2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Transparent Objects: A Corner Case in Stereo Matching

Zhiyuan Wu, Shuai Su, Qijun Chen, and Rui Fan✉

***Abstract*— Stereo matching is a common technique used in 3D perception, but transparent objects such as reflective and penetrable glass pose a challenge as their disparities are often estimated inaccurately. In this paper, we propose transparency-aware stereo (TA-Stereo), an effective solution to tackle this issue. TA-Stereo first utilizes a semantic segmentation or salient object detection network to identify transparent objects, and then homogenizes them to enable stereo matching algorithms to handle them as non-transparent objects. To validate the effectiveness of our proposed TA-Stereo strategy, we collect 260 images containing transparent objects from the KITTI Stereo 2012 and 2015 datasets and manually label pixel-level ground truth. We evaluate our strategy with six deep stereo networks and two types of transparent object detection methods. Our experiments demonstrate that TA-Stereo significantly improves the disparity accuracy of transparent objects. Our project webpage can be accessed at `mias.group/TA-Stereo`.**

I. INTRODUCTION

3D perception is a critical and foundational aspect of autonomous driving [1], [2], and stereo matching plays a vital role in this process [3]–[5]. Recent research has achieved impressive stereo matching results by leveraging state-of-the-art (SoTA) deep convolutional neural networks (DCNNs) [6], [7]. The task of estimating disparities on transparent objects is a corner case that has not been given much attention. As shown in Fig. 1, when stereo matching is formulated as a dense correspondence matching problem in SoTA DCNNs, it can incorrectly estimate disparities on transparent object pixels. Such inaccuracies can significantly impact the performance of 3D perception functionality, underscoring the need for more robust and accurate stereo matching algorithms that can handle transparent objects.

Therefore, this paper presents a transparency-aware stereo matching (TA-Stereo) strategy that significantly improves the accuracy of disparities on transparent objects. To achieve this, we first detect transparent objects using either a semantic segmentation or salient object detection network, and then adaptively homogenize the detected regions to facilitate stereo matching performance. To evaluate our approach, we create a transparent object detection dataset including 260 RGB images and their pixel-level ground truth based on the KITTI Stereo 2012 and 2015 [10], [11] datasets, and conduct extensive experiments with six SoTA deep stereo networks, five semantic segmentation networks, and two salient object detection networks. Our results validate the effectiveness of the TA-Stereo strategy and demonstrate its potential for improving 3D perception in intelligent vehicles. Furthermore, we discuss the limitations of our method and suggest possible solutions for further improvement.

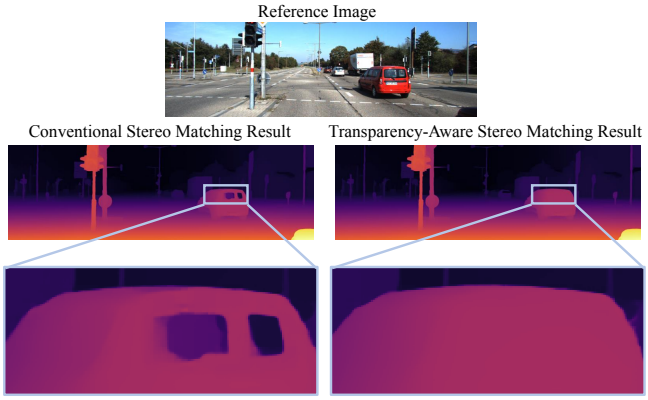

Fig. 1. Stereo matching without and with transparency awareness ability. The disparities are estimated using RAFT-Stereo [8] pre-trained on the SceneFlow [9] dataset, where the disparities of transparent objects are determined based on the light transmission, which can differ from the disparities of the surrounding objects as observed from the camera.

II. RELATED WORK

A. Stereo Matching

Explicit programming-based stereo matching algorithms have four phases: cost computation, cost aggregation, disparity optimization, and disparity refinement [12]. Local algorithms select a group of image blocks from the target image and match them with an image block selected from the reference image [13], [14]. On the other hand, global algorithms treat stereo matching as an energy minimization problem, which can be tackled by Markov random field (MRF)-based optimization approaches [15]. Semi-global matching (SGM) [16] balances stereo matching accuracy and efficiency by performing cost aggregation along all directions in the image [17]. However, explicit programming-based methods are either inaccurate (local algorithms) or computationally intensive (global algorithms) [18].

With the advancements in DCNNs, the accuracy of disparity estimation has been greatly improved. Researchers have turned their focus towards developing end-to-end approaches

This work was supported by the National Key R&D Program of China under Grant 2020AAA0108100, the National Natural Science Foundation of China under Grant 62233013, the Science and Technology Commission of Shanghai Municipal under Grant 22511104500, the Fundamental Research Funds for the Central Universities under Grants 22120220184 and 22120220214, and the Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0100. (*Zhiyuan Wu and Shuai Su contributed equally to this work.*) (*Corresponding author: Rui Fan.*)

The authors are with the Machine Intelligence & Autonomous Systems (MIAS) Group, the Robotics & Artificial Intelligence Laboratory (RAIL), the College of Electronic & Information Engineering, the State Key Laboratory of Intelligent Autonomous Systems, and the Frontiers Science Center for Intelligent Autonomous Systems, Tongji University, Shanghai 201804, P. R. China. (E-mails: {2050285, sushuai, qjchen, rfan}@tongji.edu.cn

