2023 IEEE 26th International Conference on
Intelligent Transportation Systems (ITSC)
24-28 September 2023. Bilbao, Bizkaia, Spain

# CPKD: Channel and Position-wise Knowledge Distillation for Segmentation of Road Negative Obstacles

Zhen Feng[1,2], Yanning Guo[2], and Yuxiang Sun[1,*]

2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC) | 979-8-3503-9946-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/ITSC57777.2023.10422225

***Abstract*— Segmentation of road negative obstacles is important for the safety of autonomous vehicles. Many multi-modal fusion networks have been proposed for this task. They have achieved acceptable performance. However, most of them are heavyweight, making them hard to run real-timely, especially when working with high-resolution images. To address this issue, we propose a channel and position-wise knowledge distillation framework to train a lightweight student to achieve comparable accuracy and better efficiency. Specifically, we introduce a downsampling layer at the beginning of the student network to reduce the input data size to the student network, and introduce an upsampling layer at the end to restore the resolution. We propose a channel and position-wise distillation module to transfer knowledge between different sizes of feature maps. In addition, we release an RGB-Depth dataset for negative-obstacle segmentation. Experimental results demonstrate the effectiveness of our proposed method. Our code and dataset are available at: https://github.com/lab-sun/CPKD.**

## I. Introduction

Negative obstacles (e.g., potholes, cracks) on roads can cause traffic accidents [1]. It is important for autonomous vehicles to detect or segment negative obstacles so that vehicles can plan safe paths [2]. To achieve better segmentation performance, multi-modal fusion has recently attracted great attention in the research community. It can take advantages of each modality so that different modalities can be complemented by each other. There are many multi-modal fusion networks for road negative obstacles segmentation [3], [4]. These networks mainly focus on segmentation accuracy, and less on efficiency. So, when given input images with large resolutions, the inference speed could be reduced. To accelerate the inference speed, there have been many research efforts, such as knowledge distillation [5] and pruning [6].

In this paper, we propose a Channel and Position-wise Knowledge Distillation (CPKD) framework to train a lightweight student network to achieve comparable accuracy and better efficiency. We first place a downsampling layer at the beginning of the lightweight student network to reduce the input data size of the network. An upsampling layer is placed at the end of the student network to restore the resolution. In addition, to reduce the loss of spatial information due to downsampling, we propose a Channel and Position-wise Distillation (CPD) module to transfer knowledge between feature maps with different resolutions and numbers of channels. The CPD module enables the student network to retain the spatial information. Moreover, we find that there are very limited datasets for negative obstacle segmentation, so we build a dataset in this work. Our contributions are summarized as follows:

- We propose the CPD module to transfer knowledge between two feature maps with different resolutions and channels.
- We propose the CPKD framework to transfer knowledge from the well-trained heavyweight teacher network to the lightweight student network.
- We build and release a RGB-D dataset for road-negative-obstacles segmentation with $3,000$ generated labels and 745 manually labeled labels.

## II. Related Works

### A. Negative Obstacles Segmentation and Detection

Han *et al.* [7] proposed a reflection attention unit and combined the proposed unit with the FCN-8s [8] network for road-puddle segmentation. The authors released a dataset named *Puddle-1000* for the segmentation of puddles. Bhatia *et al.* [9] designed a convolutional neural network (CNN) for the detection of potholes with thermal images. Wu *et al.* [10] proposed a scale-adaptive detection and tracking framework to detect and track road potholes. They generated 3-D point clouds of roads and detect potholes with the 3-D point clouds. Pan *et al.* [11] presented an approach to detect the potholes with multi-spectral images.

Fan *et al.* [3] introduced channel attention module, position attention module, and dual attention module to RTFNet [12] to design AA-RTFNet for road-potholes segmentation. They released an RGB-D dataset, *Pothole-600*, for the segmentation of road potholes. Feng *et al.* [4] adopted the channel attention module and dual attention module to design fusion modules to fuse RGB images and disparity images. They combined the Transformer structure and CNN to design MAFNet to segment road potholes. Fan *et al.* [13] proposed a Graph Attention Layer (GAL) and integrated the GAL into DeepLabv3+ [14] to design GAL-DeepLabv3+ for the segmentation of road potholes.

