Adapting Generic RGB-D Salient Object Detection for Specific Traffic Scenarios

Chenglizhao Chen, Mengke Song, Shanchen Pang, and Chong Peng

***Abstract*— Existing RGB-D salient object detection (SOD) models are primarily trained on general-purpose datasets, which may lead to domain shift issues when applied directly to new, specific scenes, such as stereo traffic datasets. Though "large-scale datasets (COME15K and ReDweb-S)" have been released, they only partially address the domain shift problem. From the perspective of data augmentation, this paper presents a novel solution, which follows a weakly-supervised way to adapt generic RGB-D SOD models for specific scenarios, with a focus on traffic scene imagery. Our key idea is to equip plain videos (specific scenarios, i.e., traffic scenes) with newly estimated saliency informative depth maps and pseudo-SOD GTs, enabling them to support the retraining of existing RGB-D SOD models for meeting the requirements of these specific scenes. To achieve this, we offer a fresh perspective on how depth information can be leveraged in the SOD task and introduce a new paradigm for extracting intrinsic information from optical flows derived from videos to refine RGB-D SOD models. Our method achieves a 1.2% improvement in F-measure on RGB-D datasets and a 27% enhancement on real-world street view datasets compared to baseline models. These results demonstrate the effectiveness of our approach in enhancing model adaptability for traffic scene imagery, even with limited target domain data. Codes, datasets, and results are available at https://github.com/MengkeSong/AGSS.**

***Index Terms*— RGB-D salient object detection, domain shift, deep learning, traffic image SOD.**

Received 14 August 2024; revised 10 October 2024, 9 December 2024, and 12 January 2025; accepted 26 March 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62172246 and 62276147, in part by Excellent Young Scientists Fund of Shandong Provincial Natural Science Foundation under Grant ZR2024YQ071, in part by the Fundamental Research Funds for the Central Universities under Grant 22CX06037A, and in part by Shandong Province Colleges and Universities Youth Innovation Technology Plan Innovation Team Project under Grant 2022KJ149. The Associate Editor for this article was Y. Yu. *(Corresponding author: Chong Peng.)*

Chenglizhao Chen is with Qingdao Institute of Software & College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China, also with Shandong Key Laboratory of Intelligent Oil & Gas Industrial Software, Qingdao 266580, China, and also with Jiangsu Key Laboratory of Image and Video Understanding for Social Safety, Nanjing 210094, China.

Mengke Song is with Qingdao Institute of Software and the College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China, and also with Shandong Key Laboratory of Intelligent Oil and Gas Industrial Software, Qingdao 266580, China.

Shanchen Pang is with Qingdao Institute of Software and the College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China, also with Shandong Key Laboratory of Intelligent Oil and Gas Industrial Software, Qingdao 266580, China, and also with the State Key Laboratory of Chemical Safety, Qingdao 266000, China.

Chong Peng is with the Faculty of Information Science and Engineering, Ocean University of China, Qingdao 266071, China (e-mail: pchong1991@163.com).

Digital Object Identifier 10.1109/TITS.2025.3555966

I. Introduction

IMAGE salient object detection (SOD) aims at detecting and segmenting objects which attract human attention most visually in a given scene, which is significant to downstream visual tasks such as, image retrieval [1], image segmentation [2] and driving activity [3], [4], [5], [6], [7]. Most existing works [8], [9], [10] have mainly focused on SOD for RGB images. With the rising popularity of depth (D) sensing equipment, RGB-D SOD has received intensive research attention recently. The newly available D can provide additional informative clues for potentially separating salient objects from non-salient surroundings. Such successful separation might never have been achieved if only RGB had been considered.

The rapid advancement of deep learning has facilitated the development of several RGB-D SOD models, such as those proposed in [12] and [13], which are trained on large, well-annotated datasets designed for general-purpose use. As illustrated in Fig. 1-A, these models generally perform well when tested on datasets similar to the ones they were trained on. However, a key challenge arises when such models are tested on unfamiliar RGB-D scenes. For example, a model trained on the NJUD-TR dataset [11], which predominantly contains natural scenes, performs poorly when applied to stereo traffic scenes, like those in the KITTI dataset [14] (shown on the right side of Fig. 1). This performance drop highlights a typical domain shift problem, where large differences between the source domain (natural scenes) and the target domain (stereo traffic scenes) hinder model generalization, as illustrated in Fig. 1-B.

The domain shift problem has been studied in various computer vision domains [15], [16], but it remains largely unexplored in the field of SOD, especially for multi-modal inputs like RGB-D images. Domain shift in SOD is particularly challenging because traffic scenarios, which often feature dynamic and cluttered environments, can drastically differ from the natural or indoor scenes seen in many existing training datasets. This gap in data characteristics makes it difficult for deep learning models to generalize effectively when tested in real-world traffic conditions.

One of the most intuitive approaches to address domain shift is to automatically generate large-scale, domain-specific training datasets. However, this is especially challenging for the RGB-D SOD task due to two main issues. First, obtaining RGB-D images is inherently difficult, as most publicly available images contain only RGB data without accompanying depth information. This problem is even more pronounced in

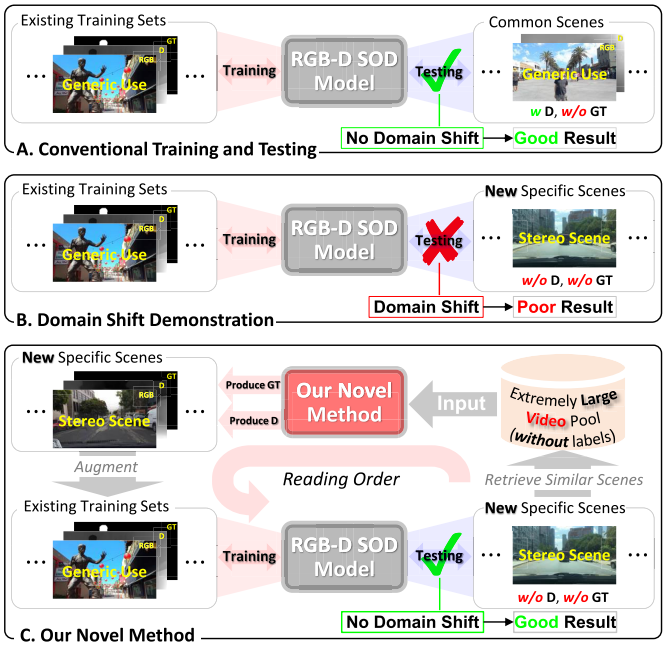

Fig. 1. Our approach highlights that an RGB-D SOD method trained on "common scenes" (e.g., NJUD-TR [11]) performs well on similar datasets (A) but struggles with specific scenes (B), such as stereo scenes in KITTI, due to domain shift. From a data augmentation perspective, our method (C) generates high-quality SOD ground truths and informative depth maps for video data, allowing us to use large amounts of plain video data from similar scenes to retrain state-of-the-art RGB-D SOD models, thereby mitigating the domain shift issue.

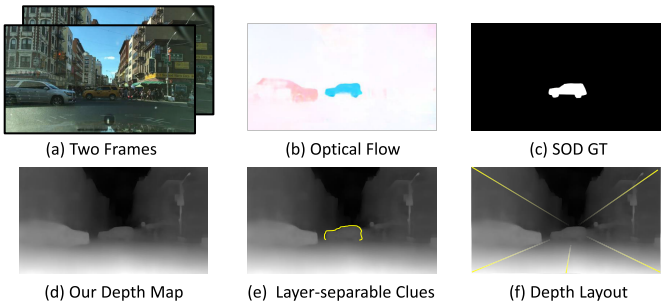

Fig. 2. NOT all information in a depth (D) map is useful for SOD. Thus, our goal is to generate **saliency informative D maps** *without* considering the goals targeted by the conventional D estimation task, *e.g.*, correct D layout, and rich details. The main reasons are two-fold. 1) Though it is still not sure if the D layout (f) can benefit the SOD task, we feel confident that a D map provides "layer-separable clues" (e) for the SOD network to separate salient objects from their non-salient surroundings. 2) Though optical flow maps (b) are physically irrelevant to real D maps, our newly-obtained **saliency informative D maps** (d) can well benefit the SOD task because moving objects tend to be salient.

traffic scenarios, where the availability of high-quality RGB-D datasets is limited. Second, SOD is a dense prediction task that requires pixel-wise saliency annotations, which are costly and time-consuming to generate. Existing auto-annotation techniques [17], [18] often produce pseudo-annotations of low quality, further complicating the problem.

To address the challenges mentioned, we propose a novel approach to automatically construct a high-quality trainable RGB-D dataset using "VIDEO" data alone. As illustrated in Fig. 1-C, we leverage the complementary spatiotemporal information present in videos to generate high-quality pseudo-GTs. In parallel, we use layer-separable clues (shown in Fig. 2-e) that are embedded in the video's temporal domain, specifically optical flow (Fig. 2-b), to enhance the RGB frames with depth information. In other words, we enrich each RGB video frame with a saliency informative depth map using the optical flow. Since objects with noticeable movement typically have higher saliency than static objects (Fig. 2-b and c), we use these motion-based saliency clues to create depth-like maps for RGB-D SOD model training, even though these maps do not represent true physical depth. Unlike conventional depth estimation methods that aim to generate accurate depth maps, our approach focuses on creating saliency-informed depth maps, as not all aspects of conventional depth maps, such as the depth layout (Fig. 2-f), are relevant for the SOD task. By using motion-based saliency, we ensure that the generated depth maps are more useful for saliency detection, even if they do not correspond to real-world depth values.

Any RGB-D SOD models can be easily upgraded with newly and automatically constructed video-based training data generated using our approach. This makes our method both generic and practical. It should be emphasized that our newly augmented trainable video data not only enhances state-of-the-art RGB-D SOD models in new, specific scenarios, such as the real-world stereo datasets BDD [14], KITTI, and CityScapes [19], but also delivers significant performance improvements on common scenes, as demonstrated on the seven widely-used RGB-D SOD benchmark datasets.

In summary, the contributions of this paper can be summarized in the following aspects:

- As the first attempt, we have presented a data augmentation method to adapt off-the-shelf SOTA RGB-D SOD models to perform well on unfamiliar RGB-D scenes and provided a novel insight towards handling the domain shift problem in the SOD task;
- We have explored how depth information could serve the learning when performing the RGB-D SOD task and proposed a simple yet effective approach to generate saliency informative depth maps;
- We have taken full advantage of videos' spatiotemporal information to generate high-quality SOD pseudo-GTs;

II. Related Work

A. RGB-D Salient Object Detection

Traditional RGB-D SOD methods [20], [21] focus on hand-crafted low-level features and thus struggle to handle complex scenes. To address this limitation, deep learning-based methods have emerged. These fusion-wise models fuse RGB and depth images in different stages, meanwhile extract high-level representations [22], [23], [24], [25], [26], [27] and obtain multi-scale features from different levels [28], [29] to improve the performance, which can be categorized into early fusion [30], [31], late fusion [32], [33], and mid fusion [34], [35], [36], [37], [38], [39]. Moreover, large-scale datasets like ReDweb-S [40] and COME15K [41] have been proposed to alleviate the issue of limited training data. Nonetheless, even with these large datasets, existing methods struggle to achieve satisfactory results in novel scenarios with domain

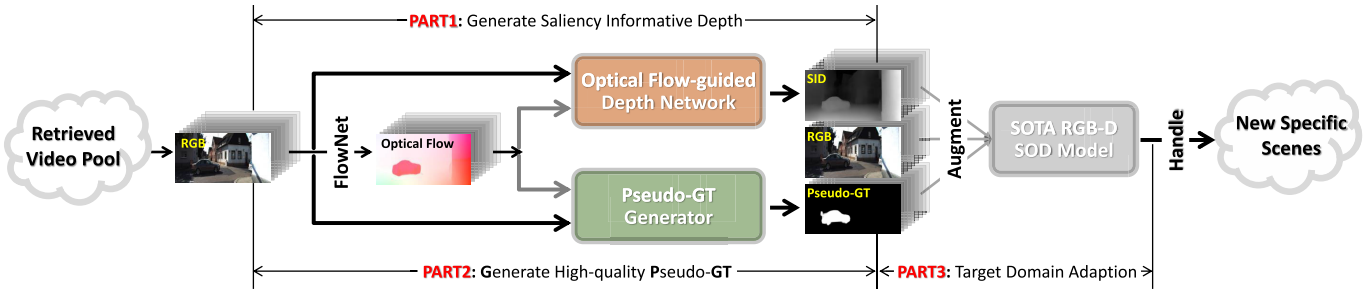

Fig. 3. Our method pipeline has three stages. Starting with retrieved video data, we extract optical flow maps, which generate saliency informative depth (SID) maps in (**PART1**) and high-quality pseudo-ground truths (pseudo-GTs) in (**PART2**). These enhanced representations augment the video data, allowing fine-tuning of a state-of-the-art (SOTA) RGB-D salient object detection (SOD) model. In (**PART3**), this refined model addresses domain adaptation challenges, ensuring strong performance in new, unseen scenes.

