MDFlow: Unsupervised Optical Flow Learning by Reliable Mutual Knowledge Distillation

Lingtong Kong and Jie Yang, Senior Member, IEEE

Abstract—Recent works have shown that optical flow can be learned by deep networks from unlabelled image pairs based on brightness constancy assumption and smoothness prior. Current approaches additionally impose an augmentation regularization term for continual self-supervision, which has been proved to be effective on difficult matching regions. However, this method also amplifies the inevitable mismatch in unsupervised setting, blocking the learning process towards optimal solution. To break the dilemma, we propose a novel mutual distillation framework to transfer reliable knowledge back and forth between the teacher and student networks for alternate improvement. Concretely, taking estimation of off-the-shelf unsupervised approach as pseudo labels, our insight locates at defining a confidence selection mechanism to extract relative good matches, and then add diverse data augmentation for distilling adequate and reliable knowledge from teacher to student. Thanks to the decouple nature of our method, we can choose a stronger student architecture for sufficient learning. Finally, better student prediction is adopted from teacher to student. Thanks to the decouple nature of our method, we can choose a stronger student architecture for sufficient learning. Finally, better student prediction is adopted from teacher to student. Thanks to the decouple nature of our method, we can choose a stronger student architecture for sufficient learning. Finally, better student prediction is adopted from teacher to student. Thanks to the decouple nature of our method, we can choose a stronger student architecture for sufficient learning. Finally, better student prediction is adopted from teacher to student. Thanks to the decouple nature of our method, we can choose a stronger student architecture for sufficient learning. Finally, better student prediction is adopted from teacher to student. Thanks to the decouple nature of our method, we can choose a stronger student architecture for sufficient learning. Finally, better student prediction is adopted from teacher to student. Thanks to the decouple nature of our method, we can choose a stronger student architecture for sufficient learning. Finally, better student prediction is adopted from teacher to student. Thanks to the decouple nature of our method, we can choose a stronger student architecture for sufficient learning. Finally, better student prediction is adopted from teacher to student. Thanks to the decouple nature of our method, we can choose a stronger student architecture for sufficient learning. Finally, better student prediction is adopted from teacher to student. Thanks to the decouple nature of our method, we can choose a stronger student architecture for sufficient learning. Finally, better student prediction is adopted from teacher to student. Thanks to the decouple nature of our method, we can choose a stronger student architecture for sufficient learning. Finally, better student prediction is adopted from teacher to student. Thanks to the decouple nature of our method, we can choose a stronger student architecture for sufficient learning. Finally, better student prediction is adopted from teacher to student. Thanks to the decouple nature of our method, we can choose a stronger student architecture for sufficient learning. Finally, better student prediction is adopted from teacher to student. Thanks to the decouple nature of our method, we can choose a stronger student architecture for sufficient learning. Finally, better student prediction is adopted from teacher to student. Thanks to the decouple nature of our method, we can choose a stronger student architecture for sufficient learning. Finally, better student prediction is adopted from teacher to student. Thanks to the decouple nature of our method, we can choose a stronger student architecture for sufficient learning. Finally, better student prediction is adopted from teacher to student. Thanks to the decouple nature of our method, we can choose a stronger student architecture for sufficient learning. Finally, better student prediction is adopted from teacher to student. Thanks to the decouple nature of our method, we can choose a stronger student architecture for sufficient learning. Finally, better student prediction is adopted from teacher to student. Thanks to the decouple nature of our method, we can choose a stronger student architecture for sufficient learning. Finally, better student prediction is adopted from teacher to student. Thanks to the decouple nature of our method, we can choose a stronger student architecture for sufficient learning. Finally, better student prediction is adopted from teacher to student. Thanks to the decouple nature of our method, we can choose a stronger student architecture for sufficient learning. Finally, better student prediction is adopted from teacher to student. Thanks to the decouple nature of our method, we can choose a stronger student architecture for sufficient learning. Finally, better student prediction is adopted from teacher to student. Thanks to the decouple nature of our method, we can choose a stronger student architecture for sufficient learning. Finally, better student prediction is adopted from teacher to student. Thanks to the decouple nature of our method, we can choose a stronger student architecture for sufficient learning. Finally, better student prediction is adopted from teacher to student. Thanks to the decouple nature of our method, we can choose a stronger student architecture for sufficient learning. Finally, better student prediction is adopted from teacher to student. Thanks to the decouple nature of our method, we can choose a stronger student architecture for sufficient learning. Finally, better student prediction is adopted from teacher to student. Thanks to the decouple nature of our method, we can choose a stronger student architecture for sufficient learning.

One alternative pipeline is to leverage countless unlabeled video sequences for unsupervised learning, which could presumably produce satisfying results without suffering from any domain mismatch. These methods usually build objective functions based on brightness constancy and spatial smoothness [18], [19], and further include occlusion handling mechanism [20], [21]. Most recently, the state-of-the-art ARFlow [22], UFlow [23] and UPFlow [24] propose to integrate augmentation regularization term into each iteration step for continuous improvement. Although their self-learning approach can improve flow accuracy, the unavoidable mismatch in predicted pseudo labels can mislead the network conversely. As a result, the optimization procedure is blocked by such match noise even with more training iterations. This puts up a question: Can we decouple such mismatching regions in augmentation regularization for further accuracy improvement?

On the other hand, comparing with recent progress in deep flow architectures [13], [14], most of recent top unsupervised approaches [22], [23], [25], [26] adopt the PWC-Net [11] as flow estimation model, for its well-behaved speed-accuracy trade-off and efficiency in real deployment. Intuitively, we can achieve better accuracy by naively replacing PWC-Net [11] with superior flow architecture, such as RAFT [13], like the current state-of-the-art method SMURF [27]. However, this is at the cost of extra computation and delay during inference. In addition, more complex structure may lead to lower performance when guided by unsupervised losses [23], which brings another question: Can unsupervised optical flow benefit from recent and future advanced flow architecture without introducing additional cost at inference?