### B. Knowledge Distillation

Knowledge distillation is mainly used to enable a lightweight student network to learn some capabilities from a heavyweight teacher network so that the efficiency could be improved while keeping accuracy [15]. Qin *et al.* [5] adopted the knowledge distillation method to design a lightweight network for medical image segmentation. They

[1]The Hong Kong Polytechnic University, Hung Hom, Hong Kong (email: zfeng94@outlook.com; yx.sun@polyu.edu.hk, sun.yuxiang@outlook.com).
[2]Harbin Institute of Technology, Harbin, Heilongjiang, China (email: zfeng94@outlook.com; guoyn@hit.edu.cn).
*Corresponding author: Yuxiang Sun.

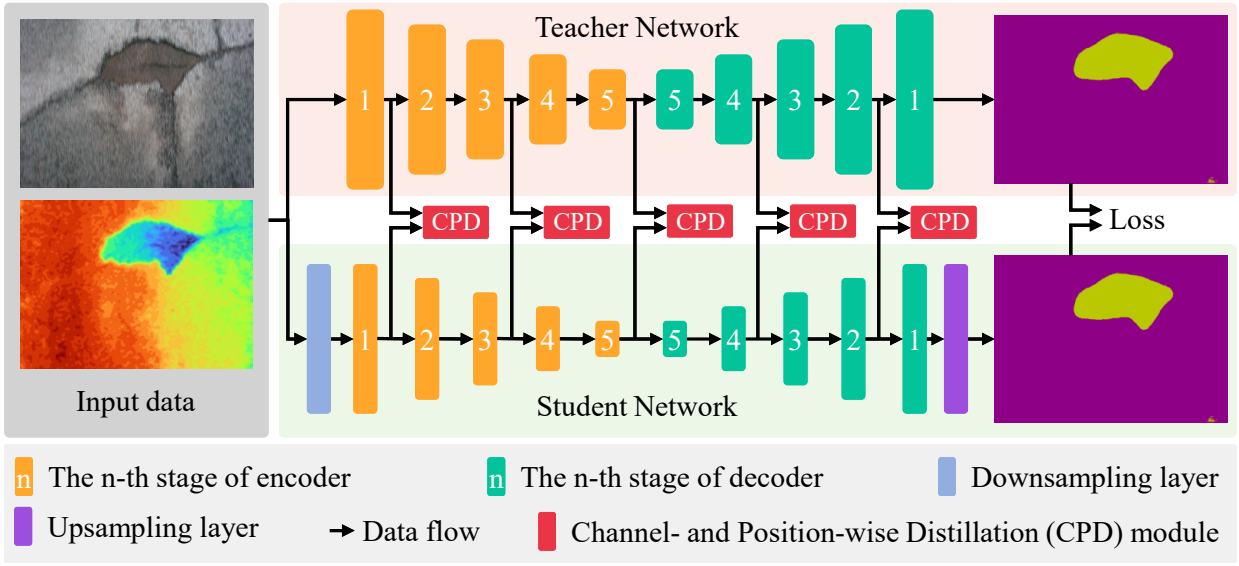


Fig. 1: The overall architecture of our proposed Channel and Position-wise Knowledge Distillation framework. The teacher network and student network share the same input data (i.e., RGB images and depth images). The student network has a similar structure with fewer parameters and layers to the teacher network. In the student network, the input data is first fed into a downsampling layer to reduce the resolution. The output of the student network is fed into an upsampling layer to restore the resolution to that of the input images. The outputs of the 1st, 3rd, and 5th stages of encoders, as well as the outputs of the 4th and 2nd of decoders, are fed into our proposed CPD module. The figure is best viewed in color.

designed a RAD module to transfer knowledge from a well-trained heavyweight teacher network to a lightweight student network. Liu *et al.* [16] designed a pair-wise distillation scheme and holistic distillation scheme to transfer structured knowledge from a teacher network to a student network. Komodakis *et al.* [17] employed attention maps to transfer knowledge from a teacher network to a student network. They proposed activation-based and gradient-based methods to generate spatial attention maps for feature maps. Shu *et al.* [18] proposed a channel-wise distillation paradigm that normalizes the activation map of each channel to generate a soft probability map. They used the soft probability maps of the teacher network and student work to transfer knowledge from the teacher network to the student work. Zheng *et al.* [19] proposed a localization distillation method to enable a student network to learn knowledge from a teacher network. Knowledge distillation is also commonly used to transfer cross-modal capabilities. Feng *et al.* [20] adopted the knowledge distillation method to transfer the edge detection capability from the edge-sharp RGB modality to the edge-blurred thermal modality.