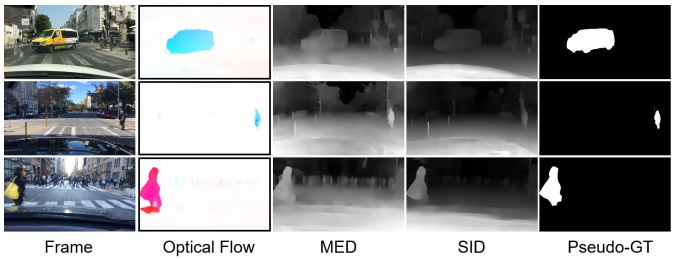

Fig. 4. We visually demonstrate the advantages of our saliency informative depth maps (SID) over monocular estimated depth (MED) on specific scenes. While our SID (column #4) lacks the detailed depth layouts of conventional monocular depth estimation (column #3), it still performs well in RGB-D SOD tasks. This is because it effectively utilizes layer-separable cues (Fig. 2-e) to distinguish between salient objects and non-salient backgrounds.

shift issues. Thus, we aim to improve performance in these specific scenarios by retrieving video data with similar scenes from a data perspective without modifying the architecture of state-of-the-art models.

B. Weakly Supervised RGB-D Saliency Detection

Since annotating pixel-wise masks is arduous and time-consuming, existing weakly-supervised methods mainly obtain saliency from low-cost annotations, *i.e.*, bounding-boxes [42], image-level labels [43], scribbles [44], and noisy labels [45]. For example, [43] adopted an iterative learning strategy to update an initial saliency map generated from unsupervised saliency methods by learning with image-level supervision. Reference [45] proposed a noise-robust adversarial learning framework to avoid error-prone predictions generated by pseudo-label noise. Reference [42] iteratively refined the predicted pixel-level pseudo-GT saliency maps with saliency bounding boxes. Reference [44] proposed a prediction consistency training method and an active scribble-boosting strategy to provide extra supervision signals with negligible annotation cost. Reference [46] leveraged unlabeled RGB images to generate depth to boost RGB-D saliency detection.

Differently from these approaches, we utilize video data, incorporating optical flow to enhance RGB-D SOD performance in novel specific scenarios.

C. Optical Flow-Guided Depth Estimation

Optical flow, a fundamental tool in video processing, captures the motion of objects between frames, offering vital information for distinguishing between static and dynamic elements. This motion data is inherently layer-separable, making it a robust feature for generating depth maps that highlight moving objects. Traditional monocular depth estimation methods often emphasize producing rich layouts and highly detailed depth maps, which may not be optimal for certain downstream tasks, such as RGB-D salient object detection (SOD). For example, Kopf et al. [47] proposed Robust Consistent Video Depth Estimation (RCVD), which ensures robust depth maps under noisy conditions but lacks motion segmentation capabilities, making it less effective for distinguishing dynamic elements. Xu et al. [48] introduced a Unified Flow, Stereo, and Depth Estimation framework that excels in feature matching but is not optimized for saliency detection. Shimada et al. [49] focused on real-time flow-to-depth estimation for drones but struggled with separating moving and static objects. Guo et al. [50] developed F2Depth for low-texture indoor scenes, yet it performs poorly in dynamic settings, while Guizilini et al. [51] introduced DRAFT for joint optical flow and depth learning but lacks saliency prioritization. Lu and Chen [52] tackled dynamic object segmentation using joint depth and flow estimation but without leveraging cross-stream refinement.

On the contrary, our proposed method integrates optical flow and RGB streams through dense connections and boundary supervision, effectively separating dynamic and static objects to generate saliency-aware depth maps. This focus on retaining salient motion cues and precise boundary segmentation gives our method an edge in distinguishing dynamic objects compared to existing methods.

D. Domain Shift

The domain shift issue arises when there are differences and gaps between the distributions of the source and target domains. To address this challenge, current methods [15], [16] typically employ domain adaptation techniques, which can be categorized into sample adaptation, feature adaptation, and model adaptation. Recent works in object detection [53], [54], semantic segmentation [55], and depth estimation [56] have introduced strategies such as domain-adversarial training, self-training, and contrastive learning to improve generalization across domains. These techniques have proven effective in handling domain shifts in various visual tasks, offering

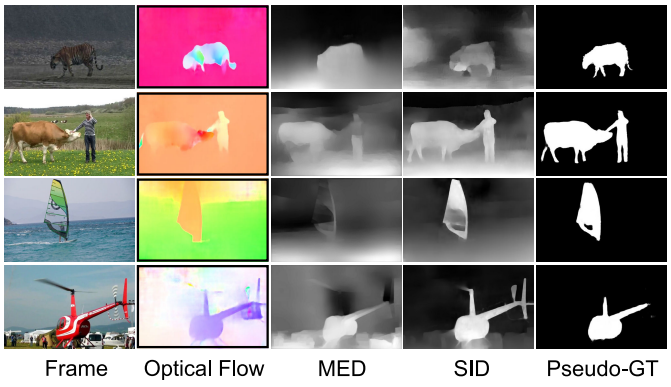

Fig. 5. Visual demonstrations to show the advantages of our SID over other MED [60] on normal scenes.

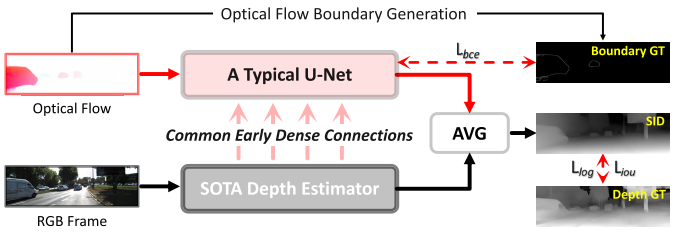

Fig. 6. The Optical Flow-guided Depth Network enhances the state-of-the-art MiDaS [60] method by adding optical flow maps as an additional stream, integrated with the RGB stream through early dense connections. We use boundary ground truth from high-quality optical flow for supervision. Our saliency informative depth maps (SID) focus primarily on salient cues rather than the complete depth layout.

potential insights for RGB-D SOD tasks. However, these approaches generally focus on adapting source knowledge to the target domains, while incremental learning [57], [58], [59] — a lifelong learning method — emphasizes maintaining performance on the source domain while incorporating new concepts from the target domain without forgetting previously learned ones.

Unlike other tasks, RGB-D SOD faces additional challenges from depth information, which makes domain adaptation harder due to variations in scene geometry and sensor noise. Our approach, which uses video data and optical flow, aims to bridge this domain gap by providing scene-specific data without altering the architecture of SOTA models. This makes it particularly valuable for scenarios with limited target domain data, improving the robustness and adaptability of models in real-world applications.

III. The Proposed Method

A. Method Pipeline

Our approach aims to enhance the performance of state-of-the-art RGB-D models in specific new scenes by leveraging plain video data (without depth or pixel-wise SOD ground truth). The overall method pipeline is depicted in Fig. 3, which consists of three main parts — **PART1**: Given a plain video, this part generates saliency informative depth maps (SID). The SID utilizes layer-separable clues from optical flow maps to improve conventional depth estimation, making it highly informative for the SOD task. However, this comes at the cost of losing some depth-related details. **PART2**: This part fully exploits the complementary spatiotemporal information from the video to generate high-quality SOD pseudo-GTs. We pre-generate a large number of pseudo-GTs, but only retain a few frames with high-quality labels. These high-quality pseudo-GTs are then used for training the SOD network. **PART3**: Now that the video data is enriched with SIDs and pseudo-GTs, this part addresses the domain shift problem by using these enhanced video inputs to improve performance in novel scenes.

By clearly separating these stages, our method efficiently handles domain shift issues and boosts model performance in specific new scenes, all without altering the architecture of existing SOTA RGB-D models.

B. Generate Saliency Informative Depth

It is well known that optical flow has been widely used in various video-related tasks to sense motions, which provides informative clues to separate moving objects from other static ones. A qualitative demonstration of 'Optical Flow' can be seen in Fig. 4 and 5.

Our first objective is to obtain video-correlated D maps to make video data applicable for SOTA RGB-D SOD models' training. However, conventional monocular depth estimation approaches [61], [62] mainly focus on generating depth maps with *good layout* and *rich details*, which might not be necessary for RGB-D SOD task (see Fig. 2). Noticed by the distinct attribute of optical flow maps (see Fig. 4 and 5) — very informative in providing layer-separable clues since they can well highlight moving objects, it is technically appropriate to use them to facilitate depth estimation in SOD task.

To enable a generic application, we propose to enhance the existing monocular depth estimation approach to generate saliency informative depth maps (SID) via optical flow maps (**PART1** of Fig. 3). As is shown in Fig. 6, our approach, coined as Optical Flow-guided Depth Network (OFDNet), is based on the SOTA monocular depth estimation method MiDaS [60]. The major difference is that our OFDNet focuses on retaining the informative layer-separable clues provided by optical flow maps. Layer-separable clues refer to features extracted from optical flow maps that help distinguish moving objects from static backgrounds, essential for saliency detection. By highlighting object boundaries and motion dynamics, these clues enable models to focus on significant motion aspects rather than comprehensive depth layouts. This emphasis allows for better identification of salient objects, as the distinct motion patterns of these objects enhance the model's differentiation between salient and non-salient areas.

First, we generate optical flow maps for each frame using the off-the-shelf RAFT [64] model, which captures motion information by highlighting layer-separable clues, such as moving objects and their boundaries. Next, we develop a bi-stream network based on MiDaS, where the RGB stream follows the depth estimation model (e.g., MiDaS [60]), focusing on extracting detailed depth information from static scenes. Simultaneously, the optical flow stream processes motion cues through a U-Net, which specializes in capturing boundary information related to moving objects. To enhance the

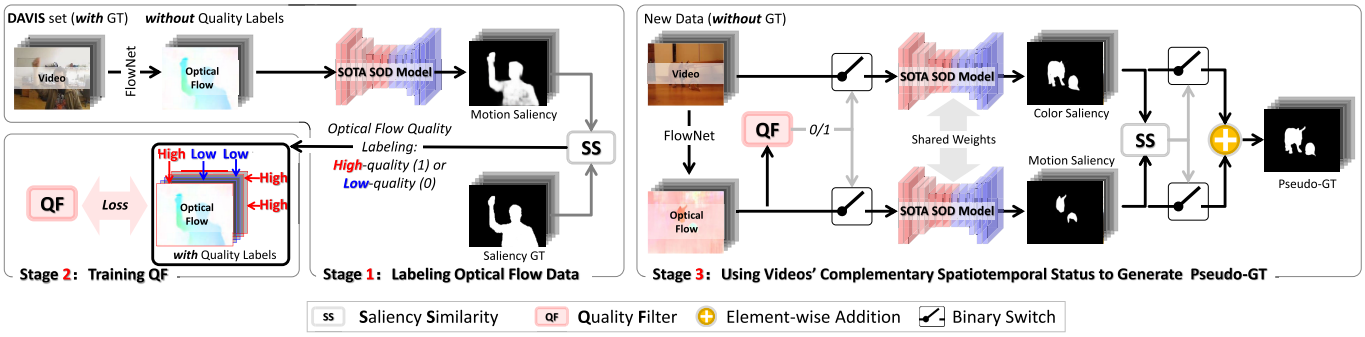

Fig. 7. Pipeline illustration of Pseudo-GT Generator. We have devised a three-stage process to mine high-quality Pseudo-GT. Stage 1 aims at assigning quality labels of optical flow maps to train quality filter (QF). Stage 2 is the training phase of QF, which is the key to obtaining high-quality Pseudo-GT. Stage 3 utilizes the complementary spatiotemporal status of videos to generate high-quality Pseudo-GT. Specifically, we use saliency similarity (SS) to measure the spatiotemporal consistency between color saliency and motion saliency, ensuring that the generated Pseudo-GTs are aligned with both visual and motion cues. Further, we apply CRF [63] operation after obtaining initial Pseudo-GT to acquire sharper ones.

interaction between the two streams, we establish dense early connections, where intermediate features from MiDaS are passed to corresponding layers in the U-Net via a *Concatenation + Convolution* operation. This allows the U-Net to leverage depth information while refining its boundary detection. Finally, the outputs from both streams are fused through a simple average (AVG) operation, resulting in saliency informative depth (SID) maps that emphasize salient regions and objects in the scene, improving depth estimation for video-based tasks. The overall dataflow of OFDNet can be formulated as follows:

$$\mathcal{O}(\mathbf{I}_i, \mathbf{OF}_i) = \mathcal{AVG}(\mathrm{MSI}(\mathbf{I}_i), \mathrm{UNet}(\mathrm{MSI}(\mathbf{I}_i), \mathbf{OF}_i)), \quad (1)$$

where $\mathbf{I}$ is the input RGB frame, $\mathbf{OF}$ is the corresponding optical flow, $\mathcal{AVG}$ is an average operation, MSI denotes the SOTA depth estimator MiDaS [60], and $\mathcal{O}$ represents our OFDNet. Notice that, compared with the conventional monocular depth estimation MiDaS, our SID can provide more discriminative information to separate salient objects from their non-salient surroundings (see Fig. 4), and thus it has the potential to facilitate the RGB-D SOD task better.