2023 IEEE International Conference on Robotics and Automation (ICRA) | 979-8-3503-2365-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICRA48891.2023.10161385

to learn dense stereo matching. Chang *et al.* [19] introduced PSMNet, a pyramid stereo matching network that leverages both spatial pyramid pooling and 3D convolutional layers. GwcNet [20] constructs the cost volume using group-wise correlation to further improve the 3D stacked hourglass network. GA-Net [21] employs a two-layer guided aggregation block to capture local and global cost dependencies. As 3D convolutions are computationally expensive, researchers have focused on enhancing both the efficiency and accuracy of stereo matching. Cheng *et al.* [22] applied neural architecture search (NAS) to stereo matching and presented the first hierarchical NAS framework for end-to-end deep stereo matching. RAFT-Stereo [8] presents a rectified stereo matching method that builds upon the optical flow estimation network RAFT [23] and demonstrates real-time performance. In an effort to increase the DCNN inference speed, researchers have also adopted coarse-to-fine paradigms, replacing 3D convolutions. Li *et al.* [24] proposed CRE-Stereo, a hierarchical network with recurrent refinement that produces more accurate disparity results. However, when these deep stereo networks are trained on datasets with incorrect labels for transparent object disparities, they often perform poorly. Regarding transparent object disparity estimation, Tsin *et al.* [25] proposed a nested-plane-sweeping method and graph cut optimization to estimate the depths of two layers, which provides valuable insights for further research.

B. Semantic Segmentation and Salient Object Detection

Semantic segmentation algorithms can be broadly categorized into two groups: single-modal and data-fusion. Single-modal algorithms, such as SegNet [26], U-Net [27], PSPNet [28], and DeepLab series [29], [30], employ end-to-end approaches to perform pixel-wise classification. SegNet introduces the encoder-decoder architecture and a pixel-wise classification layer, while U-Net improves upon it by adding skip connections to better preserve the spatial information of small objects. PSPNet, on the other hand, leverages a pyramid pooling module to gather contextual information for improved semantic segmentation performance. DeepLabv3 leverages atrous convolution and depth-wise separable convolution in both its atrous spatial pyramid pooling (ASPP) and decoder modules to achieve improved efficiency and accuracy in semantic segmentation. The use of atrous convolution allows the network to capture multi-scale contextual information while reducing computational complexity, leading to improved efficiency and accuracy in semantic image segmentation. Data-fusion algorithms have outperformed single-modal networks by incorporating multiple types of visual information. Fan *et al.* [31] proposed SNE-RoadSeg, a data-fusion DCNN that combines features from both RGB images and surface normal maps to achieve improved driving scene segmentation. Our paper provides additional experimental results to show the efficacy of SNE-RoadSeg when only one encoder of the network is utilized for semantic segmentation. Specifically, Mei *et al.* introduced GDNet [32] as a solution for glass segmentation, utilizing a large convolutional feature interpolation (LCFI) module to acquire abundant contextual cues. However, our experimental results showed that it falls short in detecting transparent objects in driving scenarios.

Salient object detection methods aim to identify the most prominent object(s) in an image. Deep learning-based salient object detection approaches consider both bottom-up and top-down saliency inferences, and attention mechanisms have recently been incorporated. For instance, PiCANet [33] is a representative prior work that employs attention mechanisms to learn both global and local contexts. Additionally, RAS [34] uses foreground and background attention maps to assist in detecting salient objects and removing non-salient ones. This paper compares the performance of semantic segmentation and salient object detection in terms of transparent object detection.

III. Transparency-Aware Stereo Matching

This section introduces our proposed TA-Stereo strategy, which greatly enhances the accuracy of disparity estimation for transparent objects. TA-Stereo aims to eliminate the disparity inconsistencies caused by transparent objects (which can be detected using either semantic segmentation or salient object detection networks) and ensure that a well-developed stereo matching algorithm can treat them as non-transparent objects. The primary objective of TA-Stereo is to homogenize the transparent objects, allowing the stereo matching algorithm to produce more accurate and consistent disparity results.

Transparent objects can be modeled as a combination of transmission and reflection scenarios. Both these scenarios can be regarded as a linear superposition at a fixed position behind the transparent object, which we call a mapping scenario. Humans can accurately identify transparent objects without the need to observe every pixel of them, because we regard them as usual opaque objects according to their geometric priors [35], [36]. Drawing inspiration from this, we homogenize the transparent objects according to the surrounding opaque objects, enabling stereo matching networks to make use of surrounding features for disparity estimation, resulting in more accurate disparity estimation results.

