

# SciFlow: Empowering Lightweight Optical Flow Models with Self-Cleaning Iterations

Jamie Menjay Lin<sup>1</sup> Jisoo Jeong<sup>2</sup> Hong Cai<sup>2</sup> Risheek Garrepalli<sup>2</sup> Kai Wang<sup>1</sup> Fatih Porikli<sup>2</sup>

<sup>1</sup>Qualcomm Technologies, Inc. <sup>2</sup>Qualcomm AI Research<sup>†</sup>

{jmlin, jisojeon, hongcai, rgarrepa, kwang, fporikli}@qti.qualcomm.com

## Abstract

Optical flow estimation is crucial to a variety of vision tasks. Despite substantial recent advancements, achieving real-time on-device optical flow estimation remains a complex challenge. First, an optical flow model must be sufficiently lightweight to meet computation and memory constraints to ensure real-time performance on devices. Second, the necessity for real-time on-device operation imposes constraints that weaken the model’s capacity to adequately handle ambiguities in flow estimation, thereby intensifying the difficulty of preserving flow accuracy.

This paper introduces two synergistic techniques, Self-Cleaning Iteration (SCI) and Regression Focal Loss (RFL), designed to enhance the capabilities of optical flow models, with a focus on addressing optical flow regression ambiguities. These techniques prove particularly effective in mitigating error propagation, a prevalent issue in optical flow models that employ iterative refinement. Notably, these techniques add negligible to zero overhead in model parameters and inference latency, thereby preserving real-time on-device efficiency.

The effectiveness of our proposed SCI and RFL techniques, collectively referred to as SciFlow for brevity, is demonstrated across two distinct lightweight optical flow model architectures in our experiments. Remarkably, SciFlow enables substantial reduction in error metrics (EPE and Fl-all) over the baseline models by up to 6.3% and 10.5% for in-domain scenarios and by up to 6.2% and 13.5% for cross-domain scenarios on the Sintel and KITTI 2015 datasets, respectively.

## 1. Introduction

Optical flow is a fundamental task that represents pixel-level correspondence between two consecutive video frames. Since optical flow provides pixel-level movement, it is widely used for a variety of video perception tasks, e.g.,

<sup>†</sup> Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

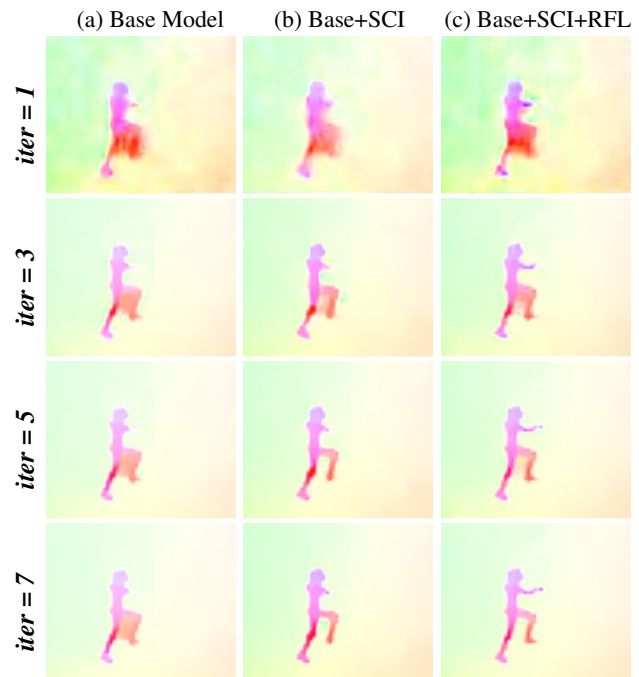


Figure 1. A zoomed-in demonstration of “Self-Cleaning Iterations (SCI)” effect against error propagation, a prevalent issue in iterative refinement for optical flow models. (a) The baseline model (RAFT-Small [39] as one choice of model architecture) suffers from error propagation over iterations, especially near the arm and legs. (b) When the SCI technique is applied to the baseline model, it demonstrates a “self cleaning” effect over iterations. This is achieved at negligible additional overhead in computation and in model size. (c) When both the SCI and RFL techniques are applied to the baseline model, the “self cleaning” effect becomes even more visible, particularly around the arm and feet. On top of “Base+SCI”, this RFL technique concerns only the loss function in training so it adds no additional overhead for inference.

action recognition [4, 26], object tracking [22, 50], video compression [28, 42], video frame interpolation [19, 25].

Thanks to the recent advances in deep learning, optical flow estimation models have become significantly more accurate by leveraging neural networks [13, 16, 39]. While

some earlier works lack a principled way to design neural networks for capturing pixel correspondences, more recently, RAFT [39] have proposed an optimization-inspired architecture that sets a new baseline for the method and the model architecture. More specifically, it constructs a global and re-usable cost volume with pairwise correlations between extracted image features of the two frames and then uses a recurrent neural module to iteratively refine the optical flow. This essentially mimics the optimization steps used conventional computer vision algorithms for solving correspondences, and has been shown to greatly improve accuracy and generalizability. As a result, most subsequent and current state-of-the-art solutions follow a similar model design strategy.

During the iterative estimation process, the predicted optical flow is prone to errors especially in the earlier stages or under high ambiguity. While the iterations can rectify many errors, other errors in some cases could persist and even be further propagated into later iterations, impacting final model accuracy. Figure 1 provides an example of this. In the top row for the first iteration, the initial flow estimates tend to be not accurate, with apparent errors near the person’s arm and legs as well as in the background. Through the iterations, the estimation on the background pixels improves, but other errors near the legs persist and the estimation on the arm becomes even less accurate. There is generally a lack of an effective solution to handle such issue of error propagation for many optical flow methods.

In this paper, we propose a novel and effective approach, Self-Cleaning Iterations (SCI), to address the issue of error propagation that is often observed during the iterative refinement process of optical flow models. We enable the network to “self-assess” the likely correctness of flow estimates during the iterative refinement. More specifically, in each iteration, we compare the feature maps of the two frames using the current estimated optical flow and warping. The pixel-wise differences provide an indication for consistency of the optical flow, which are converted into a quality range between 0 and 1. The resulting dense quality measure is consumed by the model as an additional feature channel to guide the network to “self-correct” inconsistencies in next iterations.

In addition, during training, we introduce a new loss, namely, Regression Focal Loss (RFL), to better leverage the available ground truth to improve the network’s awareness for regions of potentially incorrect estimates. Existing optical flow training schemes predominantly weight the pixel-wise loss equally, without taking into account the different prediction accuracy on each pixel at a given iteration. In contrast, our RFL gives heavier weights to regions of high residual regression errors, encouraging the network to focus its learning more on regions where it faces higher ambiguities to find feature correspondences.

