

Beyond RGB-D: Quasi-Inverse Perspective Transformation Enables High-Precision Waste Segmentation in Noisy Construction Sites

Xinxing Chen^{1 2},

¹ School of Mechanical and Electrical Engineering, Shenzhen Polytechnic University, Shenzhen, Guangdong 518055, China;

² Research Associate, Department of Electronic & Computer Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong
xchenak@szpu.edu.cn

Abstract. As a massive amount of waste generated by global construction activities, robots capable of automatically recycling construction waste have become effective tools for conserving natural resources. However, the complex environments of construction sites and the high diversity of waste materials pose challenges for robotic inspection, target recognition, and grasping. Achieving precise waste recognition is a prerequisite for stable grasping operations. In construction sites, complicated lighting conditions and cluttered backgrounds result in low feature contrast of waste materials in traditional RGB images, severely hindering recognition accuracy. To eliminate illumination interference, this study employs a depth camera to acquire scene disparity information and develops a Quasi-Inverse Perspective Transformation (QIPT) module, which enhances geometric contrast between objects and background by generating transformed disparity (TD) maps. Experimental validation within a single-modal fusion framework based on UNet demonstrates that the TD modality significantly outperforms RGB and raw disparity modalities in segmentation performance, achieving an improvement of 12.8% in the mIoU metric.

Keywords: construction waste; semantic segmentation; transformed disparity.

1. Introduction

Construction and Demolition (C&D) waste refers to surplus or damaged products and materials generated from construction, renovation, and demolition activities [1], typically constituting the largest proportion of total waste output. However, current practices often transport mixed C&D waste directly to landfills without prior sorting [2, 3]. This wasteful management approach causes air, water, and soil pollution, exerts tremendous pressure on limited landfill capacity, and has triggered multiple serious accidents. Given that C&D waste can be repurposed as raw materials, its recycling is widely recognized as an effective strategy for conserving natural resources. On-site sorting technology—which processes waste directly at construction sites—can significantly reduce transportation/storage costs and pollution risks [4]. Therefore, novel automation technologies are urgently needed to address on-site C&D waste sorting challenges.

Driven by the demand for efficient on-site recycling, our previous project developed a highly robust robotic prototype. This prototype employs advanced SLAM positioning technology and a high-precision 3D object grasping strategy to enable autonomous C&D waste recycling at construction sites. Nevertheless, persistent recognition issues caused by dynamic site conditions and waste surface contamination continue to pose formidable challenges for waste categorization and grasping. Consequently, multimodal data beyond RGB has emerged as a new research direction. While disparity images (traditional depth representations) are less susceptible to ambient lighting than RGB, they remain significantly affected by background noise.

In this context, we utilized the quasi inverse perspective transformation (QIPT) [5] to calculate the transformed disparity map (TD), which highlights the differences between the

background and construction wastes. Experimental results show that the utilization of TD leads to significantly improved performance in contrast to using RGB or disparity images.

To rigorously evaluate the architectural robustness of TD representation, we conduct comprehensive ablation studies across diverse neural network paradigms. The selection of benchmark networks is driven by three critical design dimensions: (1) feature decoding mechanisms (e.g., skip-connections vs. pooling indices), (2) context aggregation strategies (e.g., pyramid pooling vs. atrous convolution), and (3) boundary optimization approaches. This systematic variation enables probing TD's compatibility with fundamental segmentation philosophies.

In this section, we present an overview of the state-of-the-art single-modal networks on semantic segmentation, which aims to label each pixel in an image into individual classes. The Fully Convolutional Network (FCN) [6] pioneered the replacement of fully-connected layers with convolutional layers to achieve end-to-end segmentation, yet its upsampling process (e.g., $32\times$ transposed convolution) suffers from detail loss due to pooling operations. U-Net [8] significantly enhances small-object segmentation accuracy by incorporating skip connections that fuse shallow-layer features. SegNet [7] innovatively employs pooling indices to enable precise decoder upsampling while reducing memory consumption. However, these approaches lack comprehensive global scene understanding. PSPNet [9] first introduced pyramid pooling to capture multi-scale context. Concurrently, the DeepLab series [10] proposed atrous spatial pyramid pooling (ASPP), which captures multi-scale features through parallel multi-rate dilated convolutions, but suffers from inadequate feature continuity due to discrete sampling.