To achieve greater homogenization of transparent objects, we utilize their planar characteristics and transform them into uniform regions with similar features to the surrounding areas. While a direct homogenization can be accomplished by simply masking transparent objects with a single color, we explore a more effective strategy, referred to as Adaptive Homogenization. This strategy involves adapting the homogenization process to the characteristics of the connected domain of a single transparent object, which can be achieved through the use of conventional image processing algorithms. The adaptive transparent object homogenization process can be expressed as follows:

$$I(\boldsymbol{p}) = \frac{1}{k} \sum_{\boldsymbol{q} \in \boldsymbol{Q_p}} I(\boldsymbol{q}), \tag{1}$$

where I is the RGB image, $\boldsymbol{p}$ is a pixel within the detected transparent object, and $\boldsymbol{Q_p} \in (\boldsymbol{q}_1; \boldsymbol{q}_2; \cdots, \boldsymbol{q}_k)$ is a set storing

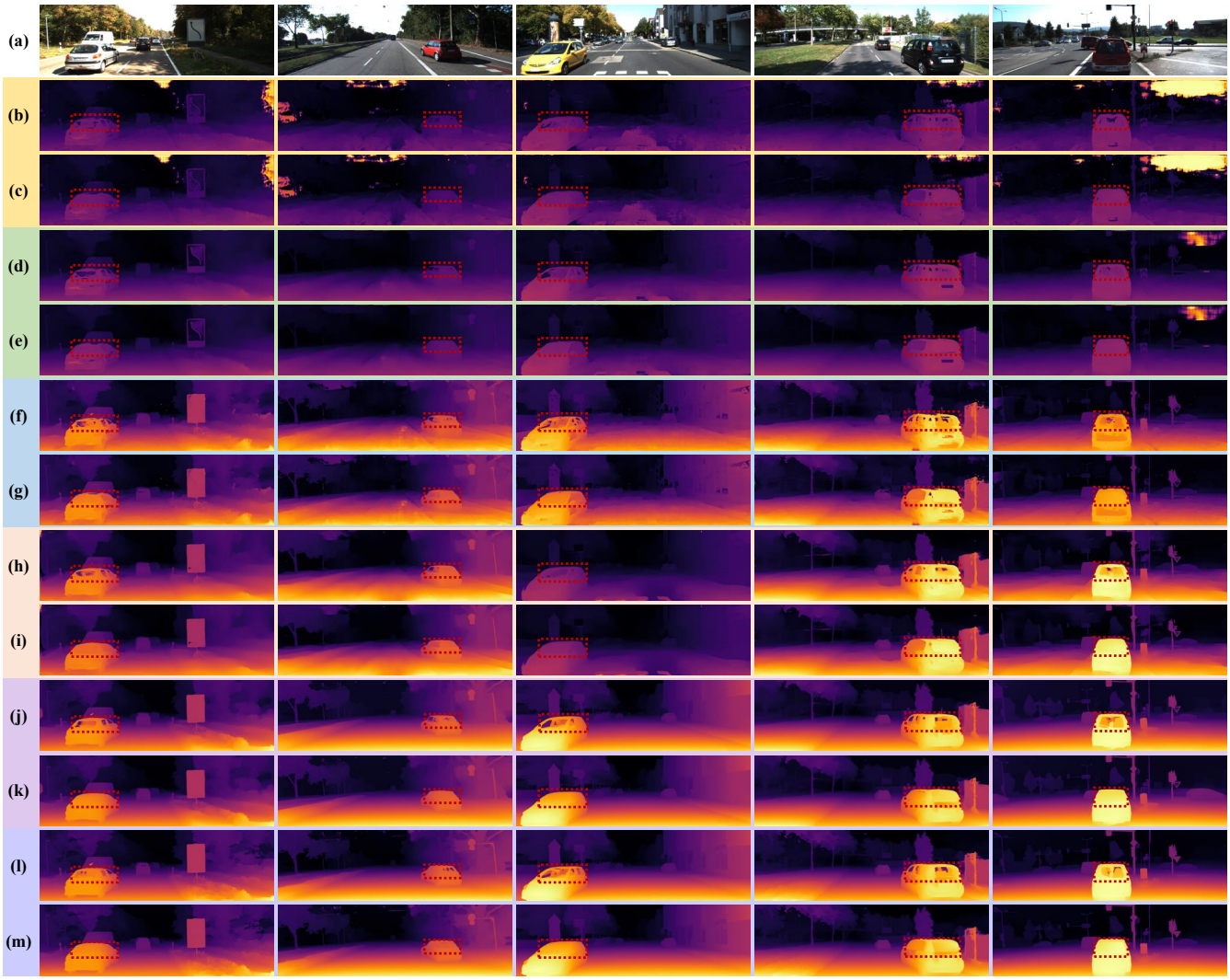

Fig. 2. Experiment results of transparency-aware stereo matching: (a) left images; (b), (d), (f), (h), (j), (l) disparity maps respectively estimated using pre-trained PSMNet [19], GwcNet [20], GA-Net [21], LEA-Stereo [22], RAFT-Stereo [8], and CRE-Stereo [24] without TA-Stereo; (c), (e), (g), (i), (k), (m) disparity maps respectively estimated using pre-trained PSMNet [19], GwcNet [20], GA-Net [21], LEA-Stereo [22], RAFT-Stereo [8], and CRE-Stereo [24] with TA-Stereo. The deep stereo networks are pre-trained on the SceneFlow [9] dataset, where the disparity estimation for transparent objects is incorrectly learned. The regions that exhibit significant improvements are marked with red dashed boxes.

the pixels on the boundary of the detected transparent object. Through adaptive homogenization, the interference caused by mapping scenarios can be eliminated. This enables the stereo matching algorithms to treat transparent objects as if they were non-transparent.