Our proposed techniques add negligible or zero computation overhead at inference time, which is particularly critical for lightweight optical flow models intended for real-time on-device targets, such as mobile phones and AR/VR devices. Specifically, SCI only requires the network to process an additional channel of the quality map and RFL only affects loss computation during training. In contrast, many of the latest state-of-the-art methods require more complex computations for accuracy improvement, including heavier models or transformer architecture. Despite being parsimonious on computation usage, our proposed approach effectively improves multiple baseline architectures, achieving the best accuracy when comparing to other existing lightweight optical flow models.

Our main contributions are summarized as follows:

- We propose a novel technique, Self-Cleaning Iteration (SCI), which enables the model to “self-assess” flow quality in current iteration, and to “self-clean” flow estimates in subsequent iterations for optical flow models. This helps resolve ambiguities in the estimation and mitigates error propagation during the iterative refinement. It is noteworthy that SCI incurs minimal computational overhead during model inference.
- In addition, we propose a Regression Focal Loss (RFL), which guides the model to focus more on regions of high residual regression errors, thus encouraging the model to learn better to improve for those challenging scenarios where feature correspondences are harder.
- We further combine both techniques, SCI and RFL and verify our proposal on two distinct optical flow baseline architectures. Our experiments demonstrate that our SCI and RFL jointly serve as an effective unified solution to handle ambiguities in both in-domain and cross-domain scenarios. Remarkably, our solution also leads to state-of-the-art accuracy results compared with existing lightweight optical flow models.

## 2. Related Work

### 2.1. Optical Flow Models with Iterative Refinement

RAFT [39] introduced a new optical flow model design and has since become the baseline architecture for many later advancements. It builds a cross-level global correlation volumes and iteratively refines the prediction using convolutional gated recurrent units (ConvGRU) [6]. GMA [20], Flowformer [13], and FlowFormer++ [35] among others keep improving model accuracy while keeping this ConvGRU baseline design.

Despite the success of ConvGRU being an effective neural ODE optimizer [5, 8, 39], the error propagation behavior due in part to irresolvable ambiguities in estimates has not been explicitly discussed. In this paper, we aim at mitigating this issue prevalent in ConvGRU-based models.

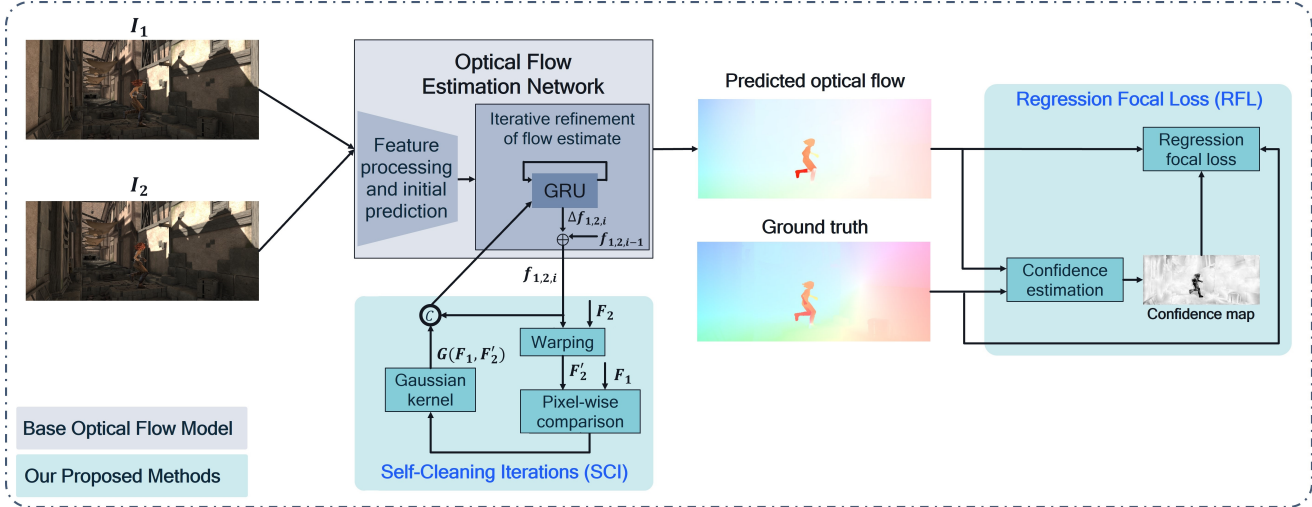


Figure 2. **Overview of our proposed approach.** **Self-Cleaning Iterations (SCI)** enables the network to “self-assess” the flow prediction quality and then to “self-clean” the flow prediction itself over the standard practice of iterative refinement process in many optical flow models. **Regression Focal Loss (RFL)** derives a confidence map and guide the network to focus more on regions of high residual regression errors during the iterations.  $\odot$  stands for the concatenation operator.

## 2.2. Uncertainty-Aware Optical Flow Estimation

Several prior works have explored incorporating confidence or uncertainty estimates into their models [11, 18, 40, 46] among others. For instance, [11] modify the network’s output layer to predict variance at intermediate layers and use assumed density filtering to propagate uncertainty across the network. [40] (PDC-Net) take a probabilistic approach, employing a mixture distribution for prediction and a separate uncertainty decoder within their multi-stage architecture to decouple flow estimation from uncertainty estimation. While PDC-Net is the closest work to ours, we differ in two key aspects. First, we do not explicitly use a dedicated uncertainty decoder. Second, we leverage geometric consistency for confidence estimation and utilize a self-cleaning mechanism for our iterative refinement.

## 2.3. Standard Loss Function for Optical Flow

FlowNet [9] uses end-point error loss, mathematically the Euclidean distance, between the ground truth and predicted flow. PWC-Net [38] uses L1 and L2 losses. L2 loss is first applied in the initial stage of training, while L1 loss is applied in subsequent finetuning. Several recent optical flow models [13, 39] apply iterative refinement by summing weighted L1 losses over multiple iterations.

$$l_i = \|f_{gt} - f_i\|_1 \quad (1)$$

where  $f_{gt}$  and  $f_i$  are the optical flow ground truth and prediction, respectively, in iteration  $i$ . A scheme below for weighted combination is then used over multiple iterations.

$$\mathcal{L}_{total} = \sum_{i=1}^N \gamma^{N-i} \cdot l_i \quad (2)$$

where  $N$  stands for the iteration index and  $0 < \gamma < 1$  is a decay factor over iterations. The whole predicted flow map of each iteration is weighted accordingly before accumulation.

Several other works adopt complementary insights or regularization objectives based on semantic segmentation, object depths, multi-frame aggregation, temporal consistency, occlusion consistency, or transformation consistency [1, 3, 7, 17, 44, 45] to further enhance their model accuracy on top of the standard optical flow loss function.