During training, the two streams utilize individual loss functions. As shown in Fig. 6, the U-Net stream takes the optical flow-based Canny boundary maps as supervision, in which we use the typical binary cross entropy loss (L_{bce}) for training. The MiDaS stream takes real depth maps for supervision, and we use the classic Log-mse loss (L_{log}) and intersection over union loss for training L_{iou}. Thus, the overall loss function (L_{all}) can be expressed as:

$$\mathrm{L}_{all} = \mathrm{L}_{bce} + \mathrm{L}_{log} + \mathrm{L}_{iou}. \quad (2)$$

We have trained our OFDNet on the KITTI training set. Unlike conventional approaches that utilize optical flow for dense depth refinement [60], our OFDNet leverages optical flow maps specifically to enhance salient region accuracy, making it a task-specific enhancement rather than a general-purpose refinement. To highlight the generic nature of our method, we have only employed the simplest network architecture here. A more sophisticated network architecture with appropriate loss functions may potentially yield further performance improvements. However, to stay focused on the main topic, we suggest leaving this as a direction for future research.

C. Generate High-Quality Pseudo-GT

We can easily obtain high-quality SID of the given plain video sequence using the abovementioned OFDNet. However, to make video data trainable for RGB-D SOD models, we are still short of pixel-wise SOD ground-truth (GT).

To solve this problem, we present a feasible way to generate high-quality Pseudo-GT (**PART2** of Fig. 3). As shown in Fig. 7, our pseudo-GT generation includes three stages: 1) stage 1 assigns optical flow maps with binary quality labels, *i.e.*, high-quality (1) or low-quality (0), 2) stage 2 trains the newly-devised quality filter (QF), and 3) stage 3 uses videos' complementary spatiotemporal status to generate pseudo-GT, where QF is the key to ensure the quality of the generated pseudo-GT.

The primary objective of stage 1 and stage 2 is to obtain the QF to exclude low-quality optical flow maps because it is almost infeasible to ensure the pseudo-GT's quality when the frame's optical flow is low-quality. Thus, in our pseudo-GT generation process, only the frames with high-quality optical flow maps are considered. The proposed QF is a typical binary classifier, which takes optical flow as input, followed by a pre-trained feature backbone (ResNet50), then a multi-layer perceptron (MLP) outputting either 1 (high-quality) or 0 (low-quality). To train QF, we propose a method to automatically label the quality of optical flow maps using saliency similarity (SS).

The rationale of SS is based on the fact that a high-quality optical flow map usually correlates to high-quality motion saliency, where the motion saliency can be easily obtained by feeding the optical flow map to any off-the-shelf RGB SOD model.[1] Therefore, we can directly measure the consistency degree between the motion saliency (MS) and real saliency GT to reveal the quality degree of the optical flow map, where we call this consistency measuring process between the two saliency maps as SS. The SS process can be detailed as

[1]We simply choose the EDN [65], which was trained on SOD datasets without equipping D (*i.e.*, DUTS [43]).

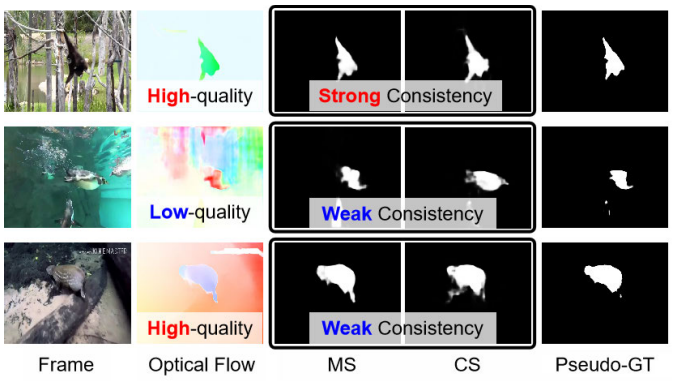

Fig. 8. Rationale demonstration of Pseudo-GT Generator. There are 4 cases when generating Pseudo-GTs, *i.e.*, 'high/low-quality optical flow + strong/weak color saliency (CS) and motion saliency (MS) consistency'. Since the low- and strong- combination is very rare in practice, we have omitted it. Only the high- and strong- case (the 1st row) can ensure a high-quality Pseudo-GT.

follows:

$$\mathrm{SS}(\mathrm{MS_i}, \mathrm{GT_i}) = \begin{cases} 1, & if \quad \mathrm{S}_m(\mathrm{MS}_i, \mathrm{GT}_i) - \gamma > 0 \\ 0, & otherwise, \end{cases} \tag{3}$$

where MS_i and GT_i are the motion saliency and GT of the i-th frame, γ is a pre-defined hard threshold, and S_m means S-measure [66] which measures the structure similarity between its two inputs. In stage 1, using the equation mentioned above, we automatically assign a binary label to each training frame to indicate its optical flow quality. The optical flow quality label reflects the accuracy and reliability of the optical flow map for the given frame. Then, in stage 2, we train the QF on the newly-labeled optical flow data. To train QF, we simply use DAVIS set [67] because all video frames in this set have been equipped with pixel-wise SOD GT.

In stage 3, we utilize trained QF to facilitate pseudo-GT generation. The most intuitive way to generate pseudo-GT for video data is to perform the average operation on color saliency and motion saliency, where the color saliency can be obtained similarly to the motion saliency, *i.e.*, input an RGB frame into an off-the-shelf SOD model. However, the pseudo-GTs generated by such a naive method are not always high-quality. Thus, stage 3 follows an innovative methodology to handle the problem, ensuring all obtained pseudo-GT are high-quality.

Our approach is based on the common attribute of the abovementioned average operation-based high-quality pseudo-GTs: 1) with high-quality optical flow and 2) with strong consistency between color and motion saliency. It implies that a pseudo-GT satisfying both items is likely to be high-quality. We have provided a visual demonstration in Fig. 8 for a better understanding.

As shown in the right part of Fig. 7, given new video data (without GT), only those frames with high-quality optical flow maps (via QF) are fed into the SOTA SOD model to produce color/motion saliency. However, since there exist massive cases in frames with high-quality optical flow maps, they are still incapable of obtaining high-quality pseudo-GT (*e.g.*, the 3rd row of Fig. 8). Therefore, we only retain those frames with high-quality optical flow maps and exhibit strong consistency between their color and motion saliency (*e.g.*, the 1st row of Fig. 8) as final pseudo-GTs. Specifically, we again employ SS to compute the similarity between motion and color saliency.

In short, given a video sequence, only those frames with non-negative 'pseudo-GT scores' will be retained, where the "pseudo-GT scores (pGTscore)" can be computed by the following equation:

$$\mathrm{pGTscore} = \mathrm{QF}(\mathrm{OF_i}) \times \mathrm{SS}(\mathrm{CS_i}, \mathrm{MS_i}), \tag{4}$$

where OF means optical flow maps, while MS and CS denote motion and color saliency, respectively. Specifically, the γ value here is set to the same as the abovementioned hard threshold (Eq. 3). Following the common thread, we employ CRF [63] to refine the obtained pseudo-GTs for slightly better quality.

D. Target Domain Adaption

As previously mentioned, existing state-of-the-art (SOTA) RGB-D SOD models are primarily trained on datasets featuring common scenes. Due to the domain shift issue, these models struggle to perform well in new, specific scenes (as illustrated on the right in Fig. 3). However, leveraging the methods introduced in **PART1** and **PART2**, we can effectively adapt the target domain (training set with domain shift issues) to the source domain (testing set with new specific scenes). This is done by using a small set of images from the new specific scenes to retrieve a large amount of video data with similar scenes, and then applying our proposed approach to equip some of this data with high-quality pseudo-GTs and saliency-informed depth maps. Finally, we can use these newly generated trainable video data to re-train the RGB-D SOD models, effectively addressing the domain shift issue, as described in **PART3**.

Technically, given a SOTA RGB-D SOD model, we combine its original training sets with our newly generated video-based training set and re-train the model on this combined dataset. All other training settings remain the same as the default choices. While there are ways to accelerate the re-training process, achieving similar quantitative results with reduced computational cost, this topic is beyond the main scope of this paper and will be explored in future work. The detailed procedure of the entire method can be found in the pseudocode (Algorithm 1).

IV. Experiments

A. Datasets and Evaluation Metrics

1) Training Sets: Our training datasets consist of two groups.

The 1st group consists of the original training sets of the targeted baseline models (4 top-tier SOTA models selected as targeted baseline models, *e.g.*, LAFB [68], SPNet [69], SSL [70], and C^{2}DF [71]) and our newly augmented trainable video data with common scenes of SOD cases. We evaluate the effectiveness of our approach on nine widely-used public benchmark datasets, *e.g.*, DUT-RGBD [72], NJUD [11],

TABLE I

QUANTITATIVE EVALUATION OF MAJOR COMPONENTS USED IN OUR APPROACH. THIS EXPERIMENT ADOPTS C^{2}DF AS THE BASELINE MODEL AND FULL DESCRIPTIONS REGARDING MARKS FROM ① TO ⑤

	Major Components										Datasets and Metrics													
	Generate SID				Generate Pseudo-GT				Training		NJUD				NLPR				LFSD				SSD			
	COG	OF	M	B	rGT	pGT	QF	SS	MED	SID	Sm↑	Fm↑	Em↑	M↓	Sm↑	Fm↑	Em↑	M↓	Sm↑	Fm↑	Em↑	M↓	Sm↑	Fm↑	Em↑	M↓
0 1	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	.907	.898	.918	.039	.914	.899	.955	.024	.851	.853	.883	.065	.869	.842	.915	.048
2	✓	✘	✘	✘	✘	✓	✓	✓	✘	✘	.910	.901	.921	.038	.915	.903	.957	.023	.851	.855	.886	.065	.870	.846	.917	.045
3	✘	✓	✘	✘	✘	✓	✓	✓	✘	✓	.915	.905	.928	.035	.918	.903	.958	.022	.854	.858	.890	.065	.872	.847	.918	.044
1 4	✘	✓	✓	✘	✘	✓	✓	✓	✘	✓	.919	.909	.932	.034	.924	.905	.960	.021	.859	.861	.897	.064	.875	.851	.919	.042
5	✘	✘	✓	✓	✘	✓	✓	✓	✘	✓	.923	.913	.937	.031	.928	.908	.962	.020	.863	.865	.902	.065	.877	.852	.921	.039
6	✘	✓	✓	✓	✘	✓	✓	✓	✘	✓	.925	.918	.940	.029	.930	.909	.963	.020	.868	.867	**.904**	.064	.880	.856	.924	.039
5 7	✘	✓	✓	✓	✘	✓	✘	✘	✘	✓	.912	.903	.922	.036	.918	.905	.956	.023	.853	.856	.888	.064	.873	.849	.918	.042
5 8	✘	✓	✓	✓	✓	✓	✘	✘	✘	✓	.915	.905	.924	.035	.918	.905	.957	.023	.855	.859	.891	.064	.875	.850	.919	.042
2 9	✘	✓	✓	✓	✓	✓	✓	✘	✘	✓	.920	.912	.930	.032	.923	.906	.959	.022	.856	.864	.896	.065	.876	.853	.919	.041
10	✘	✓	✓	✓	✓	✓	✘	✓	✘	✓	.922	.916	.936	.031	.925	.908	.961	.021	.862	.866	.900	.064	.879	.855	.921	.040
4 11	✘	✓	✓	✓	✓	✓	✓	✓	✓	✘	.925	.917	.939	.031	.929	.908	.961	.020	.865	.867	.903	**.063**	.880	.854	.923	.039
3 12	✘	✓	✓	✓	✓	✓	✓	✓	✘	✓	**.926**	**.919**	**.942**	**.028**	**.932**	**.911**	**.964**	**.019**	**.871**	**.868**	**.904**	**.063**	**.881**	**.857**	**.925**	**.037**

0 **Baseline** — **OF:** optical flow stream of our OF-guide depth network — **M:** original RGB stream of SOTA MED method
1 *Verify* **Gen SID** — **rGT:** train SOTA RGB-D models via **{DAVIS set}** + **{real-GT}** — **B:** train our OF-guide depth network using boundary GT
2 *Verify* **Pseudo-GT** — **pGT:** train SOTA RGB-D models via **{plain videos}** + **{pseudo-GT}** — **SS:** saliency similarity
3 4 *Verify* **SID** — **MED:** SOTA monocular estimated D method [**MiDaS**] — **QF:** optical flow quality filter
5 *Verify* **rGT** — **SID:** saliency informative depth maps — **COG:**simply convert optical flow maps to gray maps

NLPR [73], SIP [74], SSD [75], LFSD [76], STEREO [77], COME15K-E [41] and ReDweb-S [40]. The plain video data are collected from HMDB51 [78], UCF101 [79], GOT-10K [80], VOT2020 [81] and DAVIS (with real GT), totally 8K images (called Video8K).