2. The Proposed Method

2.1 QIPT Method

The objective of TD [14] is to perform a conversion on a disparity image, rendering it in a perspective akin to the bird's-eye view. TD mitigates planar background noise — referring to unwanted signals from flat background regions (e.g., floors) — through 3D view transformation. In this transformation, the pixels corresponding to the ground plane exhibit comparable values, in stark contrast to the pixels depicting prominent objects on the road, which display substantial deviations. Stereo cameras possess the capability to extract stereo information, which forms the fundamental basis for producing TD. Referring to [14], if the stereo rig baseline is perfectly parallel to the ground surface (as shown in Fig. 1), $\mathbf{p}^W = [x, y, z]^T$ in the world coordinate system (WCS) can be transformed into an image pixel $\mathbf{p}_{l,r}^I = [u, v]^T$ in the left and right stereo images using:

$$\begin{bmatrix} \mathbf{p}_{l,r}^I \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & o_x & 0 \\ 0 & f & o_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & \pm \frac{b_s}{2} \\ 0 & \cos\varphi & -\sin\varphi & 0 \\ 0 & \sin\varphi & \cos\varphi & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{p}^W \\ 1 \end{bmatrix}, \quad (1)$$

where f is the focal length of the stereo camera; $\mathbf{O} = [o_x, o_y]^T$ is the principal point in pixels; φ is the stereo rig pitch angle; and b_s is the stereo rig baseline. The ground surface can be regarded as a plane of form:

$$z^w = \alpha y^w + \beta. \quad (2)$$

Plugging Eq.(2) into Eq.(1) obtains:

$$d = \frac{b_s}{\beta} (f(\cos\alpha - \alpha\sin\varphi) + (o_y - v)(\sin\varphi + \alpha\cos\varphi)), \quad (3)$$

where d denotes the disparity. Eq.(3) shows that the disparities of the ground plane are projected as a straight line in the v -disparity image (Labayrade & Aubert, 2003):

$$g(\mathbf{p}) = a_0 + a_1 v, \quad (4)$$

where $\mathbf{a} = [a_0, a_1]^T$ stores the coefficients of the straight line, and $\mathbf{p} = [u, v]^T$ is a pixel in the disparity image. $\mathbf{a}$ can be estimated by minimizing the following energy:

$$E = \mathbf{d} - \mathbf{V}\mathbf{a}_2^2, \quad (5)$$

where $\mathbf{d} = [d_1, d_2, \dots, d_n]^T$ stores the disparity values; $\mathbf{V} = [1_n, \mathbf{v}]$, where 1_n represents a $n \times 1$ vector of ones and $\mathbf{v} = [v_1, v_2, \dots, v_n]^T$. The minimization of Eq.(5) has a closed-form solution:

$$\mathbf{a} = (\mathbf{V}^T \mathbf{V})^{-1} \mathbf{V}^T \mathbf{d}, \quad (6)$$

Also, combining Eq.(6) with Eq.(5) obtains the minimum energy:

$$E_{\min} = \mathbf{d}^T \mathbf{d} - \mathbf{d}^T \mathbf{V} (\mathbf{V}^T \mathbf{V})^{-1} \mathbf{V}^T \mathbf{d}. \quad (7)$$