## 2.4. Focal Loss Function for Dense Classification

Focal Loss [27] has been proven an effective technique to address the class imbalance issue in dense classification tasks, such as segmentation. It places higher emphasis on feature samples of less (or under) represented classes.

$$CE(p_t) = -\log(p_t) \quad (3)$$

$$FL(p_t) = -(1 - p_t)^\gamma \log(p_t) \quad (4)$$

where CE and FL represent the cross entropy and focal loss, respectively. And,  $p_t$  is the probability for a class. Focal loss is originally proposed for dense classification tasks and is designed to work with the cross entropy loss. The original form of focal loss is not directly applicable to optical flow estimation, a regression task without the definition of classes. Moreover, cross entropy is not used in the standard loss for optical flow estimation.

## 2.5. Lightweight Optical Flow Models

[10, 24, 48] among others are recent works on lightweight design with attention or cost volume construction in a coarse-to-fine paradigm. [10] is a lightweight version of

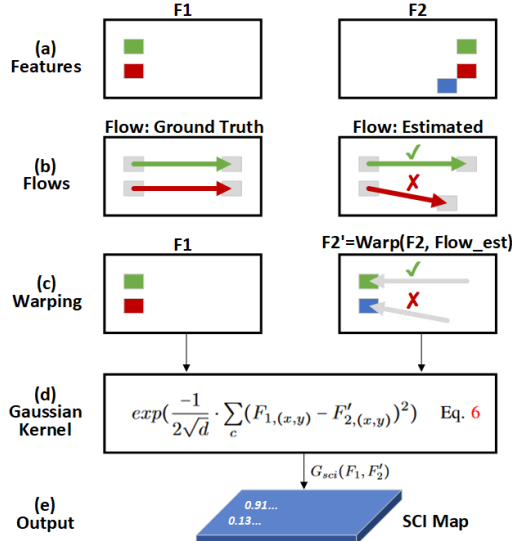


Figure 3. **Concept for SCI map creation.** (a) A pair of image features are taken as input. (b) The ground truth flows point to their matches on the left sub-figure, while the estimated flows point incorrectly for some features on the right sub-figure. (c)  $F2'$  is derived by warping  $F2$  by the estimated flows. (d)  $F1$  and  $F2'$  are taken by their tensor-wise differences for Gaussian Kernel (Eq. 6) evaluation for their affinity. (e) A dense SCI map output is derived.

[39], which adopts single level cost volume per iteration and adopts coarse-to-fine cost volumes with finest resolution being  $1/16$  and demonstrates real time performance on Snapdragon<sup>®</sup> 8 Gen 1 HTP. [48] first performs global matching at  $1/16$  resolution and then refines flow at  $1/8$  using lightweight CNN layers and demonstrates real time performance on Jetson Orion Nano. [24] also adopts coarse-to-fine and in addition uses dilated correlation layer for lighter cost volume. In addition there were additional previous works addressing light weight optical flow estimation like [14, 32].

### 3. Method

In this section, we discuss details of these two interrelated methods, Self-Cleaning Iteration (SCI) and Regression Focal Loss (RFL). The first technique, SCI, is applied for both training and inference by actively computing the similarity between the reference frame and the warped frame based on the flow estimate. RFL, the second technique as a loss function in a similar arithmetic formulation to that of SCI, is proposed to guide model learning by focusing more on regions of high residual regression errors. Figure 2 gives an overview for the system setup, including an optical flow estimation network and our proposed SCI and RFL methods.

#### 3.1. Self-Cleaning Iterations

In this section, we present the concept of Self-Cleaning Interactions (SCI) in the first half, and then the details of

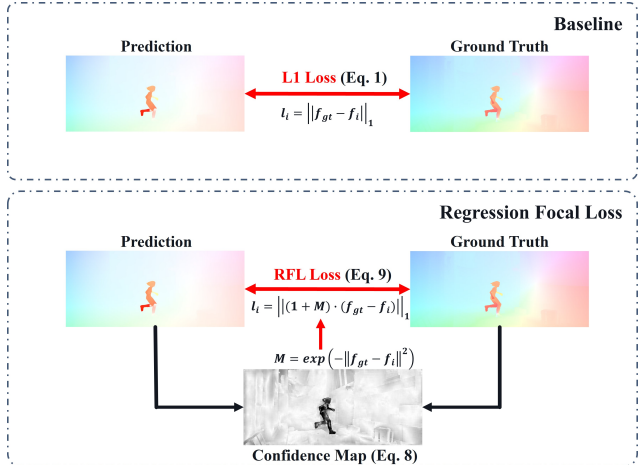


Figure 4. **Regression Focal Loss.** While equal-weight loss across all pixels is used for conventional optical flow model training (Eq. 1), Regression Focal Loss generates the confidence map using optical flow prediction and ground truth (Eq. 8) and leverages it to the optical flow loss (Eq. 9) so that model can focus on difficult areas in the dataset.

the method in the second half. Based on our observations in many iterative refinement-based models, we notice that errors made in early iterations of estimation could persist through subsequent iterations, affecting the quality of the dense flow estimates. To address this issue, SCI is designed to assess the quality of these estimates.

The core intuition behind SCI is the concept of ‘warping consistency’ of feature maps. More specifically, SCI measures the feature similarity between the warped target frame and the reference frame without any ground truth. This approach allows the model to self-assess the quality of flow estimates in an iteration, and then also allows the model to self-correct the errors in flow estimates over iterations.

Figure 3 illustrates the concept of SCI, beginning with the input of an image pair and culminating in the output of the SCI map, which represents the ‘self-assessed quality’ of flow estimates.

Next, in the second half of this section, we elaborate on how this SCI map is derived and applied.

Given input images  $I_1$  and  $I_2$ , we first encode these images into feature maps  $F_{1,0}$  and  $F_{2,0}$ . We then adopt an iterative process to estimate the dense flow field  $f_{1,2,i}$  in an iteration  $i$  for the dense pixelwise displacements between  $F_{1,i}$  and  $F_{2,i}$ .

$$F'_{2,i} = W(F_{2,i}, f_{1,2,i}) \quad (5)$$

where,  $W()$  is the standard warping operation that takes a dense input feature map  $F_{2,i}$  along with the estimated dense optical flow field  $f_{1,2,i}$  to produce a dense output feature map  $F'_{2,i}$  by reverting the pixel-wise displacements of  $F_{2,i}$  according to flow field  $f_{1,2,i}$  for each desired output coordinate point  $p_{x,y}$  of  $F_{1,i}$  and by interpolating closest neigh-

Table 1. **Cross-domain and in-domain optical flow estimation results on Sintel (train) and KITTI (train) datasets.** All models of cross-domain in this table are trained on FlyingChairs (C) and FlyingThings (T), following train protocol in [39]. All models of in-domain in this table are finetuned on Sintel (S) and KITTI (K) using pre-trained model (C+T), following [39] training protocol.