The 2nd group of our datasets includes the original training sets of the targeted baseline models and our newly augmented trainable video data, sourced from specific scenes that present domain shift challenges, such as BDD, KITTI, and CityScapes. Specifically, we selected 2,000 images from BDD, allocating 1,400 for training and 600 for testing, along with 600 images from the KITTI testing set and 855 images from the CityScapes testing set for evaluation. BDD offers diverse driving scenarios with variations in weather, time of day, and urban environments, though it underrepresents rare events and extreme conditions. KITTI provides high-quality annotations in urban and highway settings but is limited to daytime and clear weather conditions from a single geographic location. CityScapes focuses on semantic understanding with detailed pixel-level annotations across multiple German cities, yet it encompass a variety of weather conditions, different times of the day, and urban traffic patterns. The depth maps for all augmented training and testing data are generated using our proposed approach, ensuring consistency. Additionally, ground truth (GT) maps for training data are produced by our method, while those for the BDD, KITTI and CityScapes testing sets are newly labeled using the Tobii eye-fixation capturing device to accurately identify real salient objects and regions.

2) Evaluation Metrics: To perform quantitative evaluation, we have adopted 4 commonly-used evaluation metrics, including the F-measure value (Fm) [82], the E-measure (Em) [83], the mean absolute error ($\mathcal{M}$), and the structure measure value (Sm) [66]. Notice that the values of Fm, Em, and Sm are the larger, the better, while the value of $\mathcal{M}$ is smaller, the better.

B. Implementation Details

We implement our method using PyTorch on an NVIDIA GeForce RTX 3090 GPU. In the training phase, hyper-parameters such as initial learning rate, optimizer, weight decay, and input image size are all the same as targeted models. For a fair comparison, we train the targeted model from scratch using the same batch size with different composite datasets, which will degrade model performance slightly.

In addition, we also adopt a horizontal flip and random cropping method for data enhancement in the same setting as the targeted models. The inferring speed is also consistent with the targeted models, *e.g.*, C^{2}DF of 78 fps. The hyper-parameter γ in stages 1 and 3 of **PART2** are empirically set as 0.85. The ablation study is shown in Table V-**B**.

C. Component Evaluation

We have conducted an extensive component evaluation to verify the effectiveness of major components used in our approach, and the quantitative results can be seen in Table I. The 1st row denoted by mark ⓪ is the baseline model (C^{2}DF). The effectiveness of saliency informative depth (SID) towards RGB-D SOD model's training can be easily observed by mark ①. Compared with optical flow-based gray maps (convert the optical flow to grayscale directly), our SID can persistently improve all metrics, *e.g.*, the Fm metric has been enhanced from.901→.918 in NJUD testing set (line 2 *vs.* line 6).

The effectiveness of using generated Pseudo-GT for RGB-D SOD model training has been shown by mark ②, in which QF and SS are two key components to exclude frames with low-quality optical flow. By comparing lines 8 and 12, the MAE ($\mathcal{M}$) metric in NJUD has been decreased from.035→.031, showing the effectiveness of our QF and SS to generate high-quality Pseudo-GT.

TABLE II

QUANTITATIVE COMPARISON WITH CURRENT SUPERVISED SOTA RGB-D SOD MODELS. OUR APPROACH IS WEAKLY-SUPERVISED, WHICH CAN PERSISTENTLY IMPROVE TARGET SOTA MODELS SIGNIFICANTLY. WE HAVE TESTED OUR APPROACH OVER 4 TOP-TIER SOTA MODELS, *i.e.*, LAFB, C^{2}DF, SSL, AND SPNET; TARGET MODELS PROMOTED BY OUR APPROACH (VIDEO8K) ARE DENOTED BY *. THE TOP-2 RESULTS ARE MARKED IN **RED** AND **BLUE**

Set	Model	S2MA	A2dele	ICNet	D3Net	UCNet	DCF	BBS	CIRNet	MP	LSNet	MIRV	EMTr	MFUR	LAFB	LAFB*	Gain	SPNet	SPNet*	Gain	SSL	SSL*	Gain	C^{2}DF	C^{2}DF*	Gain
	Year	2020	2020	2020	2021	2021	2021	2021	2022	2023	2023	2024	2024	2024	2024	Ours	Ours	2022	Ours	Ours	2022	Ours	Ours	2022	Ours	Ours
DUT-RGBD	Sm↑	.903	.887	.852	.775	.863	.870	.870	.908	.896	.871	.875	.923	.925	.917	.929	+1.2%	.877	.895	+1.8%	.901	.917	+1.6%	.894	.923	+2.9%
	Fm↑	.901	.890	.784	.716	.857	.865	.861	.912	.889	.883	.866	.925	.928	.919	.938	+1.9%	.869	.878	+0.9%	.894	.920	+2.6%	.901	.925	+2.4%
	Em↑	.937	.924	.901	.789	.904	.911	.909	.922	.921	.913	.889	-	-	.948	.951	+0.3%	.915	.925	+1.0%	.933	.940	+0.7%	.929	.944	+1.5%
	M↓	.043	.043	.072	.097	.056	.056	.064	.046	.035	.038	.041	.029	.027	.027	.025	−0.2%	.057	.051	−0.6%	.038	.031	−0.7%	.043	.028	−1.5%
NJUD	Sm↑	.894	.871	.894	.892	.897	.906	.919	.915	.916	.911	.890	.908	.923	.907	.915	+0.8%	.914	.925	+1.1%	.909	.921	+1.2%	.907	.926	+1.9%
	Fm↑	.889	.874	.843	.863	.886	.897	.899	.897	.916	.900	.880	.918	.921	.910	.923	+1.3%	.890	.917	+2.7%	.884	.889	+0.5%	.898	.919	+2.1%
	Em↑	.930	.897	.913	.913	.915	.918	.919	.922	.929	.922	.929	-	-	.924	.933	+0.9%	.920	.931	+1.1%	.928	.939	+1.1%	.918	.942	+2.4%
	M↓	.053	.051	.052	.047	.043	.039	.037	0.35	.039	.038	.046	.038	.035	.028	.023	−0.5%	.036	.030	−0.6%	.038	.032	−0.6%	.039	.028	−1.1%
NLPR	Sm↑	.915	.898	.923	.902	.920	.919	.826	.923	.910	.918	.914	.925	.928	.918	.929	+1.1%	.926	.930	+0.4%	.909	.919	+1.0%	.914	.932	+1.8%
	Fm↑	.902	.878	.908	.857	.890	.899	.878	.914	.889	.855	.895	.921	.912	.902	.923	+1.1%	.901	.906	+0.5%	.884	.887	+0.3%	.899	.911	+1.2%
	Em↑	.953	.945	.952	.943	.953	.958	.949	.952	.954	.956	.953	-	-	.958	.965	+0.7%	.954	.961	+0.7%	.939	.954	+1.5%	.955	.964	+0.9%
	M↓	.030	.028	.028	.030	.025	.023	.028	.023	.031	.024	.025	.027	.022	.021	.017	−0.4%	.024	.021	−0.3%	.038	.027	−1.1%	.024	.019	−0.5%
SIP	Sm↑	.872	.829	.854	.860	.873	.866	.876	.888	.890	.886	.876	.887	.890	.887	.899	+1.2%	.869	.884	+1.5%	.871	.885	+1.4%	.876	.891	+1.5%
	Fm↑	.849	.825	.791	.835	.868	.862	.874	.885	.885	.884	.863	.891	.892	.883	.891	+0.8%	.872	.879	+0.7%	.875	.882	+0.7%	.873	.889	+1.6%
	Em↑	.911	.892	.900	.902	.913	.912	.915	.923	.925	.927	.924	-	-	.925	.935	+1.0%	.908	.920	+1.2%	.921	.925	+0.4%	.911	.928	+1.7%
	M↓	.058	.070	.070	.063	.051	.058	.056	.052	.042	.048	.049	.045	.049	.042	.039	−0.3%	.055	.049	−0.6%	.046	.043	−0.3%	.050	.041	−0.9%
STEREO	Sm↑	.890	.879	.891	.885	.903	.886	.909	.913	.912	.874	.882	.911	.920	.912	.925	+1.3%	.899	.914	+1.5%	.886	.913	+2.7%	.905	.915	+1.0%
	Fm↑	.882	.874	.847	.855	.885	.883	.886	.896	.896	.855	.857	.915	.919	.882	.885	+0.3%	.883	.895	+1.2%	.875	.893	+1.8%	.892	.899	+0.7%
	Em↑	.932	.915	.925	.920	.922	.926	.927	.930	.931	.915	.928	-	-	.930	.936	+0.6%	.924	.930	+0.6%	.919	.928	+0.9%	.927	.935	+0.8%
	M↓	.051	.044	.046	.046	.039	.046	.041	.038	.037	.049	.051	.042	.040	.037	.033	−0.4%	.043	.038	−0.5%	.045	.040	−0.5%	.038	.034	−0.4%
LFSD	Sm↑	.837	.834	.848	.825	.854	.842	.856	.865	.864	.833	.849	.856	.863	.858	.875	+1.7%	.847	.854	+0.7%	.834	.857	+2.3%	.851	.871	+2.0%
	Fm↑	.835	.832	.863	.810	.845	.850	.850	.852	.857	.852	.844	.881	.891	.855	.871	+1.6%	.843	.846	+0.3%	.819	.844	+2.5%	.853	.868	+1.5%
	Em↑	.873	.871	.887	.862	.891	.886	.889	.901	.886	.873	.889	-	-	.894	.906	+1.2%	.887	.892	+0.5%	.888	.898	+1.0%	.883	.904	+2.1%
	M↓	.094	.077	.075	.095	.076	.080	.074	.068	.071	.084	.072	.077	.075	.067	.060	−0.6%	.078	.074	−0.4%	.080	.072	−0.8%	.065	.063	−0.2%
SSD	Sm↑	.868	.802	.848	.847	.865	.855	.870	.868	.879	.841	.857	-	-	.875	.886	+1.1%	.865	.869	+0.4%	.855	.871	+1.6%	.869	.881	+1.2%
	Fm↑	.848	.776	.841	.815	.854	.826	.832	.829	.862	.847	.845	-	-	.840	.852	+1.2%	.830	.833	+0.3%	.833	.850	+1.7%	.842	.857	+1.5%
	Em↑	.909	.861	.902	.888	.907	.898	.904	.895	.893	.883	.913	-	-	.922	.928	+0.6%	.899	.903	+0.4%	.896	.911	+1.5%	.915	.925	+1.0%
	M↓	.052	.070	.064	.058	.049	.053	.049	.048	.041	.050	.052	-	-	.041	.035	−0.6%	.047	.044	−0.3%	.050	.042	−0.8%	.048	.037	−1.1%
COME15K-E	Sm↑	.828	.787	-	-	.822	.831	.853	.849	.868	.861	.857	-	-	.868	.870	+0.2%	.852	.861	+0.9%	.850	.857	+0.7%	.860	.865	+0.5%
	Fm↑	.789	.795	-	-	.814	.822	.834	.835	.839	.835	.827	-	-	.843	.846	+0.3%	.838	.842	+0.4%	.832	.840	+0.8%	.835	.842	+0.7%
	Em↑	.836	.838	-	-	.859	.855	.876	.872	.875	.873	.862	-	-	.885	.890	+0.5%	.869	.875	+0.6%	.863	.879	+1.6%	.877	.881	+0.4%
	M↓	.092	.092	-	-	.079	.074	.071	.064	.060	.062	.071	-	-	.060	.058	−0.2%	.063	.061	−0.2%	.068	.065	−0.3%	.064	.059	−0.5%
ReDweb-S	Sm↑	.711	.641	-	.689	.713	.690	.692	.703	.714	.715	.703	.635	.691	.671	.675	+0.4%	.709	.715	+0.6%	.710	.716	+0.6%	.715	.719	+0.4%
	Fm↑	.675	.603	-	.673	.710	.689	.647	.708	.718	.717	.709	.651	.651	.664	.666	+0.2%	.712	.719	+0.7%	.706	.712	+0.6%	.717	.723	+0.6%
	Em↑	.750	.678	-	.768	.794	.763	.709	.748	.765	.763	.754	-	-	.757	.762	+0.5%	.759	.767	+0.8%	.754	.760	+0.6%	.762	.770	+0.8%
	M↓	.140	.160	-	.149	.130	.142	.150	.132	.126	.125	.135	.172	.169	.128	.127	−0.1%	.129	.124	−0.5%	.136	.131	−0.5%	.131	.122	−0.9%

As denoted by mark ③ and ④, the advantage of our SID (line 12) against the naive optical flow-based graying (line 2) and the top-tier monocular RGB estimated D (MED) [60] (line 11) can be easily observed, *e.g.*,.865 *vs.* .871 regarding Sm in LFSD (line 12 *vs.* 11). The reason is also quite clear, *i.e.*, MEDs are mainly established from color information, which could be somewhat redundant when performing RGB-D fusion.