In this paper, we jointly address the above two questions with a novel unsupervised optical flow learning framework, termed MDFlow, by exploring reliable knowledge distillation between the teacher and student networks for alternative improvement. A high-level abstract of proposed mutual...
for decades for its fundamental role in many downstream vision tasks. FlowNet [9] and its successor FlowNet2 [10] are the first attempt to apply deep learning methods for optical flow estimation, which directly regress flow field based on the encoder-decoder U-shape network or its cascaded version. Inspired by traditional coarse-to-fine pipeline [6], [8], [31], SPyNet [32], PWC-Net [11] and LiteFlowNet [33] employ pyramid, warping and cost volume into end-to-end learning and achieve impressive real-time performance. After that, IRR-PWC [12] iteratively and jointly estimates residual flow and occlusion with shared estimators across pyramid levels. The work [34] introduces a dual self-attention module to improve original pyramid flow network. Mask-FlowNet [35] and OAS-Net [36] explore to resolve ambiguous matching caused by warping operation. As for more efficient architecture, FDFlowNet [37] introduces U-shape backbone and partial fully connected decoder for compact structure. FastFlowNet [38] constructs center dense dilated correlation layer and shuffle block decoder to reduce computation and first achieves real-time performance on embedded systems. Recently, DICL-Flow [14] transfers concatenated feature volume in stereo matching to optical flow and propose displacement-invariant cost learning. Different from above coarse-to-fine pipeline, RAFT [13] first calculates all pairs’ correspondences and then introduces a recurrent module to estimate residual flow and update multi-scale matching cost simultaneously, which achieves significant accuracy improvement. Later on, GMA [39] introduces a global motion aggregation module based on transformer to improve recurrent-based flow architecture and obtains state-of-the-art accuracy.

**B. Learning Unsupervised Optical Flow**

Due to prohibitive cost to acquire optical flow ground truth of real-world images, alternative approaches try to leverage countless unlabeled video sequences for unsupervised learning. These methods build objective functions based on brightness constancy assumption and local smoothness prior [18], [19]. Then, UnFlow [20] and OccAwareFlow [21] boost unsupervised performance by excluding residual calculation in reasoned occluded regions. However, the occluded areas are only guided by rigid smoothness constraint, that can damage overall accuracy. To handle this problem, DDFlow [40] and SelFlow [25] distill flow estimation from the teacher model to the student with random crop and occlusion hallucination in a data-driven manner, that further improves unsupervised accuracy. However, the teacher cannot benefit from the improved student. STFlow [41] integrates variational refinement with unsupervised learning and propose a self-taught framework for continual improvement. SimFlow [26] explores learnable feature similarity for regulating previous census reconstruction loss. ARFlow [22] and UFlow [23] propose to integrate augmentation regularization into each iteration step for continuous improvement of a single model. CoT-AMFlow [42] develops an adaptive modulation network and adopts a co-teaching strategy for better accuracy. Later on, UPFlow [24] improves the upsampling unit of PWC-Net and proposes a better pyramid distillation loss.

---

**II. RELATED WORK**

**A. Deep Optical Flow Architecture**

Finding dense correspondences between a pair of time adjacent frames, namely optical flow estimation, has been studied
DistillFlow [43] trains multiple teacher models and introduces a confidence based two-stage distillation approach for improvement. OffFlow [44] puts up an occlusion-inpainting framework to make full use of occlusion regions. Recently, ASPFlow [45] presents content-aware pooling and adaptive flow upsampling modules to improve pyramid-based unsupervised flow deep structure. MRDFlow [46] further introduces 4D correlation layer and recurrent decoder of RAFT to strengthen flow estimation network. At the same time, SMURF [27] replaces PWC-Net with a more powerful RAFT backbone and proposes a sequence-aware self-supervision loss to achieve state-of-the-art accuracy. However, it incurs huge computation cost and can not satisfy many real-time applications. Different from above methods, we separate and recombine multiple objective parts into a mutual distillation framework for reliable performance improvement without increasing model size and inference delay.

Another line of unsupervised optical flow methods resort to additional information beyond two input frames. Based on the setting of stereo video systems, UnOS [47] enforces geometry constraints among stereo depth, camera ego-motion and optical flow for mutual promotion. Flow2Stereo [48] introduces data distillation into the joint learning framework of optical flow and stereo matching. Most recently, the work [49] shows that feature-level collaboration of the networks for optical flow, stereo depth and camera motion can outperform previous methods that only consider loss-level joint optimization. To facilitate unsupervised optical flow in challenging scenes, such as fog, rain and night, GyroFlow [50] converts gyroscope readings into gyro field and fuse it with flow information for recovering more motion details. Different from these methods that resort to additional information, proposed MDFlow focuses on improving the basic setting by reducing matching noise and exploring better flow architecture during the decoupled bidirectional distillation procedure.

C. Knowledge Distillation and Mutual Distillation

Knowledge Distillation (KD) is usually exploited to train a compact student network by mimicking the output distribution of a pre-trained complex teacher model as well as one-hot ground-truth labels [51]. Following variants [52], [53], [54], [55] mainly focus on utilizing intermediate information of teacher, such as feature maps or attention masks as extra hints for improvement. Different from above one-way transfer between a teacher and a student in knowledge distillation, Deep Mutual Learning (DML) [56] finds that mutual learning of a collection of simple student networks can outperform distillation from a more powerful yet static teacher. The work EKD [57] introduces an evolutionary teacher that can enable more efficient knowledge transfer by minimizing the capability gap between teacher and student. Dense Cross-layer Mutual-distillation (DCM) [58] improves the two-way Knowledge Transfer (KT) scheme by adding auxiliary classifiers to hidden layers of both teacher and student networks and building dense bidirectional KD between these classifiers. The work [59] also adopts mutual distillation method to encourage information exchange between the local branch and the global branch of a single network for discovering new categories. There are also some research about applying mutual learning to computer vision tasks, such as object detection [60], [61] and video-based sign language recognition [62]. Existing mutual distillation approaches have achieved success in high-level classification tasks. However, whether mutual distillation can help the middle-level dense matching tasks has not been fully explored.

As for optical flow estimation, DDFlow [40] and SelFlow [25] distill flow estimation from teacher to student with several occlusion hallucination methods, which has been shown to be effective when inferring on occluded regions. DistillFlow [43] further improves above two-stage data distillation by introducing multiple teacher models and confidence mechanism. As for a related medical image registration task, the work CRD [63] distills knowledge from a feature-shifted teacher model with high resolution input to a student model with low resolution input for more efficiency. Different from their works, our MDFlow transfers knowledge between teacher and student networks mutually, so as to decouple matching outliers in augmentation regularization and exploit recent advanced architecture as student for better final prediction, while maintaining real-time inference.