Although the above works achieved acceptable performance, they suffer from a limitation that they transfer knowledge between feature maps with the same resolution or the same number of channels.

## III. The Proposed Method

### A. The Overall Framework

Fig. 1 shows the overall architecture of our proposed CPKD framework. There are a teacher network and a student network in this framework. The teacher network is a heavyweight network. The student network is a lightweight network with fewer parameters and layers than the teacher network. The structure of the student network is similar to that of the teacher network. The teacher network and the student network share the same input data (i.e., RGB and depth images). We increase the efficiency of the student network by reducing the input data size. Specifically, the input images are first fed into a downsampling layer to reduce the resolution. We place an upsampling layer at the last layer of the student network to restore the resolution to that of the input images.

Note that the number of channels and the output resolutions at the same level stages are different in the student and teacher networks. We employ MAFNet [4] as the teacher network and replace the first four stages of the encoder with the initial module and the first three stages of ResNet-152 [21]. We replace the first four stages of the encoder of MAFNet with the initial module and the first three stages of ResNet-18 as the student network. The numbers of channels of the output at each stage of the teacher network are $[64, 256, 512, 1024, 2048, 1024, 512, 256, 128, 2]$. The output resolutions at each stage of the teacher network are $[288 \times 512, 144 \times 256, 72 \times 128, 36 \times 64, 18 \times 32, 36 \times 64, 72 \times 128, 144 \times 256, 288 \times 512, 576 \times 1024]$. However, the numbers of channels of the outputs at each stage of the teacher are $[64, 64, 128, 256, 512, 256, 128, 64, 32, 2]$. The output resolutions at each stage of the teacher are $[144 \times 256, 72 \times 128, 36 \times 64, 18 \times 32, 9 \times 16, 18 \times 32, 36 \times 64, 72 \times 128, 144 \times 256, 288 \times 512]$.

In order to fully learn the capabilities of the teacher network, the student network learns the encoding and decod-

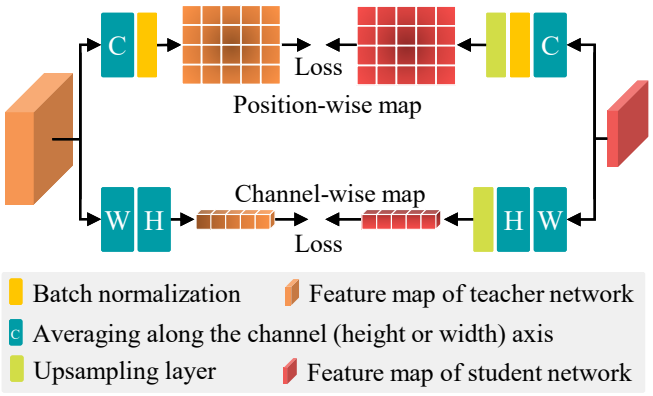


Fig. 2: The structure of our proposed CPD module. The size of the position-wise map and channel-wise map of the student network is resized to that of the teacher network by the upsampling layers. The figure is best viewed in color.

ing capabilities of the teacher network, respectively. In the encoding stage, the outputs of the 1st, 3rd, and 5th stages of both the teacher network and student network are fed into our proposed CPD module. The CPD module transfers the knowledge from the teacher network to the student network. In the decoding stage, the outputs of the 4th and 2nd stages of both networks are fed into the CPD module. The outputs of the teacher network are ground-truth labels for the student network.

### B. The CPD Module

The CPD module is used to transfer knowledge between two feature maps with different resolutions and numbers of channels. The CPD module generates position-wise maps and channel-wise maps for feature maps of the teacher network and student network. In each CPD module, we first average the feature map of the teacher network along the channel axis, and then feed the result into a batch normalization (BN) layer to generate a position-wise map. Secondly, we average the feature map of the student network along the channel axis and fed the result into a BN layer. We use an upsampling layer to adjust the resolution of the output of the BN layer to the resolution of the teacher-network position-wise map. We use the mean squared error loss to reduce the gaps between both position-wise maps to ensure that the student network can learn the capabilities of the teacher network. We average the feature map of the teacher network along the width axis and the height axis to generate the channel-wise map of the teacher network. We also first average the feature map of the student network along the width axis and the height axis. Then, we use an upsampling layer to resize the resolution of the result to that of the channel-wise map of the teacher network. The output of the upsampling layer is the channel-wise map of the student network. We also use the mean squared error loss to enable both channel-wise maps to be similar.