	Training Datasets	Method	# Params	Sintel (train)		KITTI (train)		Sintel (test)		KITTI (test)
				Clean	Final	EPE	Fl-all	Clean	Final	Fl-all
Cross-Domain	C+T	PWC-Net [37]	8.8 M	2.55	3.93	10.35	33.70	-	-	-
		LiteFlowNet2 [15]	6.4 M	2.24	3.78	8.97	25.90	-	-	-
		LiteFlowNet3 [14]	5.2 M	2.59	3.91	10.40	-	-	-	-
		FDFlowNet [23]	5.8 M	2.60	4.12	10.75	29.59	-	-	-
		FastFlowNet [24]	1.4 M	2.89	4.14	12.24	33.10	-	-	-
		MaskFlowNet-small [49]	-	2.33	3.72	-	23.58	-	-	-
		DICL [41]	-	1.94	3.77	8.70	23.60	-	-	-
		DIFT [10]	-	3.11	4.19	12.87	43.83	-	-	-
		MobileFlow <sup>1</sup>	1.5 M	1.79 (-0.0%)	3.47 (-0.0%)	8.33 (-0.0%)	22.06 (-0.0%)	-	-	-
		MobileFlow <sup>1</sup> +SCI+RFL (Ours)	1.5 M	<b>1.68 (-6.2%)</b>	<b>3.34 (-3.8%)</b>	<b>7.21 (-13.5%)</b>	<b>20.75 (-5.9%)</b>	-	-	-
In-Domain	C + T + S/K	PWC-Net [37]	8.8 M	2.02	2.08	2.16	9.80	4.39	5.04	9.60
		LiteFlowNet2 [14]	6.4 M	1.41	1.83	1.33	4.32	3.48	4.69	7.62
		LiteFlowNet3 [14]	5.2 M	1.43	1.90	1.39	4.35	2.99	4.45	7.34
		FDFlowNet [23]	5.8 M	1.80	1.93	1.56	6.36	3.71	5.11	9.38
		FastFlowNet [24]	1.4 M	2.08	2.71	2.13	8.21	4.89	6.08	11.22
		DDCNet (B1) [33]	3.0 M	1.96	2.25	2.57	15.56	6.19	6.91	38.23
		MobileFlow <sup>1</sup>	1.5 M	1.09 (-0.0%)	1.76 (-0.0%)	0.96 (-0.0%)	3.14 (-0.0%)	-	-	-
		MobileFlow <sup>1</sup> +SCI+RFL (Ours)	1.5 M	<b>1.03 (-5.5%)</b>	<b>1.65 (-6.3%)</b>	<b>0.92 (-4.2%)</b>	<b>2.81 (-10.5%)</b>	<b>2.62</b>	<b>3.80</b>	<b>5.82</b>

boring points for the queried coordinates in the source feature map  $F_{2,i}$ . Taking the pointwise differences between the original  $F_1$  and the warped  $F'_2$ , we then apply the sum of squared differences to a Gaussian kernel function with suitable normalization as follows.

$$G_{sci}(F_1, F'_2)|_{(x,y)} = e^{\frac{-1}{2\sqrt{a}} \cdot \sum_c (F_{1,(x,y)} - F'_{2,(x,y)})^2} \quad (6)$$

where, C stands for the set of elements in the channel dimension over the corresponding coordinates  $(x, y)$  of  $F_1$   $F'_2$ . The Gaussian kernel function comes with the following property for its value range.

$$0 \leq G_{sci}(F_1, F'_2)|_{(x,y)} \leq 1, \quad (7)$$

where, the maximum holds for  $\|F_{1,(x,y)} - F'_{2,(x,y)}\|_2^2 = 0$  and minimum holds for  $\|F_{1,(x,y)} - F'_{2,(x,y)}\|_2^2 = +\infty$ . For conciseness, we refer to this derived dense map  $G_{sci}(F_1, F'_2)$  as the *SCI map*. We then concatenate the SCI map with the estimated dense flow map along the channel dimension and feed them as the input to ConvGRU module for iterative refinement to derive the flow adjustment on top of the estimated flow.

### 3.2. Regression Focal Loss

In this subsection, we introduce Regression Focal Loss (RFL). Given the observation that the difficulty in predicting the pixel-wise flow can differ from one pixel to another depending on the contents at and around the pixels, we aim at helping the network focus its learning on regions that needs more improvement. Comparing with the focal loss [27] used in segmentation for handling class imbalance, our proposed RFL is intended for dense regression instead and may be considered as for "difficulty imbalance". To this end, we first derive a confidence map to facilitate the

pixel-wise weighting. We adopt the confidence map in LiteFlowNetv3 [14] as follows.

$$M_{\text{conf}}(x) = e^{-\|f_{\text{gt}}(x) - f_{\text{pred}}(x)\|^2} \quad (8)$$

Having the confidence map ready, we apply the map to  $l_i$  and replace Eq. 1 with the following.

$$l_i = \|(1 + \alpha \cdot (1 - M)^\beta) \cdot (f_{\text{gt}} - f_i)\|_1 \quad (9)$$

where  $\alpha$  and  $\beta$  are hyper parameters. The intuition of Eq. 9 is that we apply higher weighting to regions of low confidence and standard weighting to high confidence regions.

This RFL-based confidence weighting is derived by the final iteration of prediction and applied  $l_i$  of all iterations. We find this to be more effective as confidence derived in earlier iterations tends to be noisier, as we shall discuss more in our ablation study. Figure 4 compares between the baseline and the RFL approaches.

### 3.3. SciFlow: The Combination of SCI and RFL

Having individual definitions for SCI and RFL, we further discuss their relationship and our final proposal for combining them. Despite that SCI in Eq. 6 and RFL in Eq. 8 share similar arithmetic structures, their sources of the feature maps for the contrastive measures are quite different. During training, while the RFL relies on the ground truth in back propagation to focus on regions of larger residual regression errors, the SCI relies completely on the input images in the forward pass to derive the SCI map. During inference, the model continues its active computation for the SCI map to self-assess the flow estimates and to self-clean the flow ambiguities. SCI and RFL seem to be synergistic in learning to handle feature ambiguities, while they also complement each other in how their contrastive measures are used. In Section 4, we discuss more on empirical results for the combination of SCI and RFL.