III. METHOD

In this section, we first illustrate the overall pipeline of MDFlow algorithm for unsupervised optical flow, and then provide notations and unsupervised initialization approach used in our framework. Further, the reliable forward knowledge distillation process is demonstrated, where a novel confidence matching selection mechanism plays a key role. At last, we present the reliable backward knowledge distillation process for improving the accuracy of final deployed teacher model.

A. Overview

As depicted in Figure 1, our proposed MDFlow algorithm trains a teacher model $T$ and a student model $S$ interactively for progressive improvement in a three-stage manner. In the first stage, our teacher network can be trained by any unsupervised approach for generating noisy pseudo labels, which will be used later. Specifically, we adopt PWC-Net [11] as teacher and use UFlow [23] loss functions for unsupervised initialization. In the second stage, we introduce a novel reliable matching selection mechanism, that can partly remove outliers in generated pseudo labels, which will then be used to train a student network with diverse data augmentation approaches, constituting our forward knowledge distillation process. Note that we can select a stronger student architecture, such as RAFT [13], for sufficient learning thanks to the decouple nature in mutual knowledge distillation. In the last stage, better pseudo labels predicted by the student will be adopted to supervise the learning of original teacher model enriched by diverse data augmentation. Moreover, we can improve the final teacher performance by adding a weak unsupervised learning objective, formulating it into a multi-target manner. More details are shown in Figure 2.
B. Unsupervised Initialization for Teacher Model

Given two consecutive frames \(I_1, I_2\) from unlabeled image sequences \(I\), unsupervised optical flow methods aim to train a flow network \(N(.)\) based on brightness constancy assumption and spatial smoothness prior. The first goal is usually realized as a photometric reconstruction loss:

\[
L_{ph}(N) = \sum_p \rho((I_1 - I^w) \odot (1 - O_f)) + \sum_p \rho((I_2 - I^w) \odot (1 - O_b))
\]

(1)

where \(\odot\) represents element-wise multiplication, \(O_f, O_b\) are occlusion masks inferred by forward-backward consistency check [20], [64], and \(\rho(x) = (|x| + \epsilon)^q\) is a robust function with \(\epsilon = 0.01, q = 0.4\) in all our experiments. Denoting \(F_f = N(I_1, I_2), F_b = N(I_2, I_1)\) as predicted forward and backward flow fields, and \(I^w = w(I_2, F_f), I^w = w(I_1, F_b)\) are flow warped reconstruction images. Following previous research [23], [40], we use the soft Hamming distance of census transformed [28] image patches to calculate reconstruction for its robust to illuminance variation.

Due to the well-known aperture problem, above solution can be ambiguous on textureless or repetitive pattern regions, so we further introduce a loss item to constrain the spatial smoothness of predicted flow fields, which is usually reweighted by local image gradients:

\[
L_{sm}(N) = \sum_{d \in x,y} \sum_p (|\nabla^k_d F_f| e^{-\lambda |\nabla^k_d I_1|} + |\nabla^k_d F_b| e^{-\lambda |\nabla^k_d I_2|})
\]

(2)

where \(\lambda\) controls edge-aware weighting strength conditioned on reference image. \(\nabla^k_d \), \(k = 1, 2\) stands for the \(k\)-th order gradient operator. In our experiments, we set \(\lambda = 150\). For smoothness, following the experience in UFlow [23], we use first order operator on Flying Chairs and Sintel, while adopting second order operator on KITTI for their better results. The reason is that there are more parallel motion about image plane in Chairs and Sintel, while more vertical motion in regard to image plane in driving scenes of KITTI. Flow fields projected from the parallel and vertical motion about image plane satisfy the first and second order smoothness respectively.

Inspired by recent work of ARFlow [22] and UFlow [23], we also add a self-supervised loss for continual improvement. Denoting \(A\) as the data augmentor, which can jointly augment input images \(I_1, I_2\), pseudo labels \(F_f, F_b\) and valid masks \(M_f, M_b\) by applying spatial and color transformations. Indicating \(\tilde{I}_1, \tilde{I}_2, \tilde{F}_f, \tilde{F}_b\) and \(\tilde{M}_f, \tilde{M}_b\) as corresponding augmented images, pseudo labels and valid masks, the surrogate

---

Fig. 2. Detailed framework of reliable mutual knowledge distillation for unsupervised optical flow. Green line denotes the forward path and blue line denotes the backward path. Networks with the same color share weights in each stage. Reliable masks \(M_f, M_b\) generated in stage 2 are shown in the left bottom corner. Note that the inputs and outputs of the data augmentor \(A\) are detached from the gradient calculation diagram for stable learning.
supervision loss can be formulated as:

\[
L_{sup}(\mathcal{N}_1|\mathcal{N}_2,A) = \frac{\sum_p |\mathcal{N}_1(I_1,I_2) - F_b| \odot M_f}{\sum_p M_f} + \frac{\sum_p |\mathcal{N}_1(I_1,I_2) - F_b| \odot M_b}{\sum_p M_b}.
\]

(3)

Note that there are two models \(\mathcal{N}_1\) and \(\mathcal{N}_2\) in Eq. 3, where \(\mathcal{N}_1\) is optimized conditioned on the pseudo labels \(F_f, F_b\) predicted by \(\mathcal{N}_2\) and augmentor \(A\). Following previous experience, augmented training samples are detached from gradient calculation diagram for stable learning. We adopt above loss notation for clarity of following expression. In regard to the first stage of MDFlow, we employ the off-the-shelf UFlow [23] method to initialize the teacher \(T\) with following objective:

\[
L_{s1} = L_{ph}(T) + \lambda_1 L_{sm}(T) + \lambda_2 L_{sup}(T|T,A),
\]

(4)

where \(\lambda_1, \lambda_2\) are trade-off coefficients, while \(L_{sup}(T|T,A)\) means that the second forward prediction is supervised by the augmented first prediction of the same model \(T\).