IV. Experiments

A. Datasets and Evaluation Metrics

In synthetic datasets, such as SceneFlow [9], the disparities of transparent objects (*e.g.*, car windows) are determined based on the light transmission, which can differ from the disparities of the surrounding objects as observed from the camera. This disparity discrepancy can lead to errors in depth estimation and can affect the performance of deep stereo networks trained on such datasets, which rely on accurate disparity calculation for 3D geometry reconstruction. In contrast, real-world datasets, such as the KITTI Stereo 2015 [11] dataset, provide more realistic ground-truth disparities for transparent objects, as they are manually labeled based on the projections of 3D LiDAR points. Training deep stereo networks on these datasets can enable them to more accurately estimate the disparities of transparent objects. Therefore, to validate the effectiveness of our proposed TA-Stereo strategy, we created a transparent object detection dataset based on the KITTI Stereo 2012 and 2015 [10], [11] datasets. Our dataset contains a training set of 160 RGB images (selected from the KITTI Stereo 2012 [10] dataset) with our manually labeled pixel-level transparent object ground truth, and a test set of 50 pairs of stereo images (selected from the KITTI Stereo 2015 [11] dataset) and our manually labeled pixel-level transparent object ground truth. Our dataset is publicly available at `mias.group/TA-Stereo`.

In our experiments, we first evaluate the performance of six SoTA deep stereo networks: PSMNet [19], GwcNet [20],

TABLE I

COMPARISON OF SOTA DEEP STEREO NETWORKS, PRE-TRAINED ON THE SCENEFLOW [9] DATASET AND EVALUATED WITH GROUND-TRUTH TRANSPARENT OBJECT LABELS. THE BEST RESULTS ARE SHOWN IN BOLD TYPE.

Networks	EPE ↓ (pixels)			PEP 1.0 ↓ (%)			PEP 3.0 ↓ (%)		
	A	T	N	A	T	N	A	T	N
PSMNet [19]	3.60	4.42	**3.52**	62.81	71.59	**62.12**	28.76	35.72	**28.07**
PSMNet-**TA**	**3.54**	**2.48**	3.58	**62.40**	**61.75**	62.20	**28.30**	**22.68**	28.35
GwcNet [20]	2.04	4.89	**1.85**	48.01	77.72	46.33	13.60	37.77	**11.96**
GwcNet-**TA**	**1.99**	**4.66**	1.87	**46.88**	**60.90**	**45.98**	**13.01**	**29.56**	12.01
GA-Net [21]	**2.01**	**5.90**	**1.76**	44.78	74.93	43.09	12.56	**41.04**	10.76
GA-Net-**TA**	2.04	7.09	1.78	**44.18**	**69.76**	**42.74**	**12.27**	41.89	**10.62**
LEA-Stereo [22]	1.83	4.12	1.66	44.31	71.81	42.69	11.32	33.74	9.81
LEA-Stereo-**TA**	**1.70**	**2.39**	**1.66**	**43.04**	**56.29**	**42.19**	**10.50**	**21.44**	**9.69**
RAFT-Stereo [8]	1.34	4.91	1.11	27.57	72.35	25.18	6.81	39.01	4.82
RAFT-Stereo-**TA**	**1.25**	**3.16**	**1.09**	**26.03**	**42.29**	**24.86**	**5.36**	**16.21**	**4.62**
CRE-Stereo [24]	1.19	3.97	**1.01**	24.45	59.40	**22.44**	5.36	27.60	**3.98**
CRE-Stereo-**TA**	**1.08**	**1.94**	1.02	**23.99**	**46.87**	22.49	**4.76**	**14.71**	4.05

GA-Net [21], LEA-Stereo [22], RAFT-Stereo [8], and CRE-Stereo [24], pre-trained respectively on the SceneFlow [9] and the KITTI Stereo 2012 [10] datasets on our created transparent object detection dataset. We evaluated their performance on three regions of the dataset: all pixels (A), transparent object pixels only (T), and non-transparent object pixels only (N). We used two evaluation metrics: (1) the average endpoint error (EPE), and (2) the percentage of error pixels (PEP), with the threshold set to 1.0 and 3.0 pixels (hereafter referred to as PEP 1.0 and PEP 3.0), respectively.

We further evaluate the performance of five SoTA semantic segmentation networks: SegNet [26], U-Net [27], PSPNet [28], DeepLabv3+ [30], and SNE-RoadSeg [31] (with only one encoder), as well as two salient object detection networks: PiCANet [33] and RAS [34] for transparent object detection on the same dataset. To avoid any duplication between the test and training datasets, we select images that are not used in the previous experiments to train these networks. We utilize five widely used evaluation metrics: accuracy, precision, recall, F-score, and intersection over union (IoU), to quantify the performance of these networks.

B. Evaluation of Deep Stereo Networks with Transparency Awareness

Table I and Fig. 2 demonstrate the effectiveness of our TA-Stereo strategy in improving the performance of deep stereo networks for disparity estimation of transparent object pixels. With the exception of GA-Net, our approach significantly reduces the PEP and EPE for all pixels and transparent object pixels. We have also observed that applying our proposed TA-Stereo strategy to deep stereo matching results in more accurate disparities for non-transparent object pixels in most cases. Notably, our TA-Stereo strategy brings remarkable improvement in CRE-Stereo and RAFT-Stereo. When compared to their respective conventional pre-trained models, the CRE-Stereo-TA model reduces the EPE by 51.1% for transparent object pixels and 9.2% for all pixels. The PEP is also reduced by 1.9% and 11.2% (for thresholds of 1 and 3 pixels, respectively) for all pixels, and by 21.1% and 46.7% (for thresholds of 1 and 3 pixels, respectively) for transparent object pixels. Similarly, the RAFT-Stereo-TA model reduces the EPE by 45.3% for transparent object pixels and by 7.2% for all pixels compared to its conventional pre-trained model. Additionally, it reduces the PEP by 1.5% (for both thresholds of 1 and 3 pixels) for all pixels and by 30.1% and 22.8% (for thresholds of 1 and 3 pixels, respectively) for transparent object pixels. These results suggest that the RAFT-Stereo-TA and CRE-Stereo-TA models are more effective than the conventional pre-trained models for transparency-aware stereo matching.