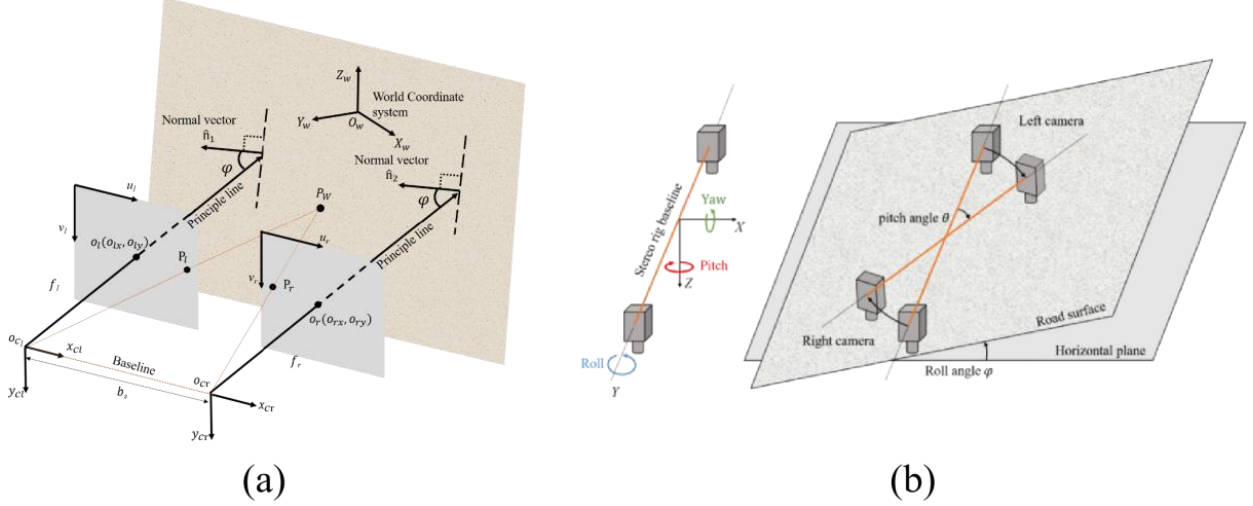


Fig. 1. (a) The coordinate system of a stereo camera; (b) The camera rotation angles θ and φ in two dimensions.

By plugging Eq.(6) into Eq.(4), we can obtain the representation of the ground plane disparity projections. However, in practice, the stereo rig baseline is not perfectly parallel to the ground plane, and therefore, introducing a non-zero roll angle θ into the disparity imaging process, as shown in Fig. 2. The introduced roll angle leads to a gradual disparity change in the horizontal direction, making the representation of ground plane disparity projections using Eq.(4) infeasible [15]. Furthermore, compared to the case that the roll angle is zero, $E_{\min}$ becomes much higher [15]. Therefore, the stereo rig roll angle has to be considered when minimizing Eq.(5).

The stereo rig roll angle θ can be introduced into the energy minimization process by simply rotating the disparity map by θ degrees around the image center. Each original image point $\mathbf{p} = [u, v]^T$ is, thus, transformed into a new image point $\mathbf{q} = [s, t]^T$:

$$\mathbf{q}(\theta, \mathbf{p}) = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \mathbf{p}. \quad (8)$$

Therefore, Eq.(7) has a more general expression:

$$E_{\min}(\theta) = \mathbf{d}^T \mathbf{d} - \mathbf{d}^T \mathbf{T}(\theta) (\mathbf{T}(\theta)^T \mathbf{T}(\theta))^{-1} \mathbf{T}(\theta)^T \mathbf{d}. \quad (9)$$

Minimizing Eq.(9) is equivalent to maximizing:

$$\begin{aligned} h(\theta) &= \mathbf{d}^T \mathbf{T}(\theta) (\mathbf{T}(\theta)^T \mathbf{T}(\theta))^{-1} \mathbf{T}(\theta)^T \mathbf{d} \\ &= \frac{p_3 + p_4 \cos 2\theta + p_5 \sin 2\theta}{p_0 \sin 2\theta + p_1 + p_2 \cos 2\theta}, \end{aligned} \quad (10)$$

where

$$p_0 = \mathbf{v}^T \mathbf{1}_n \mathbf{u}^T \mathbf{1}_n - n \mathbf{v}^T \mathbf{u}, \quad (11)$$

$$p_1 = \frac{1}{2} [n(\mathbf{v}^T \mathbf{v} + \mathbf{u}^T \mathbf{u}) - (\mathbf{v}^T \mathbf{1}_n)^2 - (\mathbf{u}^T \mathbf{1}_n)^2], \quad (12)$$

$$p_2 = \frac{1}{2} [n(\mathbf{v}^T \mathbf{v} - \mathbf{u}^T \mathbf{u}) - (\mathbf{v}^T \mathbf{1}_n)^2 + (\mathbf{u}^T \mathbf{1}_n)^2], \quad (13)$$