C. Reliable Forward Knowledge Distillation

Due to shadow, exposure, textureless and repetitive patterns existed in natural images, optimal solution of photometric reconstruction will inevitably contain mismatching areas. On the other hand, state-of-the-art unsupervised optical flow approaches [22], [23] augment predicted flow fields for a second supervision to improve performance. Concretely, they select non-occluded regions as the valid mask when calculating above self-supervised loss. As shown in Figure 3, large error regions in predicted pseudo labels, for example, the background surrounded by moving persons, are not identified by occlusion map. Thus, the augmented surrogate supervision will hinder the network from learning true displacement. To alleviate this problem, we first propose a novel confidence matching selection mechanism to find relatively reliable prediction, and then separate \(L_{sup}\) in Eq. 4 with \(L_{ph}, L_{sm}\) into a supervised knowledge distillation process. Holding the opinion that better matched image patches should have lower residuals, we explore to set a threshold to keep only a portion of reliable pseudo labels during forward knowledge distillation. In this work, we employ the soft Hamming distance of census transformed image patches for confidence measurement, and also include a robust function \(\rho\), which does not change confidence ranking. Denote \(M_f\) as forward valid mask and \(\tau\) as a threshold, the selection approach can be formulated as:

\[
M_f(p) = \begin{cases} 
1, & \rho(I_1(p) - I_2^{\rho}(p)) \leq \tau, \\
0, & \rho(I_1(p) - I_2^{\rho}(p)) > \tau.
\end{cases}
\]

(5)

In order to set meaningful threshold for outlier removal, we analyse the variation curve of average endpoint error with removal rate on Sintel and KITTI training sets, which is controlled by \(\tau\). Similar to uncertainty estimation [65], the sparsification curves are depicted in Figure 4, while several typical removal rates with corresponding thresholds and errors are listed in Table I. As can be seen, average endpoint error of remaining pseudo labels decreases as we gradually remove more prediction from original non-occluded mask, until removal rate reaches about 90%. On the other hand, larger percentage of removal means less pseudo labels for training, that can damage performance. Therefore, setting reasonable removal rate for better trade-off plays a key role in forward distillation process. We leave this question in experiment part for detailed discussion. Guided by the above confidence matching selection mask \(M_f, M_b\), reliable forward knowledge
distillation can be formulated as:
\[ L_{s2} = L_{sup}(S|T,A). \] \hspace{1cm} (6)

D. Reliable Backward Knowledge Distillation

Due to the decouple characteristic of proposed mutual distillation framework, we can employ a stronger flow architecture as student \( S \) than the original teacher \( T \), so as to learn from reliable pseudo labels sufficiently. This process is different from traditional knowledge distillation [51], [66] where a lightweight student usually distills knowledge from a large teacher model. Therefore, our approach further includes a backward distillation stage, aiming to transfer better student knowledge back to the efficient teacher model. As a result, proposed MDFlow does not add extra computation cost and inference delay in real deployment, compared with other unsupervised approaches.

To achieve this goal, one may follow the supervision loss of Eq. 6 used in stage 2, by exchanging \( T \) and \( S \). However, due to the limited learning ability of efficient optical flow network, and diverse strong augmentation imposed on pseudo labels, training data for the efficient teacher \( T \) is blended with augmentation noise, which deviates from the distribution of real-world dynamic scenes. On the other hand, we can not remove the augmentor \( A \) during backward distillation, since data augmentation has been proved to be one crucial step for excellent performance [67], [68]. To deal with this problem, we establish a multi-target learning manner for reliable backward knowledge distillation, integrating specific domain knowledge on optical flow task. Specifically, unsupervised photometric and smoothness losses based on original scene structure are introduced as regularization. In short, reliable backward distillation of stage 3 can be written as:
\[ L_{s3} = L_{sup}(T|S,A) + \lambda_3 L_{ph}(T) + \lambda_4 L_{sm}(T), \] \hspace{1cm} (7)
where \( \lambda_3, \lambda_4 \) are weight coefficients controlling unsupervised regularization intensity, and non-occluded masks \( 1 - O_f, 1 - O_b \) of \( T \) are adopted as valid masks in \( L_{sup} \).

IV. EXPERIMENTS

In this section, we first describe implementation details and training strategy of proposed MDFlow algorithm. Then, we quantitatively and qualitatively compare our results with existing state-of-the-art approaches on standard benchmark datasets, including Flying Chairs [9], MPI Sintel [29] and KITTI 2015 [30]. Further, ablation study on proposed reliable matching selection mechanism and superior student architecture in forward distillation, as well as multi-target regularized learning in backward distillation are carried out. Finally, we show superior generalization ability of our method.

A. Implementation Detail

We employ the efficient PWC-Net [11] as teacher model, and the powerful RAFT [13] as student model, whose model size, running time and computation complexity are compared in Table II. It can be seen that even RAFT contains a little less parameters than PWC-Net, however, its running time and computation complexity is \( 4 \times \) larger. All our experiments are implemented in PyTorch and conducted on 2 NVIDIA Tesla V100 GPUs. We adopt the Adam [69] and AdamW [70] optimizers to train the PWC-Net and RAFT respectively. Batch size is set to 4 for all experiments. Following previous work [23], [26], [40], we first pre-train MDFlow on Flying Chairs, and then fine-tune above teacher and student networks on each target dataset. As for Sintel, we use the combination of Clean and Final training parts, and adopt the fine-tuned teacher model for online evaluation. In regard to KITTI, we train on the KITTI 2015 multi-view extension training set, while removing frame numbers 9-12 in each sequence. Results of the teacher on KITTI 2015 test set are uploaded to KITTI website for online comparison. It takes overall about 4 days to train the three-stage MDFlow first on Chairs and then on Sintel or KITTI datasets, which also relys on the model complexity, IO speed and code parallelism degree.