Our analysis indicates that transparent objects, such as car windows, tend to exhibit a uniform texture after homogenization. This can make it difficult for stereo matching algorithms to accurately estimate disparities. However, the SoTA deep stereo networks are generally trained with large receptive fields, which enables them to accurately estimate disparities in such areas despite the lack of texture information. It is important to note that the performance of GA-Net is negatively affected when using the TA-Stereo strategy. This may be due to the fact that GA-Net incorporates a so-called local guided aggregation layer to capture local cost dependencies, which prioritizes local visual features and leads to inadequate disparity estimation in areas with low texture.

TABLE II

COMPARISON OF SEMANTIC SEGMENTATION AND SALIENT OBJECT DETECTION NETWORKS FOR TRANSPARENT OBJECT DETECTION, WHERE ONLY THE ENCODER TO EXTRACT FEATURES FROM RGB IMAGES IS USED IN SNE-ROADSEG. THE BEST RESULTS ARE SHOWN IN BOLD TYPE.

Networks	Accuracy (%)	Precision (%)	Recall (%)	F-score (%)	IoU (%)
SegNet [26]	98.5	61.2	63.7	58.5	43.9
U-Net [27]	95.4	43.2	30.2	28.4	17.8
PSPNet [28]	98.7	52.0	72.6	57.4	42.1
DeepLabv3+ [30]	98.2	28.5	74.4	35.5	23.5
SNE-RoadSeg [31]	99.0	71.9	76.2	71.6	56.9
PiCANet [33]	99.3	69.9	**90.3**	78.0	65.0
RAS [34]	**99.5**	**84.2**	86.6	**85.0**	**74.2**

C. Evaluation of Semantic Segmentation and Salient Object Detection Networks for Transparent Object Detection

To evaluate the performance of different networks in detecting transparent objects, we trained five semantic segmentation networks and two salient object detection networks for 200 epochs using their official configurations on an NVIDIA RTX 3090 GPU. As shown in Table II and Fig. 3, the single-modal semantic segmentation networks performed poorly, achieving a maximum IoU of less than 44%. The results also suggest that our previously proposed data-fusion network SNE-RoadSeg exhibits greater robustness and achieves higher F-scores and IoUs, which increase by 13.0-43.2% and