$$p_3 = \frac{1}{2} [(\mathbf{d}^T \mathbf{1}_n)^2 (\mathbf{v}^T \mathbf{v} + \mathbf{u}^T \mathbf{u}) + n((\mathbf{d}^T \mathbf{v})^2 + (\mathbf{d}^T \mathbf{u})^2)] - \mathbf{d}^T \mathbf{1}_n (\mathbf{v}^T \mathbf{1}_n \mathbf{d}^T \mathbf{v} + \mathbf{u}^T \mathbf{1}_n \mathbf{d}^T \mathbf{u}), \quad (14)$$

$$p_4 = \frac{1}{2} [(\mathbf{d}^T \mathbf{1}_n)^2 (\mathbf{v}^T \mathbf{v} - \mathbf{u}^T \mathbf{u}) + n((\mathbf{d}^T \mathbf{v})^2 - (\mathbf{d}^T \mathbf{u})^2)] - \mathbf{d}^T \mathbf{1}_n (\mathbf{v}^T \mathbf{1}_n \mathbf{d}^T \mathbf{v} - \mathbf{u}^T \mathbf{1}_n \mathbf{d}^T \mathbf{u}), \quad (15)$$

$$p_5 = \mathbf{d}^T \mathbf{1}_n (\mathbf{v}^T \mathbf{1}_n \mathbf{d}^T \mathbf{u} + \mathbf{u}^T \mathbf{1}_n \mathbf{d}^T \mathbf{v}) - (\mathbf{d}^T \mathbf{1}_n)^2 \mathbf{v}^T \mathbf{u} - n \mathbf{d}^T \mathbf{v} \mathbf{d}^T \mathbf{u}, \quad (16)$$

and $\mathbf{u} = [u_1, u_2, \dots, u_n]^T$. θ can be obtained by differentiating $h(\theta)$ with respect to θ :

$$\frac{\partial h(\theta)}{\partial \theta} = \frac{-2}{(p_1 + p_2 \cos 2\theta + p_0 \sin 2\theta)^2} ((p_4 p_0 - p_5 p_2) + (p_3 p_0 - p_5 p_1) \cos 2\theta + (p_4 p_1 - p_3 p_2) \sin 2\theta). \quad (17)$$

If the denominator of Eq.(17) does not equal zero, θ has the following expression:

$$\theta = \operatorname{argmax}[h(\arctan(\frac{p_4 p_1 - p_3 p_2 + \sqrt{\Delta}}{p_3 p_0 + p_5 p_2 - p_5 p_1 - p_4 p_0})), h(\arctan(\frac{p_4 p_1 - p_3 p_2 - \sqrt{\Delta}}{p_3 p_0 + p_5 p_2 - p_5 p_1 - p_4 p_0}))], \quad (18)$$

where

$$\Delta = (p_4 p_1 - p_3 p_2)^2 + (p_3 p_0 - p_5 p_1)^2 - (p_4 p_0 - p_5 p_2)^2 \quad (19)$$

We can obtain a better representation of ground plane disparity projections by plugging Eq.(18) into Eq.(6) and Eq.(8). Therefore, we can transform the original disparity image into a quasi inverse perspective view to make the disparities of the ground plane possess similar values:

$$\mathbf{D}_1 = \mathbf{D}_0 + g(\mathbf{p}, \theta) + \delta, \quad (20)$$

where $\mathbf{D}_0$ is the original disparity image, $\mathbf{D}_1$ is the transformed disparity image from the quasi inverse perspective view, and δ is a constant to ensure the values in $\mathbf{D}_1$ are all non-negative.

Fig. 2 illustrates the comparison between inputs and outputs of the QIPT method. The input consists of RGB color images, while the output comprises TD. The disparity maps before and after transformation clearly demonstrate enhanced visibility of target objects in the transformed images.