Table 2. Ablation study for Self-Cleaning Iteration (SCI) and Regression Focal Loss (RFL). Following same protocol as described in RAFT [39], we train all model variants on top of two baseline architectures by two combinations of datasets specified in the table and evaluate them on Sintel (S) and KITTI (K) training datasets.

	Training Datasets	Architecture	SCI	RFL	Sintel		KITTI 15	
					clean (epe)	final (epe)	Fl-epe	Fl-all
Cross-Domain	C+T	RAFT-small [39]	✓	✓	2.21 (-0.0%)	3.35 (-0.0%)	7.51 (-0.0%)	26.90 (-0.0%)
					2.29 (+3.6%)	3.52 (+5.1%)	7.44 (-0.9%)	24.88 (-7.5%)
					2.17 (-1.8%)	<b>3.33 (-0.6%)</b>	7.58 (-0.9%)	25.45 (-5.4%)
		MobileFlow <sup>1</sup>	✓	✓	1.79 (-0.0%)	3.47 (-0.0%)	8.33 (-0.0%)	22.06 (-0.0%)
					<b>1.65 (-7.8%)</b>	<b>3.30 (-4.9%)</b>	7.22 (-13.3%)	<b>20.73 (-6.0%)</b>
					1.68 (-6.1%)	3.34 (-3.7%)	<b>7.21 (-13.4%)</b>	20.75 (-5.9%)
In-Domain	C+T + S/K	RAFT-small [39]	✓	✓	<b>1.42 (-0.0%)</b>	2.09 (-0.0%)	1.21 (-0.0%)	4.68 (-0.0%)
					1.46 (+2.8%)	2.06 (-1.4%)	<b>1.20 (-0.8%)</b>	4.74 (-1.3%)
					1.46 (+2.8%)	<b>2.04 (-2.4%)</b>	<b>1.20 (-0.8%)</b>	<b>4.50 (-3.8%)</b>
		MobileFlow <sup>1</sup>	✓	✓	1.09 (-0.0%)	1.76 (-0.0%)	0.96 (-0.0%)	3.14 (-0.0%)
					1.09 (-0.0%)	1.74 (-1.1%)	0.94 (-2.1%)	2.97 (-5.4%)
					<b>1.03 (-5.5%)</b>	<b>1.65 (-6.3%)</b>	<b>0.92 (-4.2%)</b>	<b>2.81 (-10.5%)</b>

Table 3. Ablation study of Regression Focal Loss (Eq. 9). We train RAFT-small+SCI models on FlyingChairs (C) and FlyingThings (T) and evaluate on Sintel (S) and KITTI (K) training datasets.

$l_i$ loss	Sintel		KITTI 15	
	clean (epe)	final(epe)	Fl-epe	Fl-all
a. $\ (f_{gt} - f_i)\ _1$ (Eq. 1)	2.17 (-0.0%)	<b>3.33 (-0.0%)</b>	7.58 (-0.0%)	25.45 (-0.0%)
b. $\ (\alpha \cdot (1 - M)^\beta) \cdot (f_{gt} - f_i)\ _1$	2.32 (+6.9%)	3.56 (+6.9%)	<b>7.09 (-6.5%)</b>	25.27 (-0.7%)
c. $\ (1 + \alpha \cdot (M)^\beta) \cdot (f_{gt} - f_i)\ _1$	4.52 (+108.3%)	5.92 (+77.8%)	10.07 (+32.8%)	50.53 (+95.5%)
d. $\ (1 + \alpha \cdot (1 - M)^\beta) \cdot (f_{gt} - f_i)\ _1$ (Eq. 9)	<b>2.11 (-2.8%)</b>	3.34 (+0.3%)	7.22 (-4.7%)	<b>24.62 (-3.3%)</b>

## 4. Experiments

### 4.1. Experimental Setup

**Datasets:** We follow commonly adopted training and evaluation protocols in the literature [13, 20, 39, 47]. We train our model on FlyingChairs (C) [9] and FlyingThings3D (T) [29] and evaluate on training dataset of Sintel (S) [2] and KITTI (K) [12, 30, 31]. Using C+T pre-trained model, we finetune Sintel and KITTI datasets and evaluate on Sintel and KITTI datasets.

**Network Architectures and Training:** We use two lightweight models with different architectures, RAFT-small [39] and MobileFlow<sup>1</sup>, as our baselines in the experiments. In particular, MobileFlow is our model creation for a lightweight baseline architecture. In order to build a feasible architecture that fits within the limited memory and compute capacity of a smartphone, we utilize memory-efficient cost volume techniques from [21, 43, 47]. We also adopt a MobileNetV2 [34] based backbone for feature extraction and a ConvGRU module for iterative refinement that is similar to [39]. For fair comparisons, we train both RAFT-small and MobileFlow baselines along with all their variants for SCI and RFL on top of the baselines using same train framework<sup>2</sup> and dataset protocol as described in RAFT [39] to report our experiment results. We follow the training parameters all the same as for RAFT, including number of iterations and the learning rate. For additional parameters

<sup>1</sup>MobileFlow is our created lightweight baseline architecture. Please see section 4.1 "Network Architectures and Training" for more details.

<sup>2</sup>RAFT: <https://github.com/princeton-vl/RAFT>

of regression focal loss, we set both  $\alpha$  and  $\beta$  in Eq. 9 to 1.

**Evaluation Metrics:** We evaluate our models by the End-Point Error (EPE) metric, which is the Euclidean distance between the predicted flow and the ground truth flow. We also use F1-all as defined for the KITTI dataset [31]. In both cases of error metrics, the lower is the better.

### 4.2. Experimental Results

#### 4.2.1 Cross-Domain Evaluation

The top half of Table 1 shows our cross-domain evaluation results, for which the models are trained on FlyingChairs and FlyingThings, and then are evaluated on Sintel and KITTI training datasets, respectively. Our solution, MobileFlow+SciFlow (namely, with both SCI and RFL), achieves significantly higher accuracy not only over the baseline MobileFlow but also over other compared lightweight optical flow methods.

#### 4.2.2 In-Domain Evaluation

The bottom half of Table 1 shows our in-domain evaluation results, where models are trained on FlyingChairs, FlyingThings3D, and Sintel (or KITTI) and are evaluated on Sintel (or KITTI) following the protocol as in RAFT [39]. Our proposed solution demonstrates significantly improved accuracy over the baseline and even over other state-of-the-art lightweight optical flow models. Please note that, unlike LiteFlowNet model series, in our experiments MobileFlow and its variant are trained only on KITTI 2015 dataset but not also on KITTI 2012 dataset.