Also, the mark ⑤ illustrates the performance gap between models trained using {real-GT (DAVIS)} + {Pseudo-GT (plain videos)} and models trained using {Pseudo-GT (DAVIS+plain videos)} solely. As indicated by numeric results in lines 7 and 8, using real-GT could bring better performance as expected. Yet, such an advantage is marginal, which further shows the effectiveness of our approach to generating Pseudo-GT.

D. Comparisons With SOTA RGB-D SOD Methods

Our extensive experiments include two parts. For the first part, to prove the effectiveness of our approach in common scenes of SOD cases, we have compared our method on four targeted baseline models against the 17 most recent SOTA RGB-D SOD models over 9 widely-used RGB-D SOD datasets. The SOTA models include S2MA [84], A2dele [85], D3Net [74], UCNet [86], ICNet [87], DCF [88], BBS [89], CIRNet [90], MP [91], LSNet [23], MIRV [92],

TABLE III

QUANTITATIVE COMPARISONS BETWEEN FOUR TRAINING SETS ON THREE REAL-WORLD TESTING SETS, BDD, KITTI AND CITYSCAPES [19]. THESE BASELINE TRAINING SETS ARE THE TARGETED SOTA MODELS' ORIGINAL TRAINING SETS, AND THE EXPERIMENT TAKES C^{2}DF AS THE TARGETED MODEL

Target Models		SPNet				SSL				C^{2}DF			
Original		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
+COME15K-TR			✓	✓	✓		✓	✓	✓		✓	✓	✓
+Video8K				✓	✓			✓	✓			✓	✓
+BDD-TR					✓				✓				✓
BDD-TE	Sm↑	.673	.740	.758	**.885**	.688	.790	.827	**.881**	.667	.798	.820	**.892**
	Fm↑	.489	.652	.695	**.776**	.479	.601	.622	**.753**	.454	.682	.711	**.786**
	Em↑	.790	.835	.882	**.907**	.755	.830	.846	**.904**	.741	.859	.886	**.912**
	M↓	.032	.026	.024	**.019**	.039	.035	.031	**.023**	.036	.027	.022	**.017**
KITTI	Sm↑	.594	.698	.745	**.831**	.615	.748	.762	**.824**	.592	.715	.756	**.829**
	Fm↑	.334	.571	.627	**.765**	.324	.594	.643	**.718**	.307	.653	.681	**.774**
	Em↑	.706	.786	.799	**.826**	.648	.771	.790	**.822**	.662	.733	.783	**.831**
	M↓	.036	.031	.028	**.025**	.042	.037	.031	**.028**	.041	.032	.029	**.023**
CityScapes	Sm↑	.699	.715	.741	**.873**	.694	.731	.784	**.878**	.654	.768	.826	**.881**
	Fm↑	.485	.537	.698	**.771**	.416	.632	.715	**.766**	.423	.639	.718	**.782**
	Em↑	.703	.756	.772	**.869**	.687	.715	.814	**.861**	.681	.743	.831	**.874**
	M↓	.029	.026	.025	**.022**	.039	.034	.028	**.025**	.034	.031	.025	**.020**

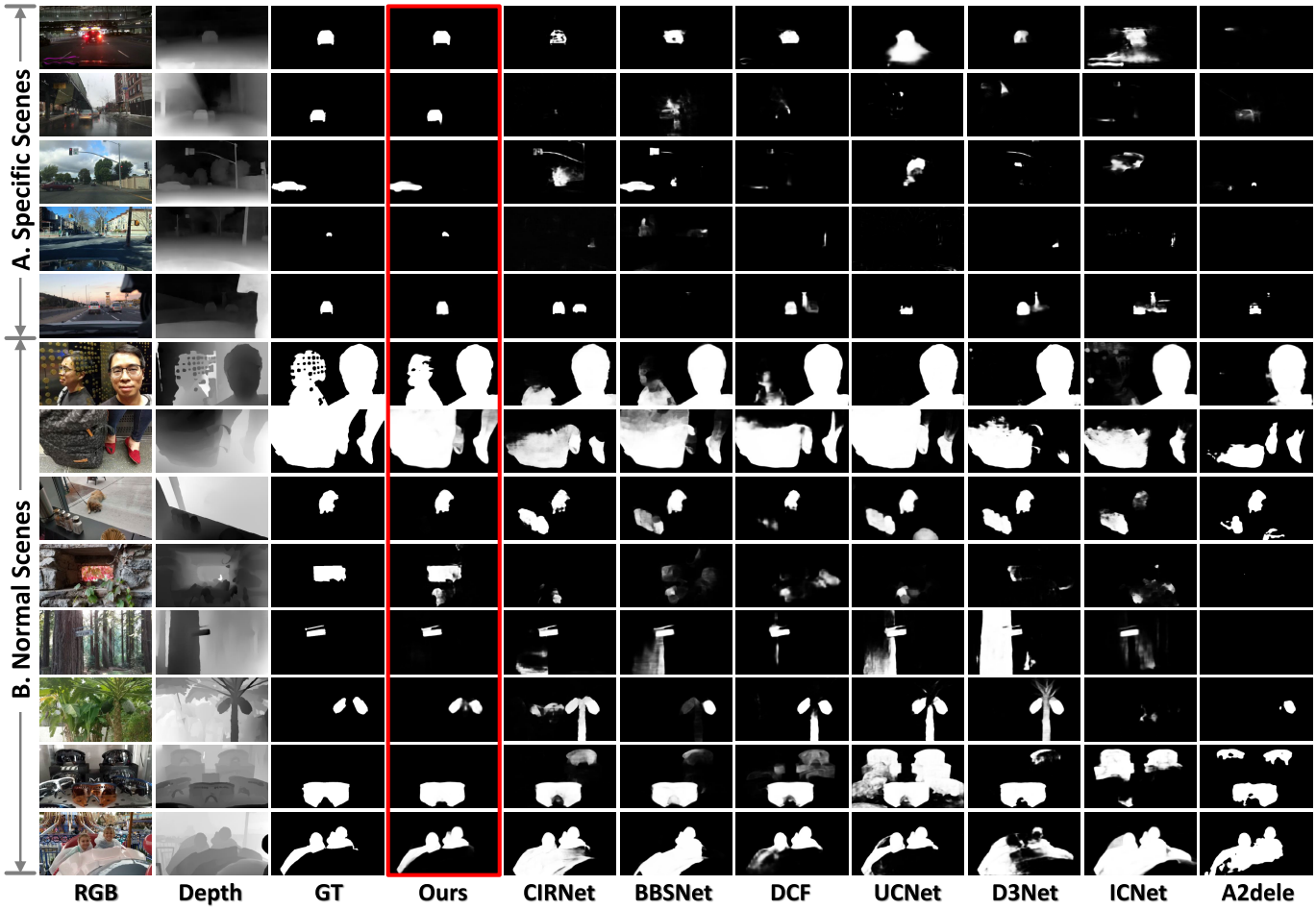

Fig. 9. Visual comparison between our method (based on C^{2}DF) and several most representative SOTA models regarding specific scenes (A) and normal scenes (B).

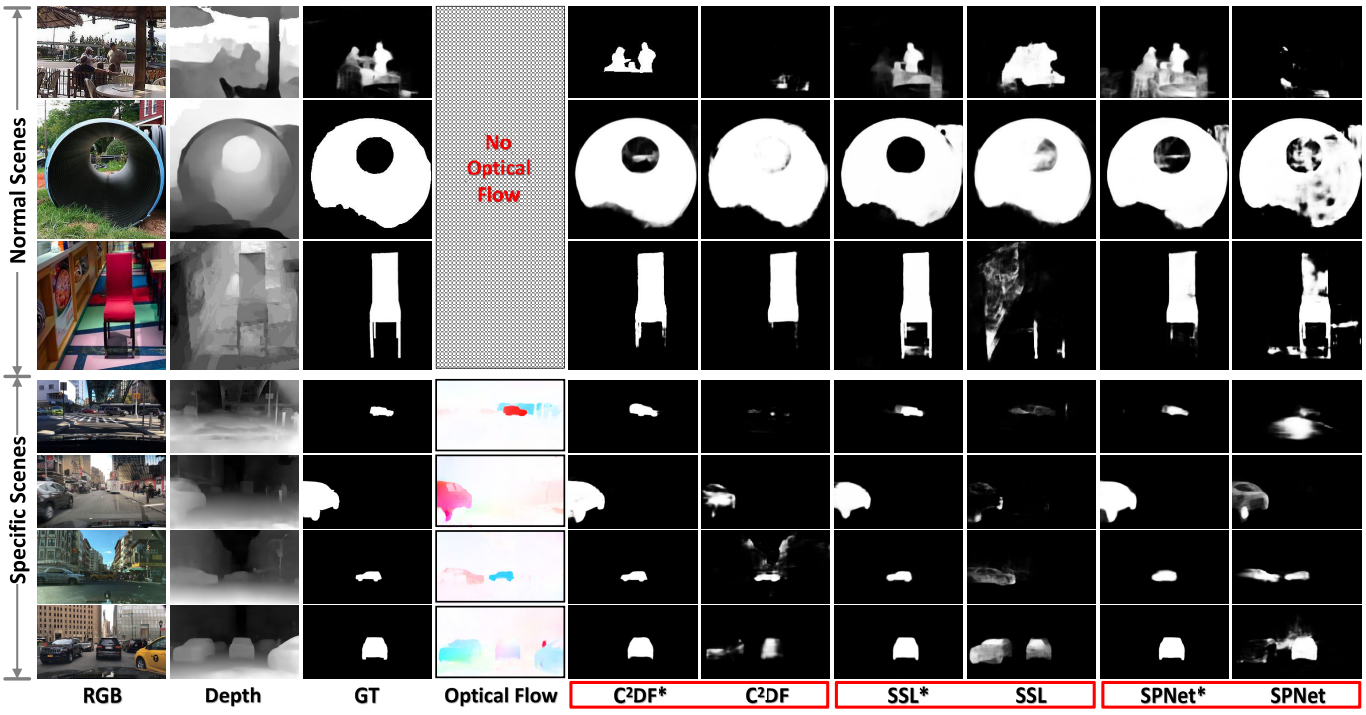

Fig. 10. Visual comparison of normal scenes (w/o optical flow) and specific scenes (BDD and KITTI sets) between three selected target SOTA models (denoted as C^{2}DF, SSL, and SPNet) and their updated versions trained by our newly augmented sets (denoted as C^{2}DF*, SSL*, and SPNet*).

Algorithm 1 Pseudocode for Domain Adaptation of RGB-D SOD in Traffic Scenarios

Part 1: Generate Saliency Informative Depth (SID)
Step 1: Generate Optical Flow Maps
for each frame $\mathbf{I}_i$ in video sequence **do**
  Compute optical flow $\mathbf{OF}_i$ using RAFT model.
  Save $\mathbf{OF}_i$ for further processing.
end for
Step 2: Depth Estimation Using MiDaS
for each frame $\mathbf{I}_i$ in video sequence **do**
  Apply MiDaS model to $\mathbf{I}_i$ to get static depth map $\mathbf{D}_i$.
end for
Step 3: Motion Boundary Extraction Using U-Net
for each optical flow map $\mathbf{OF}_i$ **do**
  Apply U-Net to $\mathbf{OF}_i$ to extract boundary information $\mathbf{B}_i$ related to moving objects.
end for
Step 4: Combine Static and Motion Information
for each frame $\mathbf{I}_i$ **do**
  Concatenate depth map $\mathbf{D}_i$ from MiDaS with boundary information $\mathbf{B}_i$ from U-Net.
  Perform convolution operation on concatenated features to combine depth and motion data.
  Average the outputs from both streams to produce the final Saliency Informative Depth (SID) map:

$$\mathbf{SID}_i = \mathcal{AVG}(\mathbf{D}_i, \mathbf{B}_i)$$

end for
Part 2: Generate High-Quality Pseudo-GT
Step 1: Assign Quality Labels to Optical Flow Maps
for each optical flow map $\mathbf{OF}_i$ **do**
  Calculate motion saliency $\mathbf{MS}_i$ using an off-the-shelf RGB SOD model (e.g., EDN).
  Compute saliency similarity (SS) between motion saliency $\mathbf{MS}_i$ and ground-truth saliency $\mathbf{GT}_i$.
  Assign binary label to $\mathbf{OF}_i$:

$$\text{Quality Label} = \begin{cases} 1, & \text{if SS}(\mathbf{MS}_i, \mathbf{GT}_i) > \gamma \\ 0, & \text{otherwise} \end{cases}$$

end for
Step 2: Train Quality Filter (QF)
Train the binary classifier (QF) using high-quality labeled optical flow maps (via ResNet50 backbone and MLP output).
QF classifies new optical flow maps as high or low quality.
Step 3: Generate Pseudo-GT for Video Sequence
for each frame $\mathbf{I}_i$ with high-quality optical flow **do**
  Generate motion saliency $\mathbf{MS}_i$ and color saliency $\mathbf{CS}_i$ for frame $\mathbf{I}_i$ using RGB SOD model.
  Compute saliency similarity score (SS) between $\mathbf{MS}_i$ and $\mathbf{CS}_i$.
  Calculate pseudo-GT score for frame $\mathbf{I}_i$:

$$\text{pGTscore}_i = \text{QF}(\mathbf{OF}_i) \times \text{SS}(\mathbf{CS}_i, \mathbf{MS}_i)$$

  If $\text{pGTscore}_i \geq 0$, retain frame $\mathbf{I}_i$ as high-quality pseudo-GT.
end for
Step 4: Refine Pseudo-GT Using CRF
Apply Conditional Random Fields (CRF) to refine the high-quality pseudo-GTs obtained in Step 3.
Part 3: Target Domain Adaptation
Step 1: Adapt Target Domain using Generated Data
Combine the original training set with newly generated video data (with pseudo-GT and SID maps).
Re-train RGB-D SOD model using this combined dataset.
Adjust model parameters to handle domain shift effectively.