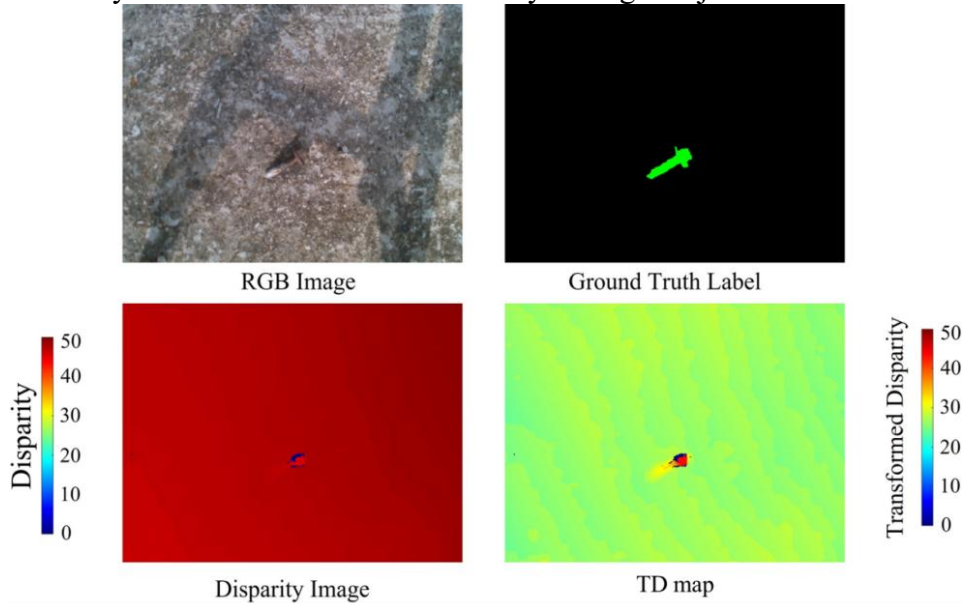


Fig. 2. Comparison of RGB image and TD before and after QIPT transformation.

3. Experimental Results and Discussions

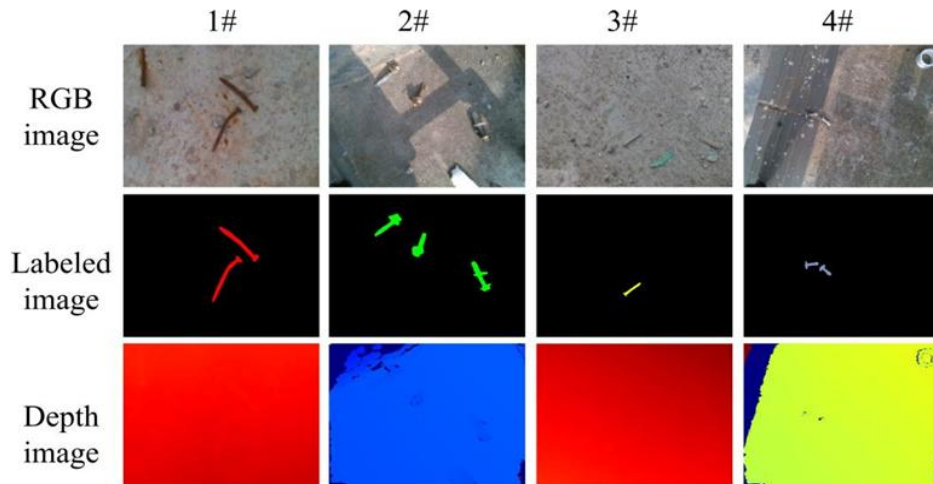


Fig. 3 Sample images in our dataset

This study will validate the proposed methodology using commonly observed construction waste objects. We captured 398 paired images (640×480 pixels) via an RGB-D camera, with each pair comprising one RGB image and one depth image. The dataset encompasses four categories of typical construction wastes, namely, iron nails (Nail #1), hexagon pan head screws (Nail #2), countersunk head self-tapping screws (Nail #3), and big flat head self-drilling screws (Nail #4), as shown in Fig. 3.

3.1 Ablation Study

To evaluate the effectiveness of TD compared to other data modalities such as RGB images and disparity images in enhancing neural network prediction accuracy, we conducted an evaluation by feeding these three modalities into the same network architecture for comparison. Specifically, we used five state-of-the-art single-modal networks, as introduced in the previous Section II. The results of this assessment are organized and are presented in TABLE 1.

