying on particular texture features. Other transformations also increase the diversity of training data, leading more accurate cross-modal flow estimation of the trained models.

Model	Mean Disparity Error
w/o. ColorJitter	0.955
w/o. ChannelJitter	0.938
w/o. ColorInvert	0.943
w/o. GradualJittier	0.914
w/o. Noise	0.967
w/o. Sharpen	0.977
w/o. Blur	0.892
w/o. Occlusion	1.062
w/o. Weather	0.925
w/o. Compression Noise	0.941
w/o. Solarize	0.972
w/o. GlassBlur	0.967
Full augumentation	**0.889**

Table 1: Experiments for different augmentation settings on the RGBNIR-Stereo dataset. The best row is marked in bold.

Cross Modal Adapter

Our goal is to estimate flow of diverse cross-modal scenarios with a single trained model. However, the feature extractors in existing flow estimation models are fixed. It generates the same feature map for the same input frame, although the other frame in the pair may change to different modalities. We argue that a fixed feature extractor cannot work well in different cross-modal scenarios.

In order to solve this problem, we modify the structure of an existing optical flow modal named RAFT (Teed and Deng 2020) to make it more suitable for cross-modal scenes. As shown in Figure 3 (b), we add a Cross Modal Adapter (CMA) to adaptively alter the features extracted from the feature encoder in RAFT. The detailed process of the CMA

is as follows:

First, we extract feature maps from the two input frames respectively, and scale them to a fixed size ($S \times S \times C$) by using the adaptive average pooling layer. Then, after reshaping the two feature maps to matrices of size $S^2 \times C$, an attention matrix M is calculated as the following formulas:

$$M = F_1 F_2^T, \tag{3}$$

$$M_{ij}^s = \frac{exp(M_{ij})}{\sum_{k=1}^{S^2} exp(M_{ik})}, \tag{4}$$

where F_1, F_2 are the two feature matrices of the first and second frame, and M^s is the row-normalized matrix get by applying softmax operation on M.

The calculated M^s is used to coarsely align the feature matrices of the two frames. The aligned features are then inputted to a small adapter-generation network G consisting of two convolution layers and a fully-connection layer to get a modality adapter matrix. The two steps can be shown as the following formulas:

$$\tilde{F_2} = M^s F_2, \tag{5}$$

$$O = G([F_1, \tilde{F_2}]), \tag{6}$$

where $\tilde{F_2}$ is the aligned F_2, $[\cdot, \cdot]$ is concatenating operation, and O is the needed adapter matrix of size $C \times C$. The O matrix changes with the inputs of different cross-modal scenarios, and patches up the original feature extractor in RAFT for better adapting to the new cross-modal input pair. The feature enhance process is shown as formulas:

$$\hat{F}_1^{RAFT} = F_1^{RAFT} O, \tag{7}$$

$$\hat{F}_2^{RAFT} = F_2^{RAFT} O, \tag{8}$$

where F_i^{RAFT} is the RAFT feature map of the i-th frame, $\hat{F}$ is the altered feature map. For brevity, we have omitted some reshape operations in the formulas.

To explain in another way, our proposed Cross-Modal Adapter generates adaptive 1×1 convolution filters for different cross-modal inputs, making the feature extractor in our CrossRAFT able to generalize to more cross-modal scenarios and even some unseen brand-new scenarios. The motivation is similar to few-shot learning, and the effectiveness of our CMA is proved in the experiment section.

Experiment

Experimental Settings

We implement our framework and models in PyTorch. All experiments are conducted on a single NVIDIA RTX2080Ti GPU with a Intel Core i7-9700K@3.60GHz CPU (32G RAM). As our models are based on the existing optical flow networks, the open-source PWC-Net (Sun et al. 2018) and RAFT (Teed and Deng 2020) code and weights are utilized. The pre-trained RAFT is chosen as the teacher model for all experiments. We use AdamW (Loshchilov and Hutter 2017) optimizer with learning-rate=0.00002 and weight-decay=0.00005 to train the models for 10k steps with batch-size=4. The weights of pre-trained optical flow models are loaded to the student models as an initialization. For our CrossRAFT, we load the RAFT weights for the unmodified layers. The data augmentations is implemented with Albumentations (Buslaev et al. 2020).