B. Training Strategy

On each dataset, we perform proposed three-stage training pipeline for mutual knowledge distillation, where the first and last stage both take 300k iterations to train the teacher, while the second stage takes 100k iterations to train the student. When training on Flying Chairs, learning rate is initially set to \( 1 \times 10^{-4} \) and decays half at 100k, 150k, 200k and 250k iterations in both stage 1 and stage 3. As for stage 2, it is initially set to \( 4 \times 10^{-4} \) and decays half at 20k, 40k, 60k and 80k iterations. When fine-tuning on Sintel and KITTI, we use the same training schedule as Chairs, however, learning rate is divided by 2 in all corresponding phases. Across stages, both the teacher and the student load latest updated checkpoint for initialization, and are random initialized from scratch before training on Chairs. The data augmentor \( A \) includes operations of random flipping, resizing, rotating, cropping and color jittering. According to hyperparameter search, weight

\begin{table}[h]
\centering
\caption{Valid Mask Selection Threshold and Endpoint Error on MPI Sintel and KITTI 2015 Training Datasets With Specific Removal Rates. Models Are Trained With Loss Function in Eq. 4 on Sintel and KITTI Training Datasets Respectively.}
\begin{tabular}{|c|c|c|c|}
\hline
Removal Rate & MPI Sintel & KITTI 2015 &
\hline
0\% & 4.66 & 1.21 & 4.66 & 1.53 \hline
10\% & 3.22 & 0.58 & 3.59 & 1.26 \hline
20\% & 2.82 & 0.46 & 3.26 & 1.12 \hline
30\% & 2.54 & 0.40 & 2.99 & 1.03 \hline
\hline
\end{tabular}
\end{table}

\begin{table}[h]
\centering
\caption{Comparison of Different Optical Flow Architectures. Running Time and Computation Complexity Are Measured on Sintel Resolution Images With One NVIDIA GTX 1080 Ti GPU.}
\begin{tabular}{|c|c|c|c|}
\hline
Model & Parameters (M) & Runtime (ms) & FLOPs (G) \\
\hline
RAFT [13] & 5.26 & 125 & 362.5 \hline
PWC-Net [11] & 8.75 & 34 & 90.8 \hline
FastFlowNet [38] & 1.37 & 11 & 12.2 \hline
\hline
\end{tabular}
\end{table}
TABLE III

QUANTITATIVE RESULTS ON FLYING CHAIRS, MPI SINTEL AND KITTI 2015 DATASETS. THE METRIC EPE IS THE AVERAGE ENDPOINT ERROR AND FL-ALL IS THE PERCENTAGE OF ERRONEOUS PIXELS OVER ALL PIXELS. RESULTS IN PARENTHESES INDICATES IT IS EVALUATED USING DATA IT IS TRAINED ON. THE UNAVAILABLE RESULTS ARE MARKED AS ‘−’. ‘C’, ‘T’, ‘S’ AND ‘K’ STAND FOR FLYING CHAIRS, FLYINGTHINGS3D, MPI SINTEL AND KITTI 2015 DATASETS RESPECTIVELY. SUPERSCRIPT ‘RAW’ MEANS THE RAW PART OF CORRESPONDING DATASETS. ‘K’ MEANS KITTI VISUAL ODOMETRY DATASET. (STEREO) DENOTES STEREO DATA IS USED DURING TRAINING. FOR EACH ITEM IN SUPERVISED SETTING (TOP PART), THE BEST RESULT IS IN BOLD. FOR EACH ITEM IN UNSUPERVISED SETTING (BOTTOM PART), THE BEST RESULT IS BOLDFACED, AND THE SECOND BEST IS UNDERLINED.

| Data | Method | C-test | S-train (EPE) | S-test (EPE) | K-15-train | K-15-test |
|------|--------|--------|---------------|--------------|------------|-----------|
|      |        | EPE    | Clean         | Final        | Clean      | Final     | EPE       | Fl-all |
| C + T | FlowNet2 [10] | - | 2.02 | 3.14 | 3.96 | 6.02 | 10.06 | - |
| S/K | FlowNet2-ft [10] | - | (1.45) | (2.01) | 4.16 | 5.74 | (2.3) | 10.41% |
| C + T | FastFlowNet [38] | - | 2.89 | 4.14 | - | - | 12.24 | - |
| S/K | FastFlowNet-ft [38] | - | (2.08) | (2.71) | 4.89 | 6.08 | (2.13) | 11.22% |
| C + T | PWC-Net [11] | 2.30 | 2.55 | 3.93 | - | - | 10.35 | - |
| S/K | PWC-Net-ft [11] | - | (1.70) | (2.21) | 3.86 | 5.13 | (2.16) | 9.60% |
| C + T | RAFT [13] | - | 1.43 | 2.71 | - | - | 5.04 | - |
| S/K | RAFT-ft [13] | - | (0.77) | (1.20) | 2.08 | 3.41 | (1.5) | 5.27% |
| SY/Kawaw | UnFlow-CSS [20] | - | - | 7.91 | - | 10.22 | 8.10 | - |
| C + S/K | OptAwareFlow [21] | 3.30 | (4.03) | (5.95) | 7.95 | 9.15 | 8.88 | 51.20% |
| C + S/K | MPCOccFlow [71] | - | (3.48) | (5.52) | 7.23 | 8.81 | 6.59 | 22.94% |
| C + S/K | EPEFlow [72] | - | (3.54) | (4.99) | 7.00 | 8.51 | 5.66 | 16.95% |
| C + S/K | DDFlow [40] | 2.97 | (2.92) | (3.98) | 6.18 | 7.40 | 5.72 | 14.29% |
| Kawaw/Kawaw | SelFlow [25] | - | (2.88) | (3.87) | 6.56 | 6.57 | 4.84 | 14.19% |
| C + S/K | STFlow [41] | 2.53 | (2.91) | (3.59) | 6.12 | 6.63 | 3.56 | 13.83% |
| Kawaw (Stereo) | UnOSS [47] | - | - | - | - | - | 5.58 | 18.00% |
| Kawaw (Stereo) | Flow2Stereo [48] | - | - | - | - | - | 3.54 | 11.10% |
| Kawaw/Kawaw | ARFlow [22] | - | 2.79 | 3.73 | 4.78 | 5.89 | 2.85 | 11.80% |
| C + S/K | SimFlow [26] | 2.69 | (2.86) | (3.57) | 5.92 | 6.92 | 5.19 | 13.38% |
| C + S/K | UPFlow [23] | 2.55 | (2.50) | (3.39) | 5.21 | 6.50 | 2.71 | 11.13% |
| S + Kawaw/Kawaw | DistillFlow [43] | - | (2.61) | (3.70) | 4.23 | 5.81 | 2.93 | 10.54% |
| C + S/K | OilFlow [44] | 2.53 | (2.44) | (3.35) | 4.26 | 5.71 | 2.57 | 9.81% |
| Kawaw | ASFlow [45] | - | (2.40) | (2.49) | 4.56 | 5.86 | 2.47 | 9.67% |
| Kawaw/Kawaw | CoT-AMFlow [42] | - | - | - | - | - | - | 10.34% |
| Kawaw + K(Stereo) | PLC [49] | - | - | - | - | - | 2.35 | 9.70% |
| Kawaw/Kawaw | UPFlow [24] | - | (2.33) | (2.67) | 4.68 | 5.32 | 2.45 | 9.38% |
| C + S/K | MDFlow-Fast (Ours) | 2.75 | (2.53) | (3.47) | 4.73 | 5.99 | 3.02 | 11.43% |
| C + S/K | MDFlow (Ours) | 2.48 | (2.17) | (3.14) | 4.16 | 5.46 | 2.45 | 8.91% |