For lightweight network such as UNet, TD achieves the largest gains of +12.8% mIoU (vs RGB) and +15.4% mIoU (vs raw disparity), demonstrating its effectiveness in noise suppression for basic architectures. Even for complex designs like DeepLabv3+, TD maintains +11.3% mIoU over raw disparity inputs, confirming compatibility with high-capacity models. TD consistently outperforms both RGB and raw disparity in all networks, with >9.3% mIoU improvements over disparity (GSCNN) and >8.3% gains over RGB (PSPNet). Notably, TD achieves the highest absolute accuracy in every architecture (bold values in Table 1), proving it is independent of backbone design.

Table 1. Single-Modality Benchmark Across Architectures (Unit:%).
Bold: best results. Δ : absolute improvement of TD over RGB/Disparity.

Neural Networks	Modality	mIoU	mFsc	IoU		Fsc	
				Δ -RGB	Δ -Disp	Δ -RGB	Δ -Disp
UNet	Rgb	35.3	51.7	N/A	2.6	N/A	2.9
	Disparity	32.7	48.8	-2.6	N/A	-2.9	N/A
	TD	48.1	64.6	12.8	15.4	12.9	15.8
DenseASPP	Rgb	35.6	52.3	N/A	-2.3	N/A	-2.1
	Disparity	37.9	54.4	2.3	N/A	2.1	N/A
	TD	48.2	64.5	12.6	10.3	12.2	10.1
DeepLabv3+	Rgb	45.4	62.1	N/A	6	N/A	6
	Disparity	39.4	56.1	-6	N/A	-6	N/A
	TD	50.7	66.7	5.3	11.3	4.6	10.6
PSPNet	Rgb	39.7	56.3	N/A	2.5	N/A	2.9
	Disparity	37.2	53.4	-2.5	N/A	-2.9	N/A
	TD	48	64.3	8.3	10.8	8	10.9
GSCNN	Rgb	37.7	54.1	N/A	-2.1	N/A	-2.6
	Disparity	39.8	56.7	2.1	N/A	2.6	N/A
	TD	49.1	65.4	11.4	9.3	11.3	8.7

While Table 1 summarizes the overall mean performance, Table 2 details the IoU/Fsc scores for each specific category (Nail #1 to #4), revealing critical modality-dependent patterns. Across each category, TD inputs consistently outperform RGB and Disparity modalities (e.g., DeepLabv3+ TD: 50.7% mean IoU vs. 45.4% for RGB), with particularly marked gains on challenging categories like Nail #4 (UNet TD: +11.9% IoU over RGB).

Table 2. Class-Wise Performance of Single-Modality Networks (Unit:%).

Neural Networks	Approach Input	Nail #1		Nail #2		Nail #3		Nail #4		Mean	
		IoU	Fsc	IoU	Fsc	IoU	Fsc	IoU	Fsc	IoU	Fsc
UNet	Rgb	43.6	60.7	41.5	58.7	32.4	48.9	23.9	38.5	35.3	51.7
	Disparity	42.0	59.2	35.6	52.5	33.1	49.7	20.2	33.7	32.7	48.8
	TD	57.0	72.6	53.3	69.5	46.4	63.3	35.8	52.8	48.1	64.6
DenseASPP	Rgb	38.1	55.2	39.6	56.7	38.3	55.4	26.5	42.0	35.6	52.3
	Disparity	46.9	63.9	43.2	60.3	36.2	53.2	25.2	40.2	37.9	54.4
	TD	58.4	73.7	51.8	68.2	47.9	64.8	34.6	51.4	48.2	64.5
DeepLabv3+	Rgb	52.8	69.1	51	67.5	43.7	60.8	34.1	50.9	45.4	62.1
	Disparity	47.2	64.2	42.1	59.3	40.4	57.6	27.8	43.5	39.4	56.1
	TD	62.6	77	56.4	72.1	47	63.9	36.7	53.7	50.7	66.7
PSPNet	Rgb	47.1	64.0	44.1	61.2	40.4	57.5	27.1	42.6	39.7	56.3
	Disparity	49.8	66.5	41.2	58.3	35.3	52.2	22.4	36.6	37.2	53.4
	TD	56.0	71.8	53.8	69.9	48.1	65.0	33.9	50.7	48.0	64.3
GSCNN	Rgb	46.7	63.7	42.2	59.4	37.5	54.6	24.2	38.9	37.7	54.1
	Disparity	43.7	60.8	45.7	62.8	40.3	57.5	29.6	45.7	39.8	56.7
	TD	59.8	74.9	52.2	68.6	48.6	65.4	35.9	52.8	49.1	65.4