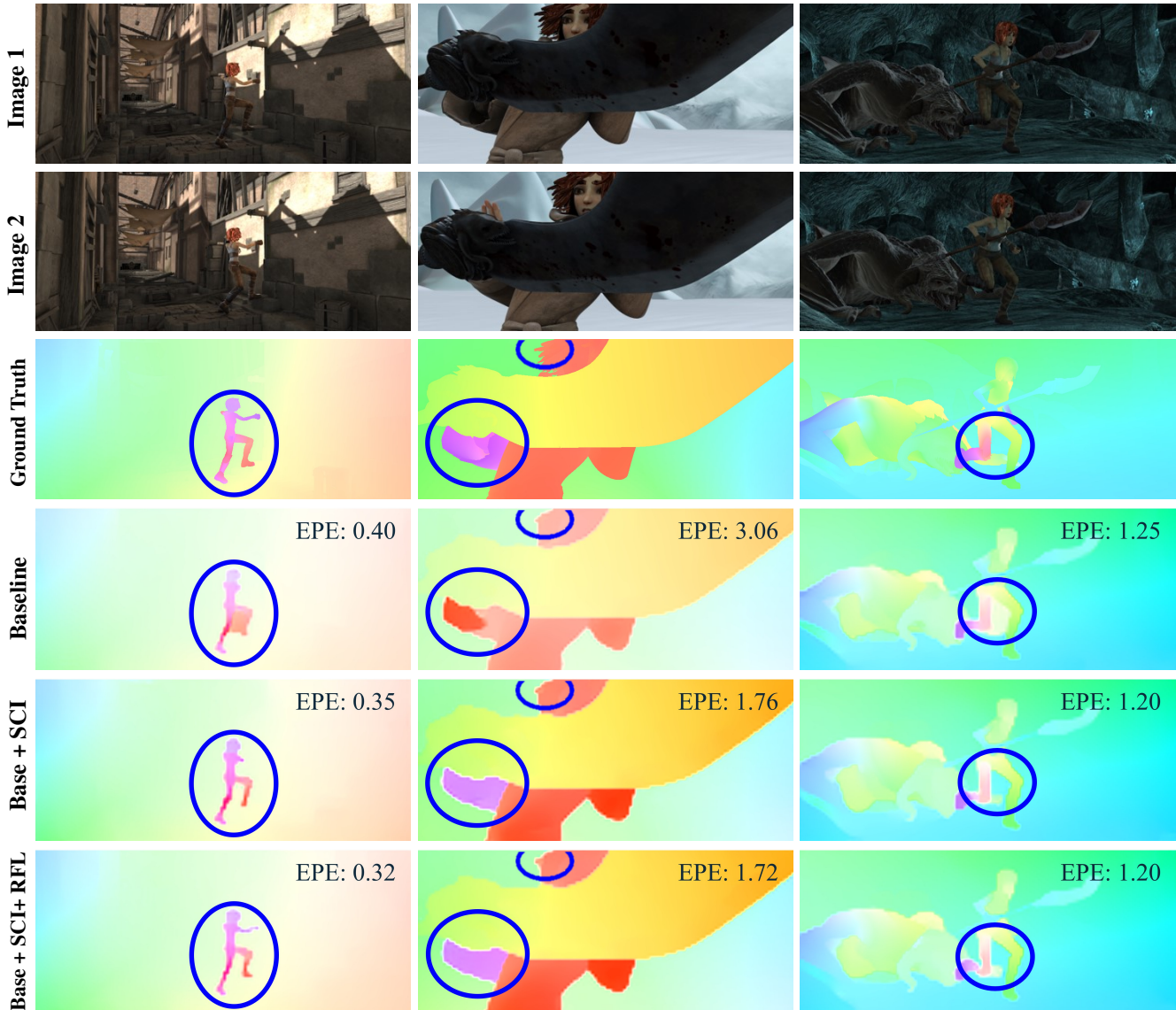


Figure 5. **Qualitative results on Sintel (train) dataset using RAFT-small architecture (trained with C+T).** First and second rows are input images. Third row is the ground truth. Fourth row is the output of RAFT-small. Fifth and sixth rows are the output of RAFT-small + SCI and RAFT-small + SCI + RFL, respectively.

### 4.2.3 Ablation Study

**SCI vs. RFL:** Table 2 summarizes our ablation study over choices and/or combinations of SCI and RFL. We applied SCI and RFL into three variants based on either the architecture of RAFT-small or the MobileFlow. Despite the fact that not all variants demonstrate improved accuracy in these experiments, the particular variants base+SCI+RFL in general demonstrate competitive accuracy over their respective baselines among minor run-to-run variations.

**Regression Focal Loss:** Table 3 summarizes our ablation study on RFL (Eq 9). Option "a" is the standard L1 loss without applying RFL. Option "b" produces inconsistent results over Sintel and KITTI, suggesting the impact

of removing the portion of the standard L1 loss. Option "c" uses the opposite focus on the regions of high confidence, which interestingly produces drastic degradation in accuracy, suggesting the wrong focus for the learning. Our proposed form in option "d" produces competitive results by combining both the L1 loss and the confidence-weighted focus on regions of higher residual errors.

**Final-Iteration Confidence Map vs. Per-Iteration Confidence Map:** Our proposed approach is to apply the single final confidence map to all iterations. When we apply instead per-iteration confidence map for each individual iteration, we see smaller gains than in the proposed approach. Table 4 lists the numbers.

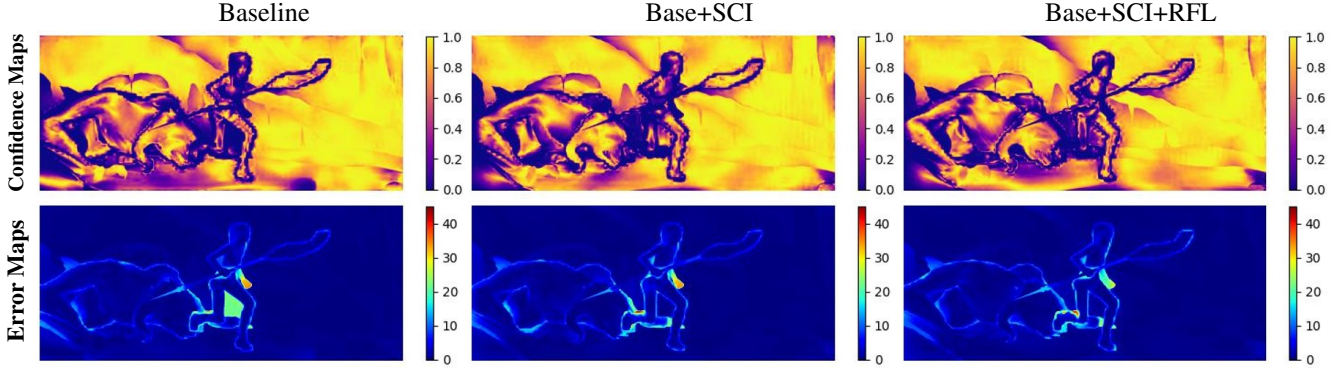


Figure 6. **Subtleties in RFL-based confidence maps and error maps.** **Top Row:** We report confidence maps derived with our RFL technique for the baseline and its variants. Moreover, SCI and RFL demonstrate their abilities to help resolve ambiguities in certain regions, as indicated by their higher confidence measures in these maps. **Bottom Row:** We report error maps for the predicted flow estimates against the ground truth for the baseline and its variants. Moreover, SCI and RFL demonstrate their ability to help resolve ambiguities in certain regions, as evidenced by their lower errors in these maps. In these examples, we use samples from the Sintel (train) dataset with RAFT-small architecture. The original input images can be found in the right column of Fig 5.