Datasets

We use YoutubeVOS dataset (Xu et al. 2018) as the training set. It contains 4,000+ video clips collected from Youtube and the corresponding high-quality segmentation annotations, while we only use the video frames. For every step, we randomly sample two frames with frame-interval less than 20. Three datasets are used to evaluate models for cross-modal flow estimation. They are RGBNIR-Stereo (Zhi et al. 2018), TriModalHuman (Palmero et al. 2016) and a dataset synthesized by ourselves named CrossKITTI, which will be described in detail when introducing the experiment results.

Ablation Study

Effect of the Modality Promotion Framework We compare different models including the off-the-shelf PWC-Net (Sun et al. 2018), the off-the-shelf RAFT (Teed and Deng 2020), the PWC-Net and RAFT fine-tuned in our Modality Promotion Framework. The evaluations are conducted on two datasets. One of them is the RGBNIR-Stereo dataset, which is taken by the vehicle-mounted RGB-NIR binocular camera. It contains 12 sequences of RGB-NIR image pairs. Among them, 4 sequences composed of 2,000 image pairs have disparity annotations of some sparse key-points. The evaluation metric is ADE (Average Disparity Error). Lower ADE means more accurate estimation. Though our models estimate the two-dimensional flow, we directly its horizontal component as the predicted disparity.

The results are shown in Table 2, and a visual example is shown in Figure 4. The off-the-shelf optical flow models perform poorly for cross-modal inputs. Meanwhile, the fine-tuned models have made significant progress, which demonstrates the effectiveness of our proposed Modality Promotion Framework.

Another evaluation is conducted on the CrossKITTI dataset, which is synthesized by ourselves by applying our composite cross-modal augmentation on the famous KITTI-2012 dataset (Geiger, Lenz, and Urtasun 2012). The KITTI-2012 dataset is collected with a set of in-vehicle sensors, and has sparse optical flow annotations for some real-world frames. We randomly select 59 image pairs with flow annotations in KITTI-2012 to construct our CrossKITTI. AEPE (Average Endpoint Error) and F1 (Percentage of Optical Flow Outliers) are used for evaluating the flow accuracy. The results are shown in Table 3, which demonstrates that our framework also works on the CrossKITTI dataset.

Effect of the Cross-Modal Adapter To verify the capability of our proposed Cross-Modal Adapter, we also conduct experiments to compare the fine-tuned RAFT model and our CrossRAFT model on the two datasets. The results can be found in Table 2 and Table 3. It shows that the CrossRAFT with the Cross-Modal Adapter achieves better performance on different cross-modal scenarios.

Model	Common	Light	Glass	Glossy	Vegetation	Skin	Clothing	Bag	**Mean**	**Reduction Ratio**
PWC-Net	0.58	0.89	1.30	1.64	2.55	1.56	1.21	1.24	**1.37**	**baseline**
PWC-Net + MPF	0.62	0.81	1.19	**1.35**	1.02	1.47	0.79	0.93	1.02	**-25.55%**
RAFT	0.81	5.22	1.17	1.37	1.73	2.13	4.64	3.49	**2.57**	**baseline**
RAFT + MPF	**0.49**	2.92	1.13	1.37	**0.98**	1.04	**0.74**	0.83	**1.19**	**-53.70%**
CrossRAFT + MPF	0.54	**0.59**	**1.10**	**1.35**	1.01	**0.95**	0.78	**0.80**	0.89	**-65.37%**

Table 2: Ablation study on the RGBNIR-Stereo dataset. MPF means our Modality Promotion Framework. The best value of each column is bold, and the second best value of each column is marked with underline. The last two bold columns are the mean ADE and the relative change rate to the baseline models.

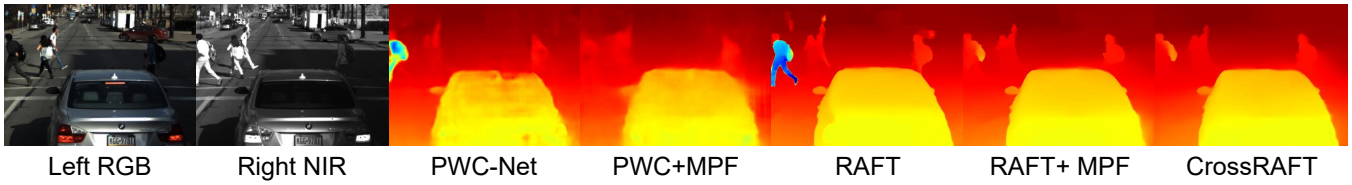


Figure 4: A visual example of the ablation study on the RGBNIR-Stereo dataset.