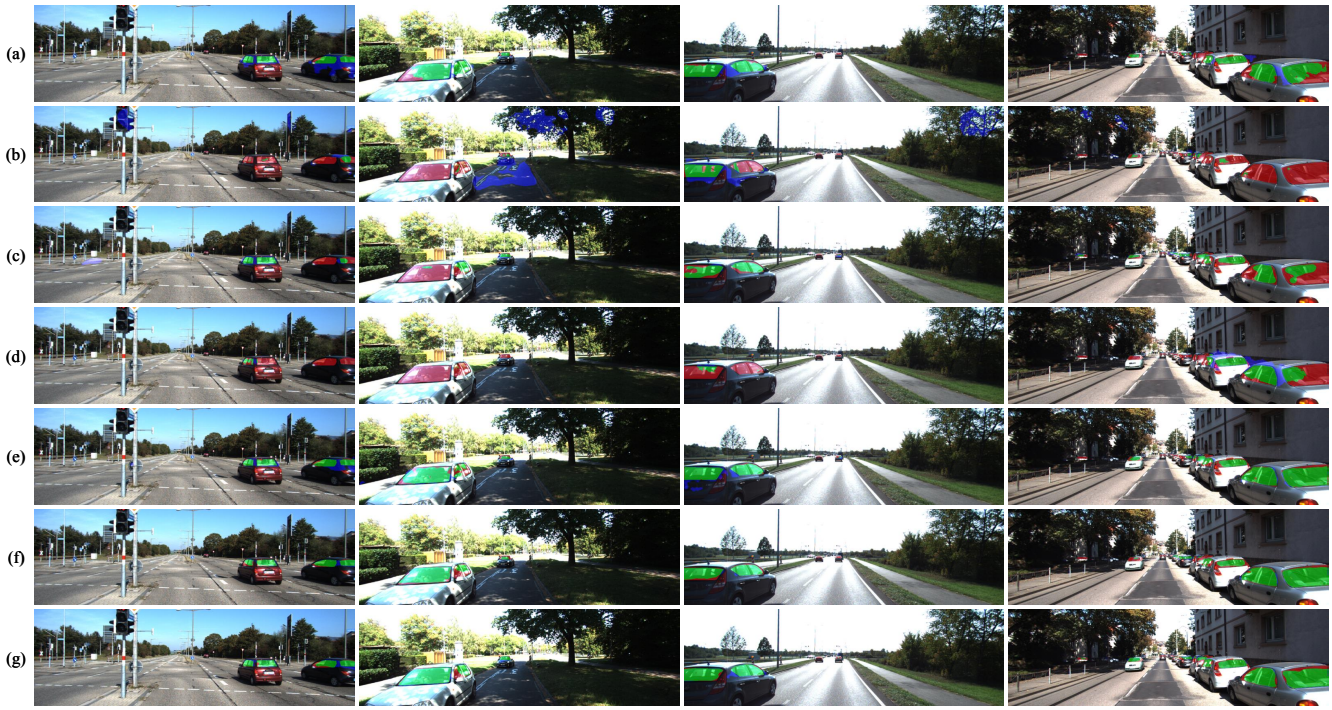

Fig. 3. Comparison of semantic segmentation and salient object detection networks for transparent object detection: (a)-(e) transparent object detection results achieved by SegNet [26], U-Net [27], PSPNet [28], DeepLabv3+ [30], and SNE-RoadSeg (only one encoder is used), respectively; (f)-(g) transparent object detection results achieved by PiCANet [33] and RAS [34]. The true-positive, false-negative, and false-positive pixels are shown in green, red, and blue, respectively. These results demonstrate the superior performance of salient object detection networks over semantic segmentation networks for transparent object detection.

13.0-39.1%, respectively, even though only a single encoder is used. The improved performance of SNE-RoadSeg is possibly attributed to the densely-connected skip connections employed in its decoder.

In addition, our findings suggest that salient object detection networks outperform semantic segmentation networks for transparent object detection. Specifically, RAS achieved the highest accuracy, precision, F-score, and IoU, while PiCANet achieved the highest recall. This is presumably due to salient object detection networks focusing on identifying one prominent object class in an image, often incorporating (bottom-up and top-down) attention mechanisms. In contrast, semantic segmentation networks apply convolutional operations to the entire image without any explicit attention mechanism, which may not be optimal for identifying transparent objects. When a transparent object is difficult to distinguish from its background, semantic segmentation networks tend to perform poorly.

D. Evaluation of Deep Stereo Networks Assisted with Salient Object Detection Network

We also utilize RAS [34], the best-performing transparent object detection method, in combination with pre-trained deep stereo networks to further evaluate the efficacy of our proposed TA-Stereo strategy. Table III presents our findings, which indicate that almost all the evaluated networks achieve higher EPEs for transparent object pixels when the TA-Stereo strategy is employed. However, only GwcNet, LEA-Stereo, and RAFT-Stereo exhibit significant improvement.

TABLE III

COMPARISON OF SOTA DEEP STEREO NETWORKS, PRE-TRAINED ON THE SCENEFLOW [9] DATASET AND EVALUATED WITH THE TRANSPARENT OBJECT DETECTION RESULTS ACHIEVED USING RAS [34]. THE BEST RESULTS ARE SHOWN IN BOLD TYPE.

Networks	EPE ↓ (pixels)			PEP 1.0 ↓ (%)			PEP 3.0 ↓ (%)		
	A	T	N	A	T	N	A	T	N
PSMNet [19]	3.60	4.42	**3.52**	**62.81**	**71.59**	**62.12**	**28.76**	**35.72**	**28.07**
PSMNet-**TA**	**3.59**	**3.41**	3.60	63.09	73.10	62.50	29.10	36.13	28.62
GwcNet [20]	2.04	4.89	**1.85**	48.01	77.72	46.33	13.60	37.77	**11.96**
GwcNet-**TA**	**2.02**	**4.13**	1.89	**47.26**	**65.47**	**46.13**	**13.52**	**32.31**	12.29
GA-Net [21]	**2.01**	**5.90**	**1.76**	**44.78**	74.93	**43.09**	**12.56**	**41.04**	**10.76**
GA-Net-**TA**	2.16	6.84	1.86	44.80	**71.75**	43.23	12.88	43.63	11.17
LEA-Stereo [22]	1.83	4.12	**1.66**	44.31	71.81	**42.69**	11.32	33.74	**9.81**
LEA-Stereo-**TA**	**1.75**	**2.82**	1.69	**43.92**	**63.78**	42.69	**11.10**	**26.53**	10.15
RAFT-Stereo [8]	1.34	4.91	**1.11**	27.57	72.35	25.18	6.81	39.01	**4.82**
RAFT-Stereo-**TA**	**1.28**	**3.29**	1.12	**26.54**	**52.31**	**25.06**	**5.79**	**20.49**	4.91
CRE-Stereo [24]	**1.19**	3.97	**1.01**	**24.45**	59.40	**22.44**	**5.36**	**27.60**	**3.98**
CRE-Stereo-**TA**	1.19	**3.31**	1.05	24.71	**57.38**	22.74	5.79	29.59	4.31

For RAFT-Stereo, the EPE is reduced by 33.0% for transparent object pixels and 4.5% for all pixels, and the PEP is reduced by 18.5%-20.0% for transparent object pixels and approximately 1.0% for all pixels. We speculate that transparent objects tend to be relatively small in size, and any inaccuracies in their detection can have a considerable impact on the overall effectiveness of the TA-Stereo strategy.

TABLE IV

COMPARISON OF SOTA DEEP STEREO NETWORKS, PRE-TRAINED ON THE KITTI STEREO 2012 [10] DATASET AND EVALUATED WITH THE TRANSPARENT OBJECT GROUND TRUTH. THE BEST RESULTS ARE SHOWN IN BOLD TYPE.