Table 4. Ablation study to compare between final-iteration confidence map and per-iteration confidence map. Here we train RAFT-large models on FlyingChairs (C) and FlyingThings (T) and evaluate on Sintel (S) and KITTI (K) training datasets.

Source of $M_{conf}$ (Eq. 8)	Sintel		KITTI 15	
	clean (epe)	final (epe)	Fl-epe	Fl-all
No confidence map	1.43 (-0.0%)	<b>2.71 (-0.0%)</b>	5.04 (-0.0%)	17.4 (-0.0%)
Per-iter confidence map	1.41 (-1.4%)	2.75 (+1.5%)	4.63 (-8.1%)	16.6 (-4.6%)
Final-iter confidence map	<b>1.38 (-3.5%)</b>	2.77 (+2.2%)	<b>4.58 (-9.1%)</b>	<b>16.2 (-6.9%)</b>

#### 4.2.4 Qualitative Results

Fig. 5 gives qualitative samples on Sintel dataset. The base+SCI variant demonstrates improved robustness over the baseline in occlusion areas. Base+SCI+FRL further shows slight improvements in several subtle visual details.

#### 4.2.5 RFL as A Confidence Measure

Figure 6 demonstrates an additional use of RFL as a confidence measure in inference.

#### 4.2.6 On-Device Evaluation

We report on-device evaluation of our models on Samsung S24 with a Snapdragon 8 Gen 3 processor and Qualcomm<sup>®</sup> Hexagon<sup>™</sup> Tensor Processor (HTP), which is an AI accelerator specialized for neural network workloads. We adopt the INT8 (W8A8) quantization based on AIMET<sup>3</sup> [36] toolkit and use the QNN-SDK<sup>4</sup> from Qualcomm<sup>®</sup> AI Stack.<sup>5</sup> Table 5 summarizes our evaluation on this target S24 device. Other than the RAFT-S and its variant that run out of memory, an expected behavior due to the all-pair cost volume space consumption for RAFT [39] architecture against limited on-target memory, the result shows

<sup>3</sup>AIMET is a product of Qualcomm Innovation Center, Inc.

<sup>4</sup><https://developer.qualcomm.com/software/qualcomm-ai-stack>

<sup>5</sup>Snapdragon and Qualcomm branded products are products of Qualcomm Technologies, Inc. and/or its subsidiaries.

Table 5. **On-device evaluation for SciFlow variants over baselines.** We report on-device performance of the baselines, RAFT-S and MobileFlow, and their variants. For fair comparisons, we ensure same on-device execution power mode and apply same number (6) of iterations to meet the real-time requirement for model variants below. "NA/OOM" indicates an out-of-memory error for the model (along with its needed memory for cost volume and activations) during inference. Subsection 4.2.6 has more details.

Model Architecture	#Params	Latency (ms)	Power (mW)
RAFT-S	1.0M	NA/OOM	NA/OOM
RAFT-S+SCI+RFL	1.0M	NA/OOM	NA/OOM
MobileFlow <sup>1</sup>	1.5M	29.02 (+0.00%)	392 (+0.00%)
MobileFlow <sup>1</sup> +SCI+RFL	1.5M	28.88 (-0.48%)	393 (+0.26%)

that our proposed SciFlow method incurs minimal additional overhead in latency and power. We observe slightly reduced latency MobileFlow+SCI+RFL compared to baseline. Though it might seem counter-intuitive, the compiler optimization may be the reason for such observation, an indication for minimal SciFlow latency overhead.

## 5. Conclusion

In this paper, we introduce two effective techniques for optical flow estimation. Specifically, we propose Self-Cleaning Iterations (SCI) to help resolve estimation ambiguities, mitigating the issue of error propagation during iterative refinement. Additionally, We propose Regression Focal Loss (RFL) to guide the model to focus on regions of high residual regression errors during training. Our experiments show that SciFlow, the combination of SCI and RFL, significantly improves accuracy of lightweight baseline models at negligible additional overhead for real-time on-device optical flow estimation. We believe our methods may benefit a wider range of model architectures and may be potentially extended to more vision use cases and tasks.