Model	AEPE (NOC)	F1 (NOC)	AEPE (ALL)	F1 (ALL)
PWC-Net	8.915	0.309	11.345	0.341
PWC-Net + MPF	4.749	0.284	8.407	0.351
RAFT	17.413	0.285	19.198	0.312
RAFT + MPF	2.216	0.102	4.392	0.147
CrossRAFT + MPF	**1.947**	**0.091**	**3.719**	**0.136**

Table 3: Ablation study on the CrossKITTI dataset. NOC means the metrics are calculated in no occlusion areas, and ALL means the metrics are calculated in all areas. The best value of each column is bold, and the second best value of each column is marked with underline.

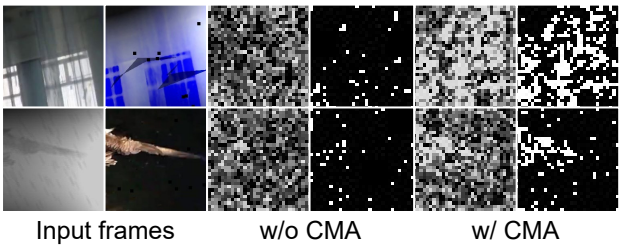


Figure 5: Examples of WTA matching of static transformed image pairs. The metric for matching is cosine distance. The correctly matched pixels in the soft/hard accuracy maps is marked in white, and the mismatched pixels are marked in gray in the soft maps. The degree of gray is determined by the matching offset.

To further demonstrate the effectiveness of the Cross-Modal Adapter, we design another experiment to show the matching accuracy between features of an image pair generated from the same static image. We use Winner-Take-All (WTA) strategy to matching pixels in local 5×5 windows with the original and modified feature maps respectively. Obviously, there is no shift between the pair of images, so the ground-truth flow is zero-flow. We show two examples in Figure 5. As we can see, the feature maps modified by the CMA lead to better matching accuracy.

Comparisons with the State-of-the-Art Methods

Evaluation on RGBNIR-Stereo We compare our CrossRAFT with CMA (Chiu, Blanke, and Fritz 2011), ANCC (Heo, Lee, and Lee 2010), DASC (Kim et al. 2015), GLU-Net (Truong, Danelljan, and Timofte 2020), RANSAC-Flow (Shen et al. 2020), DMC (Zhi et al. 2018), UCSS (Liang et al. 2019) and TBA (Walters et al. 2021). The results are shown in Table 4. The listed DMC is trained without additional segmentation annotations.

The first three methods are traditional methods, which only achieve limited performance. The next two methods are deep-learning-based single-modal matching methods, which are unable to work well in cross-modal scenarios. The following three methods are deep-learning-based cross-modal methods, they get more precise prediction results. However, these methods needs specifically training on the datasets of the same domain, which limits their scope of application. Meanwhile, our CrossRAFT is not trained or fine-tuned on the RGBNIR-Stereo dataset, but also achieves competitive performance.

Evaluation on TriModalHuman The TriModalHuman dataset contains 11,537 RGB-Depth-FIR triplets of three indoor scenes, and 5,724 of them have human segmentation annotations. Due to the lack of direct annotations of matching, we follow (Kim et al. 2021) to evaluate the quality of the warped segmentation labels instead. Two metrics are used in this evaluation: LTA (Label Transfer Accuracy) and IoU (Intersection over Union). The results are listed in Table 5.