EMTr [93], MFUR [94], LAFB [68], SPNet [69], SSL [70], and C^{2}DF [71]. For a fair comparison, we use either the code implementations with default parameter settings or saliency maps provided by the authors. Also, we have selected the four most recent top-tier SOTA models as the target baseline models (*e.g.*, LAFB, C^{2}DF, SSL, and SPNet), where we have applied our approach over them to achieve performance gain.

As shown in Table II, targeted baseline models are re-trained by the models' original training datasets (NJUD and NLPR or DUT-RGBD) and our newly-augmented Video8K with common images, denoted by *. Experimental results show that all targeted SOTA models promoted by our approaches can achieve significant performance improvements over the original versions, *i.e.*, our method can make an average of 0.85%, 1.29%, 1.15%, and 1.6% performance improvement in Fm metric of LAFB, C^{2}DF, SSL, and SPNet compared with the original versions, which proves the effectiveness of the Video8K set to potentially boost SOTA model' performance.

Further, we observed that the higher the performance of a model trained with the original training set, the less improvement it exhibits when retrained with the training set obtained using our method. Conversely, the lower the performance of a model trained with the original training set, the greater the improvement it demonstrates when retrained with the training set obtained using our method. This phenomenon can be attributed to the diminishing returns effect, where models that are already highly optimized benefit less from additional data augmentation due to their existing proficiency. On the other hand, models with lower initial performance have more room for improvement, and the introduction of our enhanced training data provides substantial benefits by addressing previously unmet needs or gaps in the data. This observation indicates that our method is particularly effective for models needing significant performance boosts, showcasing its value in improving underperforming models and making them more robust across various domains and scenarios. Qualitative results are illustrated in Fig. 9. We can find that both in specific scenes (subfigure A) and normal scenes (subfigure B), our method can generate more accurate saliency results than other SOTA methods.

For the second part, to certify the effectiveness of our approach in scenes with domain shift issues, we have conducted four scalability experiments based on the three targeted baseline models on three real-world testing sets BDD, KITTI and CityScapes [19], which encompass a variety of weather conditions, different times of the day, and urban traffic patterns. Apart from the abovementioned experiment (denoted by "+Video8K"), we added a selected BDD training set to newly re-train the targeted models (denoted by "+BDD-TR"), and re-trained the model by the large-scale set COME15K-TR (denoted by "+COME15K-TR"). As shown in Table III, SOTA models trained on BDD (without domain shift issue) perform the best on the real-world BDD, KITTI and CityScapes testing sets, while models trained with original data (denoted by "Original") perform the worst. Our augmented Video8K performing in the middle (better than

TABLE IV
QUANTITATIVE COMPARISON WITH SOTA REPRESENTATIVE DOMAIN ADAPTATION METHODS. THIS EXPERIMENT TAKES C^2DF AS THE TARGET MODEL. THE BEST RESULT IN EACH COLUMN IS BOLD

Datasets	STEREO				NLPR				ReDweb-S				BDD-TE				KITTI			
Metrics	Sm	Fm	Em	Mi	Sm	Fm	Em	Mi	Sm	Fm	Em	Mi	Sm	Fm	Em	Mi	Sm	Fm	Em	Mi
UNIQUE	.885	.877	.911	.046	.915	.885	.940	.029	.685	.889	.741	.135	.863	.764	.884	.029	.802	.750	.808	.036
ADV-DA	.892	.880	.918	.043	.918	.893	.945	.028	.697	.699	.746	.131	.870	.769	.893	.026	.806	.758	.815	.033
AsyFOD	.898	.886	.920	.041	.924	.897	.954	.025	.702	.706	.753	.128	.875	.772	.898	.022	.813	.762	.821	.029
UCBS	.906	.891	.929	.038	.927	.903	.957	.021	.708	.714	.758	.125	.884	.779	.900	.021	.818	.767	.824	.027
Ours	**.915**	**.899**	**.935**	**.034**	**.932**	**.911**	**.964**	**.019**	**.719**	**.723**	**.770**	**.122**	**.892**	**.786**	**.912**	**.017**	**.829**	**.774**	**.831**	**.023**

"+COME15K-TR") is that Video8K contains some similar scenes with the real-world BDD, KITTI and CityScapes, and has more diverse scenes than COME15K-TR, thus can slightly mitigate the domain shift problem and slightly outperforms "+COME15K-TR", but not perform well as no domain shift BDD training set. Qualitative results are illustrated in Fig. 10 (normal scenes *v.s.* specific scenes).

E. Comparisons With SOTA Representative Domain Adaptation Methods

Our method's performance was thoroughly compared against several state-of-the-art (SOTA) domain adaptation techniques, including UNIQUE [95], ADV-DA [96], AsyFOD [97], UCBS [98], across multiple datasets: STEREO, NLPR, ReDweb-S, BDD-TE, and KITTI. The results, as shown in Table IV, indicate that our method consistently outperforms existing domain adaptation techniques across all datasets and metrics. This is likely due to our method's ability to adapt well across different domains (e.g., stereo traffic scenes), which showcases its robustness and flexibility. By leveraging videos' spatiotemporal information, our method effectively generates additional high-quality training data, which enhances model performance even with domain shifts.

F. Ablation Studies

1) Different Optical Flow Methods: In fact, a more powerful optical flow tool could benefit overall performance. In this study, we have employed four representative optical flow approaches, including PWCNet [99], FlowNet 2.0 [100], SPyNet [101] and RAFT [64]. According to the quantitative results demonstrated in Table V-A, we chose RAFT in our method, where this approach outperformed other competitors in terms of all metrics, *e.g.*, the F-m metric has been improved from 0.907 (PWCNet) to 0.911 (RAFT) in NLPR set, showing the effectiveness of the optical flow maps generated by RAFT. Further, the performance results used by all these four optical flow tools are marginally different, which shows the robustness of our approach.

2) Threshold γ Adopted in QF: We have tested multiple choices regarding γ (Eq. 3), and the exact results can be found in Table V-B. As shown, the overall performance of our method is moderately sensitive to the choice of γ. Specifically, $\gamma = 0.85$ achieves the best result, and $\gamma = 0.95$ is inferior to $\gamma = 0.75$, showing a larger γ may not always bring performance gain. The main reason lies in: 1) when γ uses a large value, the amount of available training data will be small, where the model may be incompletely trained, and 2) when γ chooses a very small value, the derived training data could be redundant and will hinder the performance of the model.

3) Different Monocular Depth Estimation Methods: Though the quality of monocular estimated depth (MED) is inferior to our saliency informative depth (SID, Sec. III-B) regarding the RGB-D SOD task, the quality of MED is also a significant factor in the generation of SID (**PART1** of Fig. 3). We have conducted an extensive ablation study regarding five monocular depth estimation methods, *e.g.*, FastDepth [61], Monodepth2 [62], LapDepth [102], MiDaS [60] to figure out the monocular depth estimation method that contributes to our SID most, and the detailed quantitative results can be found in Table V-C. Compared with the other three methods, MiDaS achieved the best results in all metrics, *e.g.*, 0.723 (MiDaS) *vs.* 0.720 (the latest SOTA LapDepth) in terms of F-m in ReDweb-S set. Therefore, we have chosen MiDaS as the depth estimation method.

Further, to prove the superiority of our SID compared to MED towards the RGB-D SOD task, we have comprehensively compared two representative SOTA monocular depth estimation methods, *e.g.*, LapDepth, and MiDaS. Results in Table VI demonstrate that no matter on RGB-D SOD set ReDweb-S with normal scenes or stereo set KITTI with specific scenes, our SID can obtain the most superior saliency results compared to MED, *e.g.*, we have promoted the S-m of ReDweb-S and KITTI from 0.715→0.719 and 0.824→0.829, respectively, based on MiDaS.

4) Disscusion of Bringing in RGB SOD Datasets: During Pseudo-GT generating procedure (see in Fig. 7), we have utilized the pre-trained SOTA ISOD model, which is trained by RGB SOD datasets to produce color saliency and motion saliency. To verify the improvement of our approach is not simply brought about by RGB SOD datasets, we have conducted an extensive ablation study to prove it. Firstly, we only pre-train the RGB branch of RGB-D ISOD SOTA models, *e.g.*, C^2DF, SSL, and SPNet, by RGB SOD datasets and DAVIS, then load the pre-trained weights to train full RGB-D SOTA using RGB-D ISOD datasets (the same parameters as RGB-D ISOD SOTA models have set, denoted as "+PreRGB")). As is shown in Table VII, the C^2DF model trained by our newly augmented data without pre-trained RGB branch (denoted as "+Ours") outperforms the original models (denoted as "C^2DF") and the models trained by RGB-D ISOD datasets ("+PreRGB") with pre-trained RGB branch using RGB SOD datasets and DAVIS regarding all metrics on RGB-D sets NLPR, STEREO and ReDweb-S with normal scenes, and stereo sets BDD-TR and KITTI with specific scenes, *e.g.*, 0.592 ("C^2DF") *vs.* 0.705 ("+PreRGB") *vs.* 0.829 ("+Ours") in terms of the S-m metric in the KITTI set, which proves that the pre-training RGB branch can obtain performance gain than the original model but is inferior to our augmented datasets. Accordingly, it is reasonable to infer that RGB SOD datasets do not bring about the improvement of our approach.

5) Disscusion of Quality Filter: We also applied a quality filter (QF, Sec. III-C) during the pseudo-GT generation

TABLE V

A: COMPARISONS BETWEEN DIFFERENT OPTICAL FLOW METHODS; B: ABLATION STUDY REGRADING THRESHOLD γ (EQ. 3); C: COMPARISONS BETWEEN DIFFERENT MONOCULAR DEPTH ESTIMATION METHODS. ALL THESE EXPERIMENTS TAKE C^{2}DF AS THE TARGETED MODEL

	Datasets	STEREO				NLPR				ReDweb-S				BDD-TE				KITTI			
	Metrics	Sm	Fm	Em	M	Sm	Fm	Em	M	Sm	Fm	Em	M	Sm	Fm	Em	M	Sm	Fm	Em	M
A. Different Optical Flow Methods	PWCNet	.909	.893	.929	.037	.924	.907	.958	.023	.710	.716	.764	.129	.884	.798	.906	.021	.822	.770	.827	.020
	SPyNet	.910	.896	.930	.036	.925	.908	.959	.022	.715	.719	.767	.126	.885	.782	.909	.019	.825	.771	.828	.022
	FlowNet2	.914	.897	.932	**.034**	.929	.910	.961	.020	.717	.722	.768	.124	.888	.785	.910	.018	.826	.773	.829	.022
	RAFT	**.915**	**.899**	**.935**	**.034**	**.932**	**.911**	**.964**	**.019**	**.719**	**.723**	**.770**	**.122**	**.892**	**.786**	**.912**	**.017**	**.829**	**.774**	**.831**	**.023**
B. Different Choices of Threshold γ	γ=0.65	.907	.893	.930	.038	.928	.907	.960	.024	.711	.714	.761	.128	.884	.778	.906	.020	.820	.767	.821	.029
	γ=0.75	.909	.895	.932	.036	.929	.910	.963	.023	.714	.718	.764	.126	.887	.782	.907	.019	.823	.768	.825	.026
	γ=0.85	**.915**	**.899**	**.935**	**.034**	**.932**	**.911**	**.964**	**.019**	**.719**	**.723**	**.770**	**.122**	**.892**	**.786**	**.912**	**.017**	**.829**	**.774**	**.831**	**.023**
	γ=0.95	.910	.896	.931	.035	.931	.909	.961	.021	.718	.720	.766	.125	.891	.784	.909	.018	.828	.772	.828	.024
C. Different Choices of Monocular D Estimation Models	Monodepth2	.911	.895	.930	.037	.926	.905	.954	.023	.713	.718	.763	.128	.886	.781	.907	.014	.820	.769	.824	.026
	FastDepth	.912	.897	.932	.035	.929	.909	.955	.022	.716	.719	.765	.126	.888	.784	.910	.015	.824	.772	.828	.025
	LapDepth	.914	.898	.933	.035	.930	.910	.961	.020	.718	.720	.769	.125	.891	.785	.911	**.017**	.828	.773	.829	.024
	MiDaS	**.915**	**.899**	**.935**	**.034**	**.932**	**.911**	**.964**	**.019**	**.719**	**.723**	**.770**	**.122**	**.892**	**.786**	**.912**	**.017**	**.829**	**.774**	**.831**	**.023**