4. Summary

Construction site waste recognition faces the core challenges of cluttered background and strong noise interference, and to suppress the noise effect, this study develops the QIPT method to generate transformed parallax map (TD) modalities. On the self-constructed construction site waste dataset, the validation of five unimodal network architectures shows that the segmentation accuracy of TD is significantly better than that of RGB and the original parallax modality (up to 17.1% mIoU improvement). This demonstrates that TD can effectively improve the sharp object segmentation robustness through noise suppression and target enhancement, and the conclusion is generally applicable in the tested CNN architectures.

Acknowledgements

This work is supported by the Research Foundation of Shenzhen Polytechnic University under Grant 6023312012K.

References

- [1] Ram, V. G. and Kalidindi, Satyanarayana N., 2017, Estimation of construction and demolition waste using waste generation rates in Chennai, India, *Waste Management & Research*, Volume: 35, Issue: 6, pp. 610-617, DOI: 10.1177/0734242X17693297
- [2] Duan, H. and Li, J., 2016, Construction and demolition waste management: China's lessons, *Waste Management & Research*, Volume: 34, Issue: 5, pp.397-398, DOI:10.1177/0734242X16647603
- [3] Bovea, M.D. and Powell, J.C., 2016, Developments in life cycle assessment applied to evaluate the environmental performance of construction and demolition wastes, *Waste Management*, Volume:50, Issue: 4, pp.151-172, DOI: 10.1016/j.wasman.2016.01.036

- [4] Bao, ZK, Lee WMW, Lu, WS., 2020, Implementing on-site construction waste recycling in Hong Kong: Barriers and facilitators, *Science of the Total Environment*, Volume 747, Article Number: 141091, DOI: 10.1016/j.scitotenv.2020.141091
- [5] Zhencheng Hu, Lamosa, F., & Uchimura, K. (2005) A Complete U-V-Disparity Study for Stereovision Based 3D Driving Environment Analysis. Fifth International Conference on 3-D Digital Imaging and Modeling (3DIM'05), June 13, 1730 Massachusetts Ave., NW Washington, DC
- [6] Shelhamer, E., Long, J., & Darrell, T. (2017) Fully Convolutional Networks for Semantic Segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4): 640–651, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, DOI:10.1109/TPAMI.2016.2572683
- [7] Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017) SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39(12): 2481–2495, DOI:10.1109/TPAMI.2016.2644615
- [8] Ronneberger, O., Fischer, P., & Brox, T. (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation, *Medical Image Computing and Computer-Assisted Intervention – MICCAI*, November 18, Munich, Germany.
- [9] Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017) Pyramid Scene Parsing Network, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, December 4, Honolulu, Hawaii, USA
- [10] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018) DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40(4): 834–848, DOI:10.1109/TPAMI.2017.2699184
- [11] Yang, M., Yu, K., Zhang, C., Li, Z., & Yang, K. (2018) DenseASPP for Semantic Segmentation in Street Scenes. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, Salt Lake City, UT, USA
- [12] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018) Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *Computer Vision – ECCV*, September 8-14, Munich, Germany
- [13] Takikawa, T., Acuna, D., Jampani, V., & Fidler, S. (2019) Gated-SCNN: Gated Shape CNNs for Semantic Segmentation. *IEEE/CVF International Conference on Computer Vision (ICCV)*, October 27-November 2, Seoul, Korea.
- [14] Zhencheng Hu, Lamosa, F., & Uchimura, K. (2005) A Complete U-V-Disparity Study for Stereovision Based 3D Driving Environment Analysis. Fifth International Conference on 3-D Digital Imaging and Modeling (3DIM'05), June 13, 1730 Massachusetts Ave., NW Washington, DC
- [15] Fan, R., & Liu, M. (2020) Road Damage Detection Based on Unsupervised Disparity Map Segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 21(11), 4906–4911. DOI:10.1109/TITS.2019.2947206