coefficients of \( \lambda_1, \lambda_2, \lambda_3, \lambda_4 \) in Eq. 4 and Eq. 7 are respectively set to 1.0, 0.05, 0.1, 0.1 for Chairs and Sintel, while being 5.0, 0.05, 0.02, 0.1 for KITTI. Note that the smoothness coefficient \( \lambda_1 \) in KITTI is 5 × of that in Chairs and Sintel in the first stage, while the photometric coefficient \( \lambda_3 \) in KITTI is only 1/5 of that in Chairs and Sintel in the third stage. It is because that KITTI uses the second order smoothness, a larger coefficient can reach a reasonable regularization strength in the first stage. And in the third stage, the non-occluded region in KITTI benefits more from the distillation term rather than the noisy reconstruction term.

C. Comparison to State-of-the-Art Methods

1) Quantitative Results: As shown in Table III, we quantitatively compare MDFlow with existing supervised and unsupervised methods on three leading optical flow benchmarks with standard metrics, where our approach outperforms most pyramid-based unsupervised methods on almost all datasets. The current state-of-the-art SMURF [27] uses the unfair recurrent model during evaluation, which is not listed in Table III because of large computation and time delay. Moreover, proposed MDFlow framework is general and does not depend on particular flow architecture and unsupervised initialization. Due to the diverse training configuration used in previous work, we list their training data in the first column for reference.

On Flying Chairs, MDFlow reduces previous best EPE from 2.53 to 2.48 slightly, possibly because of the relative simple motion in this domain. As for more challenging Sintel and KITTI 2015 datasets, our method achieves obvious improvement over others. On Sintel training, we obtains the best EPE of 2.17 on the Clean pass, and only behaves a little worse than UPFlow [24] and ASFlow [45] on the Final pass. For the online test benchmarks, our method achieves EPE of 4.16 and 5.46 on Clean and Final respectively, even outperforming DistillFlow [43] (4.23 and 5.81), which is trained on Sintel raw dataset, containing the unfair test image sequences. The method CoT-AMFlow [42] behaves best on both Sintel Clean and Final test sets. We attribute the reason to be that it both employs a more complex adaptive modulation flow network and uses the unfair Sintel raw movie sequences.

In regard to KITTI 2015, we achieve the same second best EPE of 2.45 with UPFlow [24] on the training set, but without additional parameters and inference delay, which is 9.2%
improvement over comparable UFlow [23] (2.71). We further upload flow predictions of the efficient teacher on test images to the KITTI benchmark website and achieve 8.91% Fl-all error rate, outperforming the state-of-the-art pyramid-based unsupervised approach UPFlow [24], i.e., 8.91% vs 9.38%. It is worth noting that proposed MDFlow also behaves better than the state-of-the-art binocular SLAM method FLC [49], that shows the room for improvement in this research area.

To facilitate a large number of downstream video tasks, especially for embedded real-time applications, we replace the original PWC-Net [11] with a more efficient FastFlowNet [38] as the teacher model, and repeat the same experiments as before. Model size, running time and computation complexity of FastFlowNet are compared in Table II. As shown in Table III, MDFlow-Fast can approach the state-of-the-art unsupervised flow accuracy while only containing 1.37 M parameters and consuming much less computation and inference time, demonstrating the efficient flow architecture of FastFlowNet and the efficient learning framework of MDFlow.

2) Qualitative Results: Some qualitative results on MPI Sintel and KITTI 2015 test datasets are depicted in Figure 5 and Figure 6 for visual comparison respectively. It can be seen that flow fields of proposed MDFlow can keep better motion scene structure layout than other methods, especially on regions with large occlusion and complex texture.

D. Ablation Study

To verify the effectiveness of proposed contributions, we conduct thorough ablation experiments of reliable mutual knowledge distillation algorithm considering all training stages and different training settings, and our findings are reported in Table IV. All ablation experiments are trained on Sintel training or KITTI 2015 multi-view extension datasets with corresponding loss functions in their training stage. We first report results of unsupervised initialization (stage 1) for teacher model on Sintel training and KITTI 2015 training sets using Eq. 4 as the baseline, which is denoted as E1 in Table IV.
TABLE IV
ABLA TION STUDY ON MPI SINTEL AND KITTI 2015 DATASETS. → AND ← REPRESENT FORWARD AND BACKWARD DISTILLATION PROCESSES, AND NO ARROW MEANS USING A SINGLE MODEL WITHOUT DISTILLATION. ‘INIT’ STANDS FOR CHECKPOINTS THAT CORRESPONDING MODELS ARE INITIALIZED WITH. ‘C’ MEANS CHECKPOINTS PRE-TRAINED ON FLYING CHAIRS DATASET. RR STANDS FOR REMOVAL RATE IN FIGURE 4.

| Stage | ID | Mode | Init | Setting | S-train (EPE) | K-15-train |
|-------|----|------|------|---------|--------------|------------|
|       |    |      |      |         | Clean        | Final      |
| 1     | E1 | PWC  | C    | 300k Iterations | (2.39)       | 2.66       |
|       | E2 | PWC  | C    | 700k Iterations | (2.38)       | 2.64       |
| 2     | E3 | PWC→PWC | E1→C | RR = 0% | (2.43)       | 2.64       |
|       | E4 | PWC→PWC | E1→C | RR = 10% | (2.52)       | 2.54       |
|       | E5 | PWC→PWC | E1→C | RR = 20% | (2.36)       | 2.60       |
|       | E6 | PWC→PWC | E1→C | RR = 30% | (2.37)       | 2.67       |
|       | E7 | PWC→RAFT | E1 | RR = 10% | (2.33)       | 3.00       |
|       | E8 | PWC→RAFT | E1 | RR = 10% | (2.16)       | 2.31       |
| 3     | E9 | PWC→RAFT | E1→E8 | \(\mathcal{L}_{sup}\) | (2.31)       | 2.53       |
|       | E10 | PWC→RAFT | E1→E8 | \(\mathcal{L}_{sup} + \mathcal{L}_{ph} + \mathcal{L}_{sm}\) | (2.27)       | 2.48       |
| 4     | E12 | PWC→RAFT | E1→E8 | RR = 10% | (2.08)       | 2.20       |
| 5     | E13 | PWC→RAFT | E1→E12 | \(\mathcal{L}_{sup} + \mathcal{L}_{ph} + \mathcal{L}_{sm}\) | (2.16)       | 2.36       |