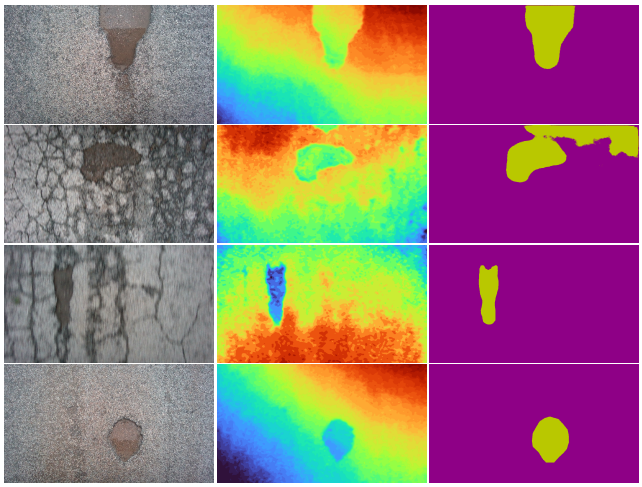

Fig. 3: Sample RGB images, depth images, and labels in our NO-4K dataset. The depth images are colored by the *jet* color map. Depth values increase from blue to red. The labels in the first two rows are generated by MAFNet [4] and the labels in the last two rows are manually labelled. ■ and ■ represent roads and negative obstacles.

## IV. The Dataset

We build and release a new RGB-D dataset for our task. We record the dataset with an on-vehicle Intel RealSense Depth Camera D455 in rural environments of Fushun City, Liaoning Province, China. The camera is looking at the road. We collect data on roads in different conditions, such as dry and wet. To increase the diversity of the data, we randomly flip the captured images along the x-axis or y-axis, as well as rotate the captured images by random angles. We collect a total of $3,745$ pairs of RGB-D images containing negative obstacles with the $576 \times 1024$ resolution.

Manually labeling of datasets for semantic segmentation is a labor-intensive task. To reduce the workload, we generate road-negative obstacle masks using a mixed dataset that contains existing datasets and a small number of manually-labeled datasets we collected. Specifically, we first manually labeled 745 images. We split these images into two categories: a training set with 245 pairs of images, and a testing set with 500 images. Secondly, we add all the images in the pothole dataset *Pothole-600*, an RGB-D road-pothole dataset similar to the data in our dataset, to the training set for training a network for the segmentation of road negative obstacles. The resolution of the images in the *Pothole-600* dataset is resized to be the same as that of the images in our dataset. Finally, we adopt the network with the best performance on the testing set to generate negative-obstacle masks for the unlabeled $3,000$ pairs of images. We adopt MAFNet which is designed for *Pothole-600* as the label-generated network.

We name our dataset as NO-4K (around 4,000 pairs of images containing negative obstacles). In our dataset, we split the $3,000$ pairs of images with generated labels into the training set, the 245 pairs of images into the validation

TABLE I: The comparative results (%) of the teacher network and student network on the testing set of our NO-4K dataset. *Student-u* means that the downsampling layer and the upsampling layer sample are removed from the student network. *Student-s* means that the student network is trained with a supervised method. *Student-c* means that the student network is trained with our CPKD method. The best results are highlighted in bold font.

| Method | Background | | | Negative Obstacles | | | mPre | F1 | mIoU | RTX 3060 | | RTX 3090 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Pre | F1 | IoU | Pre | F1 | IoU | | | | ms | FPS | ms | FPS |
| Teacher | **99.44** | 99.17 | 98.35 | 82.36 | 86.12 | 75.62 | 90.90 | 92.64 | 86.98 | 295.59 | 3.38 | 119.95 | 8.34 |
| Student-u | 99.39 | 99.04 | 98.09 | 79.49 | 84.16 | 72.66 | 89.44 | 91.60 | 85.37 | 57.33 | 17.44 | 24.00 | 41.66 |
| Student-s | 99.34 | 99.03 | 98.08 | 79.80 | 83.91 | 72.28 | 89.57 | 91.47 | 85.18 | 19.66 | 50.86 | 18.43 | 54.25 |
| Student-c (Ours) | 99.31 | **99.23** | **98.48** | **85.54** | **86.76** | **76.61** | **92.43** | **92.99** | **87.54** | **19.57** | **51.10** | **18.52** | **54.00** |