Model	Common	Light	Glass	Glossy	Vegetation	Skin	Clothing	Bag	Mean
CMA	1.60	5.17	2.55	3.86	4.42	3.39	6.42	4.63	4.00
ANCC	1.36	2.43	2.27	2.41	4.82	2.32	2.85	2.57	2.63
DASC	0.82	1.24	1.50	1.82	1.09	1.59	0.80	1.33	1.28
GLU-Net	3.12	0.76	1.35	1.56	2.58	1.49	**0.77**	0.88	1.56
RANSAC-Flow	2.80	1.56	1.50	1.78	1.78	2.50	1.77	3.89	2.20
DMC	**0.51**	1.08	1.05	1.57	**0.69**	1.01	1.22	0.90	1.00
UCSS	0.68	0.80	0.67	**1.05**	0.68	1.04	0.98	**0.80**	**0.84**
TBA	0.91	0.90	**0.64**	1.18	1.49	1.00	1.47	1.10	1.08
Ours	0.54	**0.59**	1.10	1.35	1.01	**0.95**	0.78	**0.80**	0.89

Table 4: Comparisons with the state-of-the-art cross-modal matching methods on the RGBNIR-Stereo dataset. The best value of each column is bold, and the second best value is marked with underline.

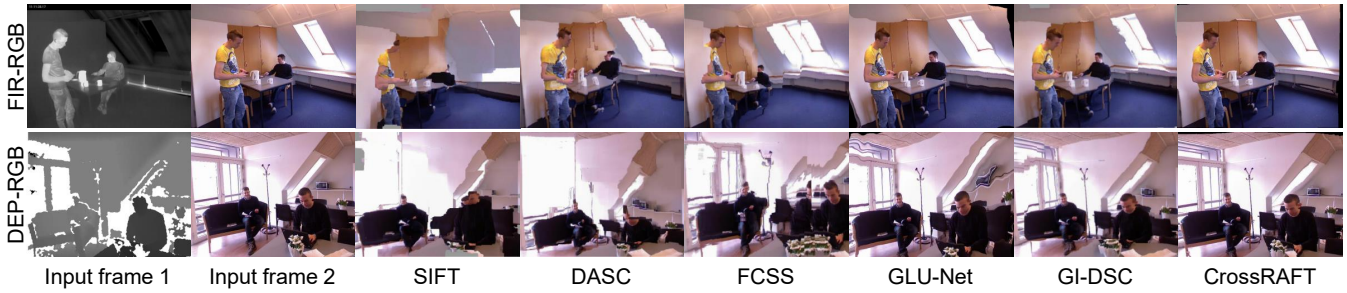


Figure 6: Qualitative comparisons between different methods. The upper row lists an example of FIR-RGB matching, and the lower row lists an example of DEP-RGB matching. The first two columns list the input frames, and rest columns list the warped RGB frames of different methods.

Method	RGB-DEP		RGB-FIR		DEP-FIR	
	1-LTA	1-IoU	1-LTA	1-IoU	1-LTA	1-IoU
DAISY	0.46	0.41	0.36	0.44	0.53	0.52
BRIEF	0.46	0.46	0.48	0.41	0.57	0.53
LSS	0.49	0.52	0.49	0.42	0.52	0.42
LIOP	0.42	0.37	0.48	0.36	0.51	0.39
DaLI	0.41	0.39	0.49	0.43	0.54	0.50
DASC	0.37	0.36	0.38	0.39	0.44	0.41
VGG	0.33	0.39	0.38	0.42	0.47	0.38
LIFT	0.39	0.47	0.44	0.49	0.51	0.53
L2-Net	0.36	0.41	0.39	0.38	0.43	0.47
FCSS	0.31	0.31	0.30	0.30	0.40	0.34
GLU-Net	0.37	0.56	**0.18**	0.26	0.21	0.28
RANSAC-Flow	0.19	0.29	0.23	0.35	0.31	0.47
SSC	0.30	0.29	0.31	0.31	0.43	0.36
DSC	0.26	0.24	0.29	0.27	0.36	0.27
GI-DSC	0.23	**0.19**	0.27	**0.24**	0.31	**0.21**
Ours	**0.15**	**0.19**	**0.18**	0.27	**0.17**	0.27

Table 5: Comparisons on the TriModalHuman dataset. The best value of each column is bold, and the second best value is marked with underline. Higher LTA and IoU indicate better performance, while we list 1-LTA and 1-IoU as the error rates.