1) Reliable Matching Selection: As shown in the middle part of Table IV, we conduct an ablation to explore the effectiveness of proposed confidence matching selection mechanism. According to Figure 4, removal rate is set to four typical values as 0%, 10%, 20% and 30%, denoted by E3, E4, E5 and E6 in Table IV, where 0% means that original non-occluded map is adopted as valid mask during forward knowledge distillation. In this group, both the teacher and the student models are instantiated as PWC-Net. Compared with the baseline E1, forward distillation with \(RR = 0\%\) has almost no improvement on the student, due to the mismatched regions in pseudo labels. When we remove 10% prediction of the teacher according to proposed confidence ranking mechanism in Eq. 5, the student yields the lowest endpoint error on both Sintel and KITTI 2015 training sets. As more uncertain labels being removed, the EPE gets larger, while the Fl-all metric on KITTI 2015 behaves the best with \(RR = 20\%\). Therefore, we use \(RR = 10\%\) in MDFlow, that is equivalent to adopting \(\tau\) of 3.22 and 3.59 on Sintel and KITTI respectively in Eq. 5 according to Table I. Speak of here, one may ask whether the distillation procedure is necessary, since we can directly apply proposed confidence mask to \(\mathcal{L}_{sup}(T|T, A)\) when optimizing Eq. 4. To explore this, we further carry out an experiment named ‘PWC’ in stage 2 marked as E7 in Table IV. As can be seen, it can approach the result of distillation counterpart on Sintel, while the results behave much worse on KITTI, whose reason may be the instability when combining the sparse valid mask with unsupervised loss in challenging scenes.

2) Stronger Student Model: Thanks to the decouple nature of our mutual distillation framework, we can employ any optical flow architecture as student without worrying about introducing additional cost in real deployment. Inspired by recent success on supervised optical flow task, for the first time, we try to explore whether a stronger student model can achieve better results when training on pseudo labels with the same amount of noise. As expected, when we employ advanced RAFT [13] as the student, endpoint error is reduced by 6.9%, 4.0% and 9.1% on Sintel Clean, Final and KITTI 2015 respectively, that is shown in E8 of stage 2. Significantly improved results demonstrate the effectiveness of proposed reliable forward knowledge distillation process. Moreover, E9 is the corresponding backward distillation experiment to E4, which jointly constitute a complete mutual distillation procedure but do not employ a stronger student. Compare E9 and E10, we can conclude that a stronger student can improve the final performance of the teacher.

3) Multi-Target Backward Distillation: Superior accuracy in stage 2 is at the cost of inference delay, thus transferring the better student knowledge back to the efficient teacher model makes sense. One can directly regard student’s prediction as pseudo labels, and train the teacher in a supervised manner, like what has been done in stage 2. However, as E10 shown in Table IV, performance of final teacher model declines back distinctly compared with E8, whose reason may be the limited learning capability of efficient network and relatively difficult task due to strong data augmentation. To deal with this problem, we formulate this step as a multi-target learning pipeline by introducing an unsupervised objective for regularization. As M11 listed in Table IV, proposed approach gets better results than pure supervision counterpart, and can well maintain or even surpass the powerful student model.

TABLE V
SENSITIVITY ANALYSIS OF THE PHOTOMETRIC AND THE SMOOTHNESS OBJECTIVE FUNCTIONS IN THE BACKWARD DISTILLATION STAGE

| \(\mathcal{L}_{sup} + \alpha \lambda_3 \mathcal{L}_{ph} + \beta \lambda_4 \mathcal{L}_{sm}\) | S-train (EPE) | K-15-train |
|-----------------------------|--------------|------------|
| \(\alpha = 1.0, \beta = 1.0\) | (2.17)       | 2.45       |
| \(\alpha = 0.5, \beta = 1.0\) | (2.21)       | 2.46       |
| \(\alpha = 1.5, \beta = 1.0\) | (2.18)       | 2.48       |
| \(\alpha = 1.0, \beta = 0.5\) | (2.19)       | 2.49       |
| \(\alpha = 1.0, \beta = 1.5\) | (2.23)       | 2.43       |
Furthermore, we do sensitivity analysis of the photometric loss $\mathcal{L}_{ph}$ and the smoothness loss $\mathcal{L}_{sm}$ in backward distillation stage according to Eq. 7, where original $\lambda_3 \mathcal{L}_{ph}$ and $\lambda_4 \mathcal{L}_{sm}$ are multiplied by a scale factor of $\alpha$ and $\beta$ respectively. As shown in Table V, $\alpha$ and $\beta$ are alternately set to 0.5 and 1.5 to perform a perturbation for weighting hyperparameters $\lambda_3$ and $\lambda_4$. It can be seen that a relatively large coefficient for $\mathcal{L}_{ph}$ is more benefit on Sintel, while a relatively large coefficient for $\mathcal{L}_{sm}$ is more helpful on KITTI. The reason may be Sintel dataset includes more non-rigid motion, while brightness noise is more obvious in real world KITTI dataset.

In summary, in a reasonable range of values for $\lambda_3$ and $\lambda_4$, mutli-target backward distillation behaves better than the pure supervision counterpart, demonstrating the robustness of proposed knowledge distillation approach.

4) Necessity of Multi-Stage Distillation: To demonstrate the necessity of multi-stage distillation, we conduct an experiment named E2 by enlarging training iterations of stage 1 to be equal to proposed three-stage learning procedure, where stage 1 is optimized with total 700k iterations for fair comparison. Compare E2 and E1 in Table IV, flow accuracy on diverse datasets has almost no improvement. We attribute the reason to be that error of pseudo label in augmentation regularization term has impeded the optimization of Eq. 4 to trap into local minimum. Therefore, multi-stage distillation is necessary.