Networks	EPE ↓ (pixels)			PEP 1.0 ↓ (%)			PEP 3.0 ↓ (%)		
	A	T	N	A	T	N	A	T	N
PSMNet [19]	**1.35**	**1.18**	**1.35**	**55.99**	**42.83**	**56.45**	**3.66**	**4.72**	**3.52**
PSMNet-**TA**	1.39	1.61	1.38	57.17	59.78	57.15	4.17	10.11	3.80
GwcNet [20]	**0.86**	**0.94**	**0.86**	**22.36**	**28.45**	**21.64**	**2.56**	**3.34**	**2.47**
GwcNet-**TA**	0.89	1.08	0.87	22.73	33.95	21.77	2.94	6.62	2.59
GA-Net [21]	**0.54**	**0.48**	**0.54**	**10.95**	**8.75**	**10.90**	**1.13**	**0.66**	**1.15**
GA-Net-**TA**	0.61	0.89	0.59	13.28	26.37	12.22	1.72	4.27	1.47
LEA-Stereo [22]	**0.80**	**0.84**	**0.80**	**21.06**	**28.79**	**20.48**	**2.13**	**1.33**	**2.17**
LEA-Stereo-**TA**	0.82	0.99	0.81	21.88	32.93	20.91	2.49	4.56	2.30

TABLE V

EVALUATION OF TA-STEREO WITH SGM [6]. THE BEST RESULTS ARE SHOWN IN BOLD TYPE.

Methods	EPE ↓ (pixels)			PEP 1.0 ↓ (%)			PEP 3.0 ↓ (%)		
	A	T	N	A	T	N	A	T	N
SGM	**7.26**	**13.28**	**6.75**	37.26	65.07	35.53	**19.66**	37.64	**18.45**
SGM-**TA**	7.49	15.88	6.84	**36.95**	**59.02**	**35.49**	19.78	**37.61**	18.62

E. Additional Evaluation of Our TA-Stereo Strategy

As discussed in Sec. IV-A, real-world datasets, including the KITTI Stereo 2012 and 2015 [10], [11] datasets, provide more realistic ground truth for transparent objects. Therefore, training deep stereo networks on these datasets can improve their ability to estimate the disparities of transparent object pixels. Table IV demonstrates that the performance of TA-Stereo is compromised when deep stereo networks are pre-trained on the KITTI Stereo 2012 dataset and subsequently tested on the KITTI Stereo 2015 dataset. This outcome can be attributed to the fact that the training dataset provides accurate ground-truth disparities for transparent objects, thereby allowing the networks to acquire transparency-aware capabilities.

Furthermore, we employ SGM to evaluate the effectiveness of our proposed TA-Stereo strategy. The results presented in Table V demonstrate that SGM outperforms SGM-TA in terms of achieving higher EPEs for all pixels, transparent object pixels, and non-transparent object pixels. Although there is a slight improvement in the PEP 1.0 achieved by SGM-TA, its overall performance is worse than SGM. Our analysis indicates that the reason for this is probably due to the fact that conventional stereo matching algorithms rely on explicit programming to determine disparities, wherein matching costs are calculated using a mathematical equation, such as the sum of absolute differences or the normalized cross-correlation. As a result, these algorithms fail to handle homogenized transparent objects (texture-less regions).

V. DISCUSSION

Estimating disparities for transparent objects in stereo matching remains a difficult task due to the multiple reflections and transmissions that can occur within the object [25]. When light interacts with a transparent object, it is not only reflected but also transmitted through the surface [37]. Conventional stereo matching algorithms lack the transparency awareness ability, and as a result, they may estimate the disparities of transparent objects based on light transmission, which may differ from the disparities of surrounding objects as observed from the camera. This discrepancy may have safety implications for autonomous vehicles. Therefore, it is crucial to develop transparency-aware stereo matching algorithms to ensure safe and reliable autonomous driving.

However, our proposed TA-Stereo strategy has some limitations that should be addressed. First, it seems to be effective only for stereo matching approaches with a strong global context inference ability. Local context-based approaches may struggle to handle homogenized transparent objects. As a result, alternative strategies to process transparent object pixels can potentially enhance the performance of stereo matching algorithms that focus on local context. Second, the efficacy of our TA-Stereo strategy heavily relies on the accuracy of transparent object detection. It is worth noting that salient object detection networks tend to perform better than semantic segmentation networks in terms of transparent object detection, as these objects are typically small in size. Therefore, it is crucial to explore and incorporate small salient object detection algorithms for improved stereo matching performance. This aspect deserves more attention in future research. Third, this work treats transparent object detection and stereo matching as separate tasks. However, we believe that a multi-task learning framework that combines these two tasks can lead to even more promising results. By jointly optimizing both tasks, the model can leverage the inter-dependencies between them to achieve improved performance in both transparent object detection and stereo matching. Such a framework represents an interesting avenue for future research.

VI. CONCLUSION

This paper discussed the issue of transparent objects in stereo matching, which poses unique challenges due to the presence of multiple reflections and transmissions inside the object. To address this issue, we proposed the TA-Stereo strategy, which leverages a semantic segmentation or salient object detection network to detect transparent objects and then homogenizes them to facilitate stereo matching. Our experimental results demonstrate that the incorporation of the TA-Stereo strategy significantly enhances the performance of global context-based stereo matching algorithms. Finally, we also discussed the limitations of our approach and suggested possible solutions for further improvement.