## References

- [1] Shubhankar Borse, Debasmit Das, Hyojin Park, Hong Cai, Risheek Garrepalli, and Fatih Porikli. Dejavu: Conditional regenerative learning to enhance dense prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19466–19477, 2023. [3](#)
- [2] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *Proceedings of the European Conference on Computer Vision*, pages 611–625. Springer, 2012. [6](#)
- [3] Hong Cai, Janarbek Matai, Shubhankar Borse, Yizhe Zhang, Amin Ansari, and Fatih Porikli. X-distill: Improving self-supervised monocular depth via cross-task distillation. In *British Machine Vision Conference*, 2021. [3](#)
- [4] Zixi Cai, Helmut Neher, Kanav Vats, David A Clausi, and John Zelek. Temporal hockey action recognition via pose and optical flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. [1](#)
- [5] Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *NeurIPS*, pages 6572–6583, 2018. [2](#)
- [6] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. [2](#)
- [7] Debasmit Das, Shubhankar Borse, Hyojin Park, Kambiz Azarian, Hong Cai, Risheek Garrepalli, and Fatih Porikli. Transadapt: A transformative framework for online test time adaptive semantic segmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. [3](#)
- [8] Edward De Brouwer, Jaak Simm, Adam Arany, and Yves Moreau. Gru-ode-bayes: Continuous modeling of sporadically-observed time series. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. [2](#)
- [9] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2758–2766, 2015. [3](#), [6](#)
- [10] Risheek Garrepalli, Jisoo Jeong, Rajeswaran C Ravindran, Jamie Menjay Lin, and Fatih Porikli. Dift: Dynamic iterative field transforms for memory efficient optical flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2219–2228, 2023. [3](#), [5](#)
- [11] Jochen Gast and Stefan Roth. Lightweight probabilistic deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3369–3378, 2018. [3](#)
- [12] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. [6](#)
- [13] Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer: A transformer architecture for optical flow. In *Proceedings of the European Conference on Computer Vision*, 2022. [1](#), [2](#), [3](#), [6](#)
- [14] Tak-Wai Hui and Chen Change Loy. Liteflownet3: Resolving correspondence ambiguity for more accurate optical flow estimation. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 169–184. Springer, 2020. [4](#), [5](#)
- [15] Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. A lightweight optical flow cnn—revisiting data fidelity and regularization. *IEEE transactions on pattern analysis and machine intelligence*, 43(8):2555–2569, 2020. [5](#)
- [16] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2462–2470, 2017. [1](#)
- [17] Jisoo Jeong, Jamie Menjay Lin, Fatih Porikli, and Nojun Kwak. Imposing consistency for optical flow estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18–24, 2022*, pages 3171–3181. IEEE, 2022. [3](#)
- [18] Jisoo Jeong, Hong Cai, Risheek Garrepalli, and Fatih Porikli. Distractflow: Improving optical flow estimation via realistic distractions and pseudo-labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13691–13700, 2023. [3](#)
- [19] Jisoo Jeong, Hong Cai, Risheek Garrepalli, Jamie Menjay Lin, Munawar Hayat, and Fatih Porikli. Ocai: Improving optical flow estimation by occlusion and consistency aware interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. [1](#)
- [20] Shihao Jiang, Dylan Campbell, Yao Lu, Hongdong Li, and Richard Hartley. Learning to estimate hidden motions with global motion aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9772–9781, 2021. [2](#), [6](#)
- [21] Shihao Jiang, Yao Lu, Hongdong Li, and Richard I. Hartley. Learning optical flow from a few matches. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16587–16595, 2021. [6](#)
- [22] Kiran Kale, Sushant Pawar, and Pravin Dhulekar. Moving object tracking using optical flow and motion vector estimation. In *2015 4th international conference on reliability, infocom technologies and optimization (ICRITO)(trends and future directions)*, pages 1–6. IEEE, 2015. [1](#)
- [23] Lingtong Kong and Jie Yang. Fdflownet: Fast optical flow estimation using a deep lightweight network. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 1501–1505. IEEE, 2020. [5](#)
- [24] Lingtong Kong, Chunhua Shen, and Jie Yang. Fastflownet: A lightweight network for fast optical flow estimation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10310–10316. IEEE, 2021. [3](#), [4](#), [5](#)

- [25] Lingtong Kong, Boyuan Jiang, Donghao Luo, Wenqing Chu, Xiaoming Huang, Ying Tai, Chengjie Wang, and Jie Yang. Ifrnet: Intermediate feature refine network for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1969–1978, 2022. 1
- [26] Myunggi Lee, Seungeui Lee, Sungjoon Son, Gyutae Park, and Nojun Kwak. Motion feature network: Fixed motion filter for action recognition. In *Proceedings of the European Conference on Computer Vision*, pages 387–403, 2018. 1
- [27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3, 5
- [28] Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Chunlei Cai, and Zhiyong Gao. Dvc: An end-to-end deep video compression framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11006–11015, 2019. 1
- [29] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4040–4048, 2016. 6
- [30] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3061–3070, 2015. 6
- [31] Moritz Menze, Christian Heipke, and Andreas Geiger. Joint 3d estimation of vehicles and scene flow. In *ISPRS Workshop on Image Sequence Analysis (ISA)*, 2015. 6
- [32] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4161–4170, 2017. 4
- [33] Ali Salehi and Madhusudhanan Balasubramanian. Ddcnet: Deep dilated convolutional neural network for dense prediction. *Neurocomputing*, 523:116–129, 2023. 5
- [34] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks, 2018. cite arxiv:1801.04381. 6
- [35] Xiaoyu Shi, Zhaoyang Huang, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer++: Masked cost volume autoencoding for pretraining optical flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1599–1610, 2023. 2
- [36] Sangeetha Siddegowda, Marios Fournarakis, Markus Nagel, Tijmen Blankevoort, Chirag Patel, and Abhijit Khobare. Neural network quantization with ai model efficiency toolkit (aimet), 2022. 8
- [37] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *CVPR*, 2018. 5
- [38] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2018. 3
- [39] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision*, pages 402–419. Springer, 2020. 1, 2, 3, 4, 5, 6, 8
- [40] Prune Truong, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning accurate dense correspondences and when to trust them. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5714–5724, 2021. 3
- [41] Jianyuan Wang, Yiran Zhong, Yuchao Dai, Kaihao Zhang, Pan Ji, and Hongdong Li. Displacement-invariant matching cost learning for accurate optical flow estimation. *Advances in Neural Information Processing Systems*, 33, 2020. 5
- [42] Chao-Yuan Wu, Nayan Singhal, and Philipp Krahenbuhl. Video compression through image interpolation. In *Proceedings of the European Conference on Computer Vision*, pages 416–431, 2018. 1
- [43] Haofei Xu, Jiaolong Yang, Jianfei Cai, Juyong Zhang, and Xin Tong. High-resolution optical flow from 1d attention and correlation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10498–10507, 2021. 6
- [44] Rajeev Yasarla, Hong Cai, Jisoo Jeong, Yunxiao Shi, Risheek Garrepalli, and Fatih Porikli. Mamo: Leveraging memory and attention for monocular video depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8754–8764, 2023. 3
- [45] Rajeev Yasarla, Manish Kumar Singh, Hong Cai, Yunxiao Shi, Jisoo Jeong, Yinhao Zhu, Shizhong Han, Risheek Garrepalli, and Fatih Porikli. Futuredepth: Learning to predict the future improves video depth estimation. *arXiv preprint arXiv:2403.12953*, 2024. 3
- [46] Zhichao Yin, Trevor Darrell, and Fisher Yu. Hierarchical discrete distribution decomposition for match density estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6044–6053, 2019. 3
- [47] Feihu Zhang, Oliver J Woodford, Victor Adrian Prisacariu, and Philip HS Torr. Separable flow: Learning motion cost volumes for optical flow estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10807–10817, 2021. 6
- [48] Zhiyong Zhang, Huaizu Jiang, and Hanumant Singh. Neuflow: Real-time, high-accuracy optical flow estimation on robots using edge devices, 2024. 3, 4
- [49] Shengyu Zhao, Yilun Sheng, Yue Dong, Eric I Chang, Yan Xu, et al. Maskflownet: Asymmetric feature matching with learnable occlusion mask. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6278–6287, 2020. 5
- [50] Huizhong Zhou, Benjamin Ummenhofer, and Thomas Brox. Deeptam: Deep tracking and mapping. In *Proceedings of the European Conference on Computer Vision*, pages 822–838, 2018. 1