DAISY (Tola, Lepetit, and Fua 2009), BRIEF (Butler et al. 2012), LSS (Shechtman and Irani 2007), LIOP (Wang, Fan, and Wu 2011), DaLI (Simo-Serra, Torras, and Moreno-Noguer 2015) and DASC (Kim et al. 2015) are earlier traditional matching methods. VGG (Simonyan and Zisserman 2014), LIFT (Yi et al. 2016), L2-Net (Tian, Fan, and Wu 2017) and FCSS (Kim et al. 2017) are four deep-learning models trained for finding matching descriptors. GLU-Net (Truong, Danelljan, and Timofte 2020) and RANSAC-Flow (Shen et al. 2020) are deep-learning methods for dense matching. SSC, DSC and GI-DSC are the state-of-the-art cross-modal descriptors proposed in (Kim et al. 2021). Our method is on par with the state-of-the-art methods when considering the IoU metric, and achieves the best performance under the LTA metric for all three cross-modal scenarios.

Furthermore, we qualitatively compare the methods in Figure 6. Unlike other methods, the results of our CrossRAFT are not only accurate but also smooth. On the contrary, GLU-Net cannot handle cross-modal scenarios well, thus leads to mismatches in some area. Meanwhile, the rest methods only give descriptors and need a post-processing with SIFT-flow (Liu, Yuen, and Torralba 2010), which may cause obvious artifacts.

Conclusion

In this paper, we propose a Modality Promotion Framework to promote the off-the-shelf single-modal optical flow networks for cross-modal flow estimation. Our framework adopts a self-supervision manner and does not need the specific cross-modal datasets for training. Furthermore, we propose the CrossRAFT model with a Cross-Modal-Adapter as a plugin to RAFT, which can enhance the cross-modal feature extraction ability for RAFT. The experiments demonstrate that our proposed framework and CrossRAFT are effective for cross-modal flow estimation.

References

Arar, M.; Ginger, Y.; Danon, D.; Bermano, A. H.; and Cohen-Or, D. 2020. Unsupervised multi-modal image registration via geometry preserving image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13410–13419.

Buslaev, A.; Iglovikov, V. I.; Khvedchenya, E.; Parinov, A.; Druzhinin, M.; and Kalinin, A. A. 2020. Albumentations: Fast and Flexible Image Augmentations. *Information*, 11(2).

Butler, D. J.; Wulff, J.; Stanley, G. B.; and Black, M. J. 2012. A naturalistic open source movie for optical flow evaluation. In *European conference on computer vision*, 611–625. Springer.

Chiu, W. W.-C.; Blanke, U.; and Fritz, M. 2011. Improving the Kinect by Cross-Modal Stereo. In *BMVC*, volume 1, 3. Citeseer.

Dosovitskiy, A.; Fischer, P.; Ilg, E.; Hausser, P.; Hazirbas, C.; Golkov, V.; van der Smagt, P.; Cremers, D.; and Brox, T. 2015. FlowNet: Learning Optical Flow With Convolutional Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Duan, B.; Fu, C.; Li, Y.; Song, X.; and He, R. 2020. Cross-spectral face hallucination via disentangling independent factors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7930–7938.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, R.; Cao, J.; Song, L.; Sun, Z.; and Tan, T. 2019. Adversarial cross-spectral face completion for NIR-VIS face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 42(5): 1025–1037.

Heo, Y. S.; Lee, K. M.; and Lee, S. U. 2010. Robust stereo matching using adaptive normalized cross-correlation. *IEEE Transactions on pattern analysis and machine intelligence*, 33(4): 807–822.

Hui, T.-W.; Tang, X.; and Loy, C. C. 2018. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8981–8989.

Ilg, E.; Mayer, N.; Saikia, T.; Keuper, M.; Dosovitskiy, A.; and Brox, T. 2017. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2462–2470.

Janai, J.; Guney, F.; Ranjan, A.; Black, M.; and Geiger, A. 2018. Unsupervised Learning of Multi-Frame Optical Flow with Occlusions. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Jason, J. Y.; Harley, A. W.; and Derpanis, K. G. 2016. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In *European Conference on Computer Vision*, 3–10. Springer.

Jeong, S.; Kim, S.; Park, K.; and Sohn, K. 2019. Learning to find unpaired cross-spectral correspondences. *IEEE Transactions on Image Processing*, 28(11): 5394–5406.

Jiang, X.; Ma, J.; Xiao, G.; Shao, Z.; and Guo, X. 2021. A review of multimodal image matching: Methods and applications. *Information Fusion*.

Kim, S.; Min, D.; Ham, B.; Jeon, S.; Lin, S.; and Sohn, K. 2017. Fcss: Fully convolutional self-similarity for dense semantic correspondence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6560–6569.