5) One More Mutual Knowledge Distillation: Since proposed approach can bring progressive improvement between the teacher and the student, we carry out an experiment to perform one more mutual distillation between these two networks, whose forward and backward distillation stages are denoted as E12 and E13 in Table IV, respectively. As is expected, compare E12 with E8, and compare E13 with E11, we can see that both the student and the teacher have achieved better flow accuracy against their previous optimization stage. However, the reduction of average endpoint error on both Sintel and KITTI are almost less than 0.1, while it requires double training iterations. Therefore, it shows that the three-stage MDFlow has already achieved relatively good results and been saturated.

In short, our proposed MDFlow (E11) improves the baseline (E1) results by 9.2%, 6.0% and 16.0% on Sintel Clean, Sintel Final and KITTI 2015 training sets respectively, that is consistent with the improvement over the comparable state-of-the-art UFlow [23] on multiple test benchmarks.

E. Cross Domain Generalization

Finally, we conduct cross domain experiments to compare with recent top-performing unsupervised methods. As illustrated in Table VI, our method consistently performs best when training on one dataset and testing on another one, which demonstrates the state-of-the-art generalization ability of MDFlow among pyramid-based approaches. We attribute its excellent generalization performance to the efficient utilization of data augmentation on reliable pseudo labels during the decoupled mutual knowledge distillation process. Figure 7 presents qualitative results of the efficient teacher network on the cross domain DAVIS dataset [73]. It can be seen that proposed unsupervised learning approaches can estimate relatively good optical flow with clear motion boundary in diverse dynamic scenes.

| Method       | C-test | S-train (EPE) | K-15-train |
|--------------|--------|---------------|-----------|
|              | EPE    | Clean | Final | EPE | F-1-all |
| Chairs       |        |       |       |     |       |
| DDFlow [40]  | 2.97   | 4.83  | 4.85  | 17.26 |         |
| SimFlow [26] | 2.69   | 3.66  | 4.67  | 16.99 |         |
| UFlow [23]   | 2.55   | 3.43  | 4.17  | 11.27 | 30.31%  |
| MDFlow (Ours)| 2.48   | 2.89  | 4.00  | 9.60  | 25.87%  |
| Sintel       |        |       |       |     |       |
| DDFlow [40]  | 3.46   | (2.92) | (3.98) | 12.69 |         |
| ARFlow [22]  | 3.50   | 2.79  | 3.73  | 9.04  |         |
| SimFlow [26] | 3.01   | (2.86) | (3.87) | 12.75 |         |
| UFlow [23]   | 3.25   | (2.50) | (3.39) | 9.40  | 20.02%  |
| MDFlow (Ours)| 2.79   | (2.17) | (3.14) | 5.92  | 16.00%  |
| KITTI        |        |       |       |     |       |
| DDFlow [40]  | 6.35   | 6.20  | 7.08  | 5.72  |         |
| SimFlow [26] | 4.32   | 5.49  | 7.24  | 5.19  |         |
| UFlow [23]   | 5.05   | 5.58  | 6.31  | 2.71  | 9.05%   |
| MDFlow (Ours)| 4.14   | 4.04  | 5.34  | 2.45  | 8.10%   |
multi-target backward distillation procedure is built to transfer better student knowledge back to the efficient teacher without sacrificing generalization ability. Experiments on Flying Chairs, MPI Sintel and KITTI 2015 datasets show that our framework achieves state-of-the-art real-time accuracy and generalization performance. In the future, we plan to explore proposed reliable mutual knowledge distillation approach on other unsupervised matching tasks, such as Structure from Motion (SM) and scene flow estimation.

REFERENCES

[1] C. Wang, X. Chen, S. Min, J. Wang, and Z.-J. Zha, “Structure-guided deep video inpainting,” IEEE Trans. Circuits Syst. Video Technol., vol. 31, no. 8, pp. 2953–2965, Aug. 2021.

[2] L. Kong et al., “IFRNet: Intermediate feature refine network for efficient frame interpolation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 1969–1978.

[3] J. Liu, L. Kong, and J. Yang, “ATCA: An arc trajectory based model with curvature attention for video frame interpolation,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Aug. 2022, pp. 1–5.

[4] J. Dong and H. Liu, “Video stabilization for strict real-time applications,” IEEE Trans. Circuits Syst. Video Technol., vol. 27, no. 4, pp. 716–724, Apr. 2017.

[5] B. K. P. Horn and B. G. Schunck, “Determining optical flow,” Artif. Intell., vol. 17, nos. 1–3, pp. 185–203, Aug. 1981.

[6] T. Brox, C. Bregler, and J. Malik, “Large displacement optical flow,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2009, pp. 41–48.

[7] D. Sun, S. Roth, and M. J. Black, “Secrets of optical flow estimation and their principles,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Jun. 2010, pp. 2432–2439.

[8] C. Bailer, B. Taetz, and D. Stricker, “Flow fields: Dense correspondence fields for highly accurate large displacement optical flow estimation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 8, pp. 1879–1892, Aug. 2019.

[9] A. Dosovitskiy et al., “FlowNet: Learning optical flow with convolutional networks,” in Proc. IEEE Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 2758–2766.

[10] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, “FlowNet 2.0: Evolution of optical flow estimation with deep networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 2462–2470.

[11] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, “PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8934–8943.

[12] J. Hur and S. Roth, “Iterative residual refinement for joint optical flow and scene flow estimation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 5754–5763.

[13] Z. Teed and J. Deng, “RAFT: Recurrent all-pairs field transforms for optical flow,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2020, pp. 402–419.

[14] J. Wang, Y. Zhong, Y. Dai, K. Zhang, P. Ji, and H. Li, “Displacement-invariant matching cost learning for accurate optical flow estimation,” in Proc. Adv. Neural Inf. Process. Syst., 2020, pp. 15220–15231.

[15] N. Mayer et al., “Learning optical flow with a bidirectional census loss,” in Proc. Computer Vision—ECCV 2020, Cham, Switzerland: Springer, 2020.

[16] L. Kong, X. Yang, and J. Yang, “OAS-Net: Occlusion aware sampling framework achieves state-of-the-art real-time accuracy and generalization performance. In the future, we plan to explore proposed reliable mutual knowledge distillation approach on other unsupervised matching tasks, such as Structure from Motion (SM) and scene flow estimation.