TABLE VI

QUANTITATIVE COMPARISONS BETWEEN THE MONOCULAR ESTIMATED DEPTH (MED) AND SALIENCY INFORMATIVE DEPTH (SID) TOWARDS THE RGB-D SOD TASK. THIS EXPERIMENT TAKES C^{2}DF AS THE TARGET MODEL. THE BEST RESULT IN EACH COLUMN IS BOLD

Datasets	ReDweb-S				KITTI			
Metrics	Sm	Fm	Em	M	Sm	Fm	Em	M
LapDepth(MED)	.713	.717	.761	.127	.823	.770	.822	.027
LapDepth (SID)	**.718**	**.720**	**.769**	**.125**	**.828**	**.773**	**.829**	**.024**
MiDaS (MED)	.715	.721	.774	.124	.824	.769	.825	.026
MiDaS (SID)	**.719**	**.723**	**.770**	**.122**	**.829**	**.774**	**.831**	**.023**

TABLE VII

QUANTITATIVE COMPARISON RESULTS OF TARGET RGB-D SOTA MODELS C^{2}DF, SSL AND SPNET (ORIGINAL MODEL), +PRERGB (FIRST PRE-TRAIN RGB BRANCH OF RGB-D SOTA MODELS WITH RGB SOD DATASETS, THEN TRAIN FULL RGB-D MODEL WITH RGB-D SOD DATASETS) AND +OURS (TRAINING RGB-D SOTA MODELS WITH OUR AUGMENTED DATASETS). THE BEST RESULT IN EACH COLUMN IS BOLD

Datasets	STEREO				NLPR				ReDweb-S				BDD-TE				KITTI			
Metrics	Sm	Fm	Em	M	Sm	Fm	Em	M	Sm	Fm	Em	M	Sm	Fm	Em	M	Sm	Fm	Em	M
SPNet	.899	.883	.924	.043	.926	.901	.954	.024	.709	.712	.759	.129	.673	.489	.790	.032	.594	.334	.706	.036
+PreRGB	.905	.887	.927	.040	.927	.904	.959	.023	.711	.713	.762	.128	.710	.598	.852	.026	.668	.551	.799	.033
+Ours	**.914**	**.895**	**.930**	**.038**	**.930**	**.906**	**.961**	**.021**	**.715**	**.719**	**.767**	**.124**	**.885**	**.776**	**.907**	**.019**	**.831**	**.765**	**.826**	**.025**
SSL	.886	.875	.919	.045	.909	.884	.939	.038	.710	.706	.754	.136	.688	.479	.755	.039	.615	.324	.648	.042
+PreRGB	.895	.879	.922	.043	.913	.885	.944	.034	.712	.708	.755	.134	.698	.556	.812	.035	.710	.534	.731	.036
+Ours	**.913**	**.893**	**.928**	**.040**	**.919**	**.887**	**.954**	**.027**	**.716**	**.712**	**.760**	**.131**	**.881**	**.753**	**.904**	**.023**	**.824**	**.718**	**.822**	**.028**
C^{2}DF	.905	.892	.927	.038	.914	.899	.955	.024	.715	.717	.762	.131	.667	.454	.741	.036	.592	.307	.662	.041
+PreRGB	.908	.893	.929	.037	.921	.903	.958	.022	.717	.720	.764	.128	.747	.590	.818	.031	.705	.582	.764	.034
+Ours	**.915**	**.899**	**.935**	**.034**	**.932**	**.911**	**.964**	**.019**	**.719**	**.723**	**.770**	**.122**	**.892**	**.786**	**.912**	**.017**	**.829**	**.774**	**.831**	**.023**

procedure (see Fig. 7), but not in the depth generation process (see Fig. 6). The reasons are as follows. Firstly, the quality filter is specifically designed to ensure the generation of high-quality pseudo-ground truths (pseudo-GTs) by retaining frames with high-quality optical flow maps and high consistency of color-motion saliency. This approach is crucial for pseudo-GTs because they directly impact the training efficacy of the RGB-D SOD models. However, when generating saliency informative depth maps, the primary goal is to extract layer-separable clues from the optical flow, rather than achieving precise depth layouts with rich details. The depth generation process is designed to be more robust to variations in the quality of optical flow maps because the optical flow primarily provides the relative motion information needed for saliency detection, rather than absolute depth accuracy.

By focusing on the essential motion information that distinguishes salient objects, the depth generation process can tolerate some degree of noise or lower quality in the optical flow maps. The subsequent fusion with RGB information in the OFDNet further mitigates the impact of any low-quality optical flow maps, ensuring that the generated depth maps remain effective for the saliency detection task.

6) Disscusion of Practical Implications: Our method offers significant practical benefits in the context of traffic scene analysis, particularly in improving the robustness of SOD models when applied to diverse traffic environments. By leveraging optical flow-based cues to generate saliency-informed depth, our approach enhances the ability of models to distinguish significant objects even in challenging dynamic traffic scenarios. This capability can directly contribute to improving safety and efficiency in traffic monitoring systems, such as autonomous vehicles, traffic surveillance, and smart city infrastructure. Additionally, the video-based data augmentation strategy we propose allows for the creation of training datasets that better reflect real-world variability, which is essential for the development of more adaptive and reliable SOD models.

G. Limitations and Error Analysis

The primary limitations of our approach include data quality issues and challenges in depth estimation. In terms of model performance, while increasing training data generally boosts accuracy, it eventually leads to diminishing returns due to redundancy and gaps between generated and testing datasets. Additionally, reliance on pseudo-ground truths can introduce noise, impacting performance. Real-world factors like lighting, weather, and sensor noise further affect video and depth

map quality, complicating model reliability. Finally, optical flow, a core component for depth estimation, struggles in low-motion scenes and under adverse conditions. Integrating semantic cues and temporal consistency could partially mitigate these issues, but achieving robust performance across diverse scenarios remains challenging.

H. Ethical Considerations

Regarding ethical and privacy considerations, we utilize publicly sourced or consented video data and have implemented anonymization processes to protect individual privacy. By ensuring that all data is either openly available or obtained with proper consent, and by removing any identifiable information, we address potential privacy concerns associated with the use of optical flow and saliency maps derived from real-world traffic videos. Consequently, there are no significant data security or ethical issues related to our dataset.

V. Conclusion

This paper has introduced a data augmentation solution that specifically addresses the domain shift problem in SOD tasks for traffic scenes. Our method uniquely leverages optical flow-based layer-separable cues to generate saliency-informed depth. It exploits the spatiotemporal complementary nature of video data to produce pseudo-GTs that enhance model training. While our approach shows promise in specific contexts, it is essential to note that results may vary depending on the characteristics of the target domain. By constructing a video-based training set with diverse scenes, we enable the re-training of off-the-shelf RGB-D SOD models, thereby providing a method to mitigate domain shift challenges.

Crucially, by enhancing the adaptability and robustness of SOD models in analyzing complex traffic scenes, our approach has the potential to contribute to improvements in traffic-related image analysis accuracy. However, further validation in varied environments is necessary. In future work, we plan to explore the integration of our method with 3D traffic simulation models to improve the representation of dynamic, real-world traffic scenarios. Additionally, hybrid SOD approaches that combine both traditional and deep learning-based techniques could be investigated to further enhance model performance. These extensions could broaden the applicability of our method and inspire future research in the field of traffic image analysis.

References

[1] Y. Sun, J. Dai, Z. Ren, Q. Li, and D. Peng, "Relaxed energy preserving hashing for image retrieval," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 7, pp. 7388–7400, Jul. 2024.

[2] S. Ren and Q. Liu, "Small target augmentation for urban remote sensing image real-time segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 2, pp. 2076–2088, Feb. 2024.

[3] J. Wang, G. Li, G. Qiu, G. Ma, J. Xi, and N. Yu, "Depth-assisted semi-supervised RGB-D rail surface defect inspection," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 7, pp. 8042–8052, Jul. 2024.

[4] T. Deng, H. Yan, L. Qin, T. Ngo, and B. S. Manjunath, "How do drivers allocate their potential attention? Driving fixation prediction via convolutional neural networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 5, pp. 2146–2154, May 2020.

[5] J. Li, Z. Qu, S.-Y. Wang, and S.-F. Xia, "YOLOX-RDD: A method of anchor-free road damage detection for front-view images," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 10, pp. 14725–14739, Oct. 2024.

[6] L. Zhang, M. Chen, B. Tu, Y. Li, and Y. Xia, "Retargeting HR aerial photos under contaminated labels with application in smart navigation," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 1, pp. 349–358, Jan. 2024.

[7] S. Dong, W. Zhou, C. Xu, and W. Yan, "EGFNet: Edge-aware guidance fusion network for RGB-thermal urban scene parsing," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 1, pp. 657–669, Jan. 2024.

[8] Y. Liu, L. Zhou, G. Wu, S. Xu, and J. Han, "TCGNet: Type-correlation guidance for salient object detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 7, pp. 6633–6644, Jul. 2024.

[9] L. Qin et al., "ID-YOLO: Real-time salient object detection based on the driver's fixation region," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 15898–15908, Sep. 2022.

[10] N. Jia, Y. Sun, and X. Liu, "TFGNet: Traffic salient object detection using a feature deep interaction and guidance fusion," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 3, pp. 3020–3030, Mar. 2024.

[11] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 1115–1119.

[12] K. Fu, J. He, and X. Yang, "Few-shot learning-based RGB-D salient object detection: A case study," *Neurocomputing*, vol. 512, pp. 142–152, Nov. 2022.

[13] K. Fu, D.-P. Fan, G.-P. Ji, Q. Zhao, J. Shen, and C. Zhu, "Siamese network for RGB-D salient object detection and beyond," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5541–5559, Sep. 2022.

[14] F. Yu et al., "BDD100K: A diverse driving dataset for heterogeneous multitask learning," 2018, *arXiv:1805.04687*.

[15] Y. Tsai, W. Hung, S. Schulter, K. Sohn, M. Yang, and M. Chandraker, "Learning to adapt structured output space for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7472–7481.

[16] Z. Deng, K. Zhou, D. Li, J. He, Y. Song, and T. Xiang, "Dynamic instance domain adaptation," *IEEE Trans. Image Process.*, vol. 31, pp. 4585–4597, 2022.

[17] J. Li et al., "Joint semantic mining for weakly supervised RGB-D salient object detection," in *Proc. NeurIPS*, vol. 34, Dec. 2021, pp. 11945–11959.

[18] Y. Piao, J. Wang, M. Zhang, and H. Lu, "MFNet: Multi-filter directive network for weakly supervised salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Oct. 2021, pp. 4136–4145.

[19] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[20] R. Cong, J. Lei, C. Zhang, Q. Huang, X. Cao, and C. Hou, "Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion," *IEEE Signal Process. Lett.*, vol. 23, no. 6, pp. 819–823, Jun. 2016.

[21] L. Wu, Z. Liu, H. Song, and O. Le Meur, "RGBD co-saliency detection via multiple kernel boosting and fusion," *Multimedia Tools Appl.*, vol. 77, no. 16, pp. 21185–21199, Aug. 2018.

[22] W. Zhou, F. Sun, and W. Qiu, "MSNet: Multiple strategy network with bidirectional fusion for detecting salient objects in RGB-D images," *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 4341–4353, 2025.

[23] W. Zhou, Y. Zhu, J. Lei, R. Yang, and L. Yu, "LSNet: Lightweight spatial boosting network for detecting salient objects in RGB-thermal images," *IEEE Trans. Image Process.*, vol. 32, pp. 1329–1340, 2023.

[24] W. Zhou, F. Sun, Q. Jiang, R. Cong, and J.-N. Hwang, "WaveNet: Wavelet network with knowledge distillation for RGB-T salient object detection," *IEEE Trans. Image Process.*, vol. 32, pp. 3027–3039, 2023.

[25] W. Zhou, Q. Guo, J. Lei, L. Yu, and J.-N. Hwang, "ECFFNet: Effective and consistent feature fusion network for RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1224–1235, Mar. 2022.

[26] W. Zhou, Y. Zhu, J. Lei, J. Wan, and L. Yu, "CCAFNet: Crossflow and cross-scale adaptive fusion network for detecting salient objects in RGB-D images," *IEEE Trans. Multimedia*, vol. 24, pp. 2192–2204, 2021.