Kim, S.; Min, D.; Ham, B.; Ryu, S.; Do, M. N.; and Sohn, K. 2015. DASC: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2103–2112.

Kim, S.; Min, D.; Lin, S.; and Sohn, K. 2021. Dense Cross-Modal Correspondence Estimation With the Deep Self-Correlation Descriptor. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7): 2345–2359.

Liang, M.; Guo, X.; Li, H.; Wang, X.; and Song, Y. 2019. Unsupervised cross-spectral stereo matching by learning to synthesize. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 8706–8713.

Liu, C.; Yuen, J.; and Torralba, A. 2010. Sift flow: Dense correspondence across scenes and its applications. *IEEE transactions on pattern analysis and machine intelligence*, 33(5): 978–994.

Liu, L.; Zhang, J.; He, R.; Liu, Y.; Wang, Y.; Tai, Y.; Luo, D.; Wang, C.; Li, J.; and Huang, F. 2020. 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)*.

Liu, P.; King, I.; Lyu, M. R.; and Xu, J. 2019a. Ddflow: Learning optical flow with unlabeled data distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 8770–8777.

Liu, P.; Lyu, M.; King, I.; and Xu, J. 2019b. Selflow: Self-supervised learning of optical flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4571–4580.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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

Meister, S.; Hur, J.; and Roth, S. 2018. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Palmero, C.; Clapés, A.; Bahnsen, C.; Møgelmose, A.; Moeslund, T. B.; and Escalera, S. 2016. Multi-modal rgb–depth–thermal human body segmentation. *International Journal of Computer Vision*, 118(2): 217–239.

Ren, Z.; Yan, J.; Ni, B.; Liu, B.; Yang, X.; and Zha, H. 2017. Unsupervised deep learning for optical flow estimation. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Shechtman, E.; and Irani, M. 2007. Matching local self-similarities across images and videos. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8. IEEE.

Shen, X.; Darmon, F.; Efros, A. A.; and Aubry, M. 2020. Ransac-flow: generic two-stage image alignment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, 618–637. Springer.

Simo-Serra, E.; Torras, C.; and Moreno-Noguer, F. 2015. DaLI: deformation and light invariant descriptor. *International journal of computer vision*, 115(2): 136–154.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sun, D.; Yang, X.; Liu, M.-Y.; and Kautz, J. 2018. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8934–8943.

Teed, Z.; and Deng, J. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, 402–419. Springer.

Tian, Y.; Fan, B.; and Wu, F. 2017. L2-net: Deep learning of discriminative patch descriptor in euclidean space. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 661–669.

Tola, E.; Lepetit, V.; and Fua, P. 2009. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *IEEE transactions on pattern analysis and machine intelligence*, 32(5): 815–830.

Truong, P.; Danelljan, M.; and Timofte, R. 2020. GLU-Net: Global-local universal network for dense flow and correspondences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6258–6268.

Walters, C.; Mendez, O.; Johnson, M.; and Bowden, R. 2021. There and Back Again: Self-supervised Multispectral Correspondence Estimation. *arXiv preprint arXiv:2103.10768*.

Wang, Y.; Yang, Y.; Yang, Z.; Zhao, L.; Wang, P.; and Xu, W. 2018. Occlusion Aware Unsupervised Learning of Optical Flow. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, Z.; Fan, B.; and Wu, F. 2011. Local intensity order pattern for feature description. In *2011 International Conference on Computer Vision*, 603–610. IEEE.

Xu, N.; Yang, L.; Fan, Y.; Yang, J.; Yue, D.; Liang, Y.; Price, B.; Cohen, S.; and Huang, T. 2018. YouTube-VOS: Sequence-to-Sequence Video Object Segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Yang, G.; and Ramanan, D. 2019. Volumetric correspondence networks for optical flow. *Advances in neural information processing systems*, 32: 794–805.

Yi, K. M.; Trulls, E.; Lepetit, V.; and Fua, P. 2016. Lift: Learned invariant feature transform. In *European conference on computer vision*, 467–483. Springer.

Zhi, T.; Pires, B. R.; Hebert, M.; and Narasimhan, S. G. 2018. Deep material-aware cross-spectral stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1916–1925.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.