REFERENCES

[1] Z. Liu, S. Zhou, C. Suo, P. Yin, W. Chen, H. Wang, H. Li, and Y.-H. Liu, "LPD-Net: 3D point cloud learning for large-scale place recognition and environment analysis," in *Proceedings of the IEEE/CVF*

International Conference on Computer Vision (ICCV), 2019, pp. 2831–2840.

[2] H. Wang, R. Fan, P. Cai, M. Liu, and L. Wang, "UnDAF: A general unsupervised domain adaptation framework for disparity or optical flow estimation," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 01–07.

[3] N. Qian, "Binocular disparity and the perception of depth," *Neuron*, vol. 18, no. 3, pp. 359–368, 1997.

[4] R. Fan, U. Ozgunalp, B. Hosking, M. Liu, and I. Pitas, "Pothole detection based on disparity transformation and road surface modeling," *IEEE Transactions on Image Processing*, vol. 29, pp. 897–908, 2019.

[5] R. Fan and M. Liu, "Road damage detection based on unsupervised disparity map segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 11, pp. 4906–4911, 2020.

[6] H. Hirschmuller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2. IEEE, 2005, pp. 807–814.

[7] J. Zbontar and Y. LeCun, "Computing the stereo matching cost with a convolutional neural network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1592–1599.

[8] L. Lipson, Z. Teed, and J. Deng, "RAFT-Stereo: Multilevel recurrent field transforms for stereo matching," in *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021, pp. 218–227.

[9] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4040–4048.

[10] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 3354–3361.

[11] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3061–3070.

[12] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, no. 1, pp. 7–42, 2002.

[13] R. Fan, X. Ai, and N. Dahnoun, "Road surface 3D reconstruction based on dense subpixel disparity map estimation," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 3025–3035, 2018.

[14] H. Hirschmuller and D. Scharstein, "Evaluation of stereo matching costs on images with radiometric differences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 9, pp. 1582–1599, 2008.

[15] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001.

[16] Y. Lee, M.-G. Park, Y. Hwang, Y. Shin, and C.-M. Kyung, "Memory-efficient parametric semiglobal matching," *IEEE Signal Processing Letters*, vol. 25, no. 2, pp. 194–198, 2017.

[17] R. Fan, U. Ozgunalp, Y. Wang, M. Liu, and I. Pitas, "Rethinking road surface 3-D reconstruction and pothole detection: From perspective transformation to disparity map segmentation," *IEEE Transactions on Cybernetics*, vol. 52, no. 7, pp. 5799–5808, 2022.

[18] H. Wang, R. Fan, P. Cai, and M. Liu, "PVStereo: Pyramid voting module for end-to-end self-supervised stereo matching," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4353–4360, 2021.

[19] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5410–5418.

[20] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3273–3282.

[21] F. Zhang, V. Prisacariu, R. Yang, and P. H. Torr, "GA-Net: Guided aggregation net for end-to-end stereo matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 185–194.

[22] X. Cheng, Y. Zhong, M. Harandi, Y. Dai, X. Chang, H. Li, T. Drummond, and Z. Ge, "Hierarchical neural architecture search for deep stereo matching," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 22 158–22 169, 2020.

[23] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 402–419.

[24] J. Li, P. Wang, P. Xiong, T. Cai, Z. Yan, L. Yang, J. Liu, H. Fan, and S. Liu, "Practical stereo matching via cascaded recurrent network with adaptive correlation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 16 263–16 272.

[25] Y. Tsin, S. B. Kang, and R. Szeliski, "Stereo matching with reflections and translucency," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1. IEEE, 2003, pp. I–I.

[26] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 2015, pp. 234–241.

[28] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2881–2890.

[29] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[30] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European Conference on computer vision (ECCV)*, 2018, pp. 801–818.

[31] R. Fan, H. Wang, P. Cai, and M. Liu, "SNE-RoadSeg: Incorporating surface normal information into semantic segmentation for accurate freespace detection," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 340–356.

[32] H. Mei, X. Yang, Y. Wang, Y. Liu, S. He, Q. Zhang, X. Wei, and R. W. Lau, "Don't hit me! glass detection in real-world scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3687–3696.

[33] N. Liu, J. Han, and M.-H. Yang, "PiCANet: Learning pixel-wise contextual attention for saliency detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3089–3098.

[34] S. Chen, X. Tan, B. Wang, H. Lu, X. Hu, and Y. Fu, "Reverse attention-based residual network for salient object detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 3763–3776, 2020.

[35] Y.-C. Guo, D. Kang, L. Bao, Y. He, and S.-H. Zhang, "NeRFReN: Neural radiance fields with reflections," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 18 409–18 418.

[36] J. Hou, S. Xie, B. Graham, A. Dai, and M. Nießner, "Pri3D: Can 3D priors help 2D representation learning?" in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 5693–5702.

[37] Z. Rao, M. He, Y. Dai, Z. Zhu, B. Li, and R. He, "NLCA-Net: a non-local context attention network for stereo matching," *APSIPA Transactions on Signal and Information Processing*, vol. 9, 2020.