[27] W. Zhou, Q. Guo, J. Lei, L. Yu, and J.-N. Hwang, "IRFR-net: Interactive recursive feature-reshaping network for detecting salient objects in RGB-D images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 3, pp. 4132–4144, Mar. 2021.

[28] C. Chen, J. Wei, C. Peng, and H. Qin, "Depth-quality-aware salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 2350–2363, 2021.

[29] Z. Chen, R. Cong, Q. Xu, and Q. Huang, "DPANet: Depth potentiality-aware gated attention network for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 7012–7024, 2021.

[30] Z. Liu, S. Shi, Q. Duan, W. Zhang, and P. Zhao, "Salient object detection for RGB-D image by single stream recurrent convolution neural network," *Neurocomputing*, vol. 363, pp. 46–57, Oct. 2019.

[31] J. Ren, X. Gong, L. Yu, W. Zhou, and M. Y. Yang, "Exploiting global priors for RGB-D saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 25–32.

[32] N. Wang and X. Gong, "Adaptive fusion for RGB-D salient object detection," *IEEE Access*, vol. 7, pp. 55277–55284, 2019.

[33] Y. Ding, Z. Liu, M. Huang, R. Shi, and X. Wang, "Depth-aware saliency detection using convolutional neural networks," *J. Vis. Commun. Image Represent.*, vol. 61, pp. 1–9, May 2019.

[34] Q. Zhang, Q. Qin, Y. Yang, Q. Jiao, and J. Han, "Feature calibrating and fusing network for RGB-D salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 3, pp. 1493–1507, Mar. 2024.

[35] W. Gao, G. Liao, S. Ma, G. Li, Y. Liang, and W. Lin, "Unified information fusion network for multi-modal RGB-D and RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2091–2106, Apr. 2022.

[36] Z. Liu, Y. Tan, Q. He, and Y. Xiao, "SwinNet: Swin transformer drives edge-aware RGB-D and RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4486–4497, Jul. 2022.

[37] Y. Yang, Q. Qin, Y. Luo, Y. Liu, Q. Zhang, and J. Han, "Bi-directional progressive guidance network for RGB-D salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5346–5360, Aug. 2022.

[38] G. Chen et al., "Modality-induced transfer-fusion network for RGB-D and RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 4, pp. 1787–1801, Apr. 2023.

[39] B. Tang, Z. Liu, Y. Tan, and Q. He, "HRTransNet: HRFormer-driven two-modality salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 2, pp. 728–742, Feb. 2023.

[40] N. Liu, N. Zhang, L. Shao, and J. Han, "Learning selective mutual attention and contrast for RGB-D saliency detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9026–9042, Dec. 2022.

[41] J. Zhang et al., "RGB-D saliency detection via cascaded mutual information minimization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4318–4327.

[42] Y. Liu, P. Wang, Y. Cao, Z. Liang, and R. W. H. Lau, "Weakly-supervised salient object detection with saliency bounding boxes," *IEEE Trans. Image Process.*, vol. 30, pp. 4423–4435, 2021.

[43] L. Wang et al., "Learning to detect salient objects with image-level supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3796–3805.

[44] Y. Xu, X. Yu, J. Zhang, L. Zhu, and D. Wang, "Weakly supervised RGB-D salient object detection with prediction consistency training and active scribble boosting," *IEEE Trans. Image Process.*, vol. 31, pp. 2148–2161, 2022.

[45] Y. Piao, W. Wu, M. Zhang, Y. Jiang, and H. Lu, "Noise-sensitive adversarial learning for weakly supervised salient object detection," *IEEE Trans. Multimedia*, vol. 25, pp. 2888–2897, 2022.

[46] X. Wang et al., "Boosting RGB-D saliency detection by leveraging unlabeled RGB images," *IEEE Trans. Image Process.*, vol. 31, pp. 1107–1119, 2022.

[47] J. Kopf, X. Rong, and J. Huang, "Robust consistent video depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1611–1621.

[48] H. Xu et al., "Unifying flow, stereo and depth estimation," *IEEE TPAMI*, vol. 45, no. 11, pp. 13941–13958, Nov. 2023.

[49] T. Shimada, H. Nishikawa, X. Kong, and H. Tomiyama, "Fast and high-quality monocular depth estimation with optical flow for autonomous drones," *Drones*, vol. 7, no. 2, p. 134, Feb. 2023.

[50] X. Guo, H. Zhao, S. Shao, X. Li, and B. Zhang, "F2Depth: Self-supervised indoor monocular depth estimation via optical flow consistency and feature map synthesis," *Eng. Appl. Artif. Intell.*, vol. 133, Jul. 2024, Art. no. 108391.

[51] V. Guizilini, K.-H. Lee, R. Ambrus, and A. Gaidon, "Learning optical flow, depth, and scene flow without real-world labels," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 3491–3498, Apr. 2022.

[52] Z. Lu and Y. Chen, "Joint self-supervised depth and optical flow estimation towards dynamic objects," *Neural Process. Lett.*, vol. 55, no. 8, pp. 10235–10249, Dec. 2023.

[53] V. S. Vibashan, P. Oza, and V. M. Patel, "Towards online domain adaptive object detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 478–488.

[54] K. Wang, X. Fu, Y. Huang, C. Cao, G. Shi, and Z.-J. Zha, "Generalized UAV object detection via frequency domain disentanglement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1064–1073.

[55] M. B. Colomer et al., "To adapt or not to adapt? Real-time adaptation for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 16502–16513.

[56] A. Lopez-Rodriguez and K. Mikolajczyk, "DESC: Domain adaptation for depth estimation via semantic consistency," *Int. J. Comput. Vis.*, vol. 131, no. 3, pp. 752–771, Mar. 2023.

[57] B. Cui, G. Hu, and S. Yu, "DeepCollaboration: Collaborative generative and discriminative models for class incremental learning," in *Proc. AAAI*, vol. 35, May 2021, pp. 1175–1183.

[58] S. Dong, X. Hong, X. Tao, X. Chang, X. Wei, and Y. Gong, "Few-shot class-incremental learning via relation knowledge distillation," in *Proc. AAAI*, vol. 35, May 2021, pp. 1255–1263.

[59] J.-Y. Kim and D.-W. Choi, "Split-and-bridge: Adaptable class incremental learning within a single neural network," in *Proc. AAAI*, vol. 35, May 2021, pp. 8137–8145.

[60] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1623–1637, Mar. 2022.

[61] D. Wofk, F. Ma, T.-J. Yang, S. Karaman, and V. Sze, "FastDepth: Fast monocular depth estimation on embedded systems," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 6101–6108.

[62] C. Godard, O. M. Aodha, M. Firman, and G. Brostow, "Digging into self-supervised monocular depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3827–3837.

[63] P. Krhenbühl and V. Koltun, "Efficient inference in fully connected crfs with Gaussian edge potentials," in *Proc. NeurIPS*, 2011, pp. 109–117.

[64] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *Proc. Eur. Conf. Comput. Vis.* Glasgow, U.K.: Springer, Aug. 2020, pp. 402–419.

[65] Y.-H. Wu, Y. Liu, L. Zhang, M.-M. Cheng, and B. Ren, "EDN: Salient object detection via extremely-downsampled network," *IEEE Trans. Image Process.*, vol. 31, pp. 3125–3136, 2022.

[66] D. Fan, M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," *IJCV*, vol. 129, pp. 2622–2638, Jun. 2021.

[67] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 724–732.

[68] K. Wang, Z. Tu, C. Li, C. Zhang, and B. Luo, "Learning adaptive fusion bank for multi-modal salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 7344–7358, Aug. 2024.

[69] T. Zhou, H. Fu, G. Chen, Y. Zhou, D.-P. Fan, and L. Shao, "Specificity-preserving RGB-D saliency detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4661–4671.

[70] X. Zhao, Y. Pang, L. Zhang, H. Lu, and X. Ruan, "Self-supervised pretraining for RGB-D salient object detection," in *Proc. AAAI*, vol. 36, Jun. 2022, pp. 3463–3471.

[71] M. Zhang, S. Yao, B. Hu, Y. Piao, and W. Ji, "C^2DFnet: Criss-cross dynamic filter network for RGB-D salient object detection," *IEEE TMM*, vol. 25, pp. 5142–5154, 2022.

[72] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7253–7262.

[73] H. Peng, L. Bing, W. Xiong, W. Hu, and R. Ji, "Rgbd salient object detection: A benchmark and algorithms," in *Proc. ECCV*, 2014, pp. 92–109.

[74] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, "Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 5, pp. 2075–2089, May 2021.

[75] G. Li and C. Zhu, "A three-pathway psychobiological framework of salient object detection using stereoscopic technology," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 3008–3014.

[76] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2806–2813.

[77] Y. Niu, Y. Geng, X. Li, and F. Liu, "Leveraging stereopsis for saliency analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 454–461.

[78] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2556–2563.

[79] K. Soomro, A. Roshan Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*.

[80] L. Huang, X. Zhao, and K. Huang, "GOT-10k: A large high-diversity benchmark for generic object tracking in the wild," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1562–1577, May 2021.

[81] M. Kristan et al., "The eighth visual object tracking vot2020 challenge results," in *Proc. ECCVW*, 2020, pp. 547–601.

[82] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1597–1604.

[83] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 698–704.

[84] N. Liu, N. Zhang, and J. Han, "Learning selective self-mutual attention for RGB-D saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13753–13762.

[85] Y. Piao, Z. Rong, M. Zhang, W. Ren, and H. Lu, "A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9057–9066.

[86] J. Zhang et al., "Uncertainty inspired RGB-D saliency detection," *IEEE Trans. Pattern Analysis Mach. Intell.*, vol. 44, no. 9, pp. 5761–5779, Sep. 2021.

[87] G. Li, Z. Liu, and H. Ling, "ICNet: Information conversion network for RGB-D based salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 4873–4884, 2020.

[88] W. Ji et al., "Calibrated RGB-D salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9466–9476.

[89] Z. Zhang, Z. Lin, J. Xu, W.-D. Jin, S.-P. Lu, and D.-P. Fan, "Bilateral attention network for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 1949–1961, 2021.

[90] R. Cong et al., "CIR-Net: Cross-modality interaction and refinement for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 31, pp. 6800–6815, 2022.

[91] X. Wang, S. Li, C. Chen, A. Hao, and H. Qin, "Modality profile—A new critical aspect to be considered when generating RGB-D salient object detection training set," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 3355–3364.

[92] A. Li, Y. Mao, J. Zhang, and Y. Dai, "Mutual information regularization for weakly-supervised RGB-D salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 1, pp. 397–410, Jan. 2024.

[93] G. Chen et al., "EM-Trans: Edge-aware multimodal transformer for RGB-D salient object detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 2, pp. 3175–3188, Feb. 2025.

[94] Z. Feng, W. Wang, W. Li, G. Li, M. Li, and M. Zhou, "MFUR-net: Multimodal feature fusion and unimodal feature refinement for RGB-D salient object detection," *KBS*, vol. 299, no. 5, Sep. 2024, Art. no. 112022.

[95] W. Zhang, K. Ma, G. Zhai, and X. Yang, "Uncertainty-aware blind image quality assessment in the laboratory and wild," *IEEE Trans. Image Process.*, vol. 30, pp. 3474–3486, 2021.

[96] S. Song et al., "Multi-spectral salient object detection by adversarial domain adaptation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 12023–12030.

[97] Y. Gao, K.-Y. Lin, J. Yan, Y. Wang, and W.-S. Zheng, "AsyFOD: An asymmetric adaptation paradigm for few-shot domain adaptive object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 3261–3271.

[98] Y. Zhang and C. Wu, "Unsupervised camouflaged object segmentation as domain adaptation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2023, pp. 4336–4346.

[99] 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.

[100] 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. 1647–1655.

[101] A. Ranjan and M. J. Black, "Optical flow estimation using a spatial pyramid network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2720–2729.

[102] M. Song, S. Lim, and W. Kim, "Monocular depth estimation using Laplacian pyramid-based depth residuals," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 11, pp. 4381–4393, Nov. 2021.

Chenglizhao Chen received the Ph.D. degree from Beihang University in 2017. He is currently a Professor with the College of Computer Science and Technology, China University of Petroleum (East China). His research interests include virtual reality, computer vision, deep learning, and pattern recognition.

Mengke Song is currently pursuing the Ph.D. degree with the College of Computer Science and Technology and Qingdao Institute of Software, China University of Petroleum (East China). His research interests include computer vision and deep learning,

Shanchen Pang is currently a Professor with the College of Computer Science and Technology, China University of Petroleum (East China). His research interests include software formalization, edge computing, artificial intelligence, and computer vision.

Chong Peng is currently a Professor with the College of Computer Science and Technology, Ocean University of China. His research interests include machine learning, data mining, and computer vision.