TABLE II: The results (%) of the ablation study on the position of the CPD module. ✓means the output of the stage is fed into the CPD module. The best results are highlighted in bold font.

| Variant | Encoder | | | | | Decoder | | | | | mPre | mF1 | mIoU |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 5 | 4 | 3 | 2 | 1 | | | |
| A | | | | | | | | | ✓ | | 89.97 | 91.94 | 85.89 |
| B | | | | | | | ✓ | | ✓ | | 90.80 | 92.28 | 86.42 |
| C | | | | | ✓ | | ✓ | | ✓ | | 91.01 | 92.56 | 86.86 |
| D | | | ✓ | | ✓ | | ✓ | | ✓ | | 91.16 | 92.66 | 87.02 |
| E | ✓ | | ✓ | | ✓ | | ✓ | | ✓ | | **92.43** | **92.99** | **87.54** |
| F | ✓ | | | | | | | | | | 90.87 | 92.58 | 86.88 |
| G | ✓ | | ✓ | | | | | | | | 91.39 | 92.62 | 86.95 |
| H | ✓ | | ✓ | | ✓ | | | | | | 91.20 | 92.64 | 86.98 |
| I | ✓ | | ✓ | | ✓ | | ✓ | | | | 91.23 | 92.63 | 86.97 |
| E | ✓ | | ✓ | | ✓ | | ✓ | | ✓ | | **92.43** | **92.99** | **87.54** |

set, and the 500 pairs of images into the testing set. Some samples of our dataset are shown in Fig. 3. Similar to the *Pothole-600* dataset, we colored the depth images using the jet color scheme to highlight the negative-obstacle regions. Due to factors, such as camera tilt caused by uneven roads or vehicle vibrations, different places of the road are at different distances from the camera. Some places are even at a larger distance from the camera than the depth of the negative obstacles. This makes it more difficult to segment the negative obstacles only based on depth images. From the second row in Fig. 3, we can see that the generated negative-obstacle labels are not accurate in wet areas, which is also a challenge for our dataset.

## V. Experimental Results and Discussions

### A. Training Details

We implement our network with PyTorch, which is trained and tested on a PC with an NVIDIA RTX 3090 graphics card. We also test the inference speed of the network on a PC with an NVIDIA RTX 3060 graphics card. The training scheme of MAFNet is employed to train our network.

We first train the teacher network to achieve the best performance on the testing set of our NO-4K dataset. Secondly, we fix the well-trained weights of the teacher network to train the student network.

TABLE III: The results (%) of the ablation study on the structure of the CPD module. The best results are highlighted in bold font.

| Variant | mPre | mF1 | mIoU |
|---|---|---|---|
| Position | 90.42 | 92.25 | 86.37 |
| Channel | 91.94 | 92.89 | 87.38 |
| Position & Channel | **92.43** | **92.99** | **87.54** |

### B. Ablation Study

*1) Ablation Study on the Position of the CPD Module:* We conduct an ablation study to find the best position for the CPD module. We use two approaches to design variants. Firstly, we design variants by placing a different number of CPD modules in the network sequentially, starting at the end of the network. Secondly, we design variants by placing a different number of CPD modules in the network sequentially, starting at the beginning of the network. The details of each variant are shown in Tab. II. We use the Precision (Pre), the F-score (F1), and the Intersection over Union (IoU) to evaluate the performance of the network. The calculation of the metrics can be found in [4]. The results of each variant are displayed in Tab. II. Comparing the results of the variants A, B, C, D, and E, we can find that, in general, the more CPD modules in the network, the better the performance is achieved. We can also get the same conclusion from the results of variants F, G, H, I, and E. Comparing the results of variants A and F, we can find that the network achieves better performance with a CPD module placed at the beginning than that placed at the end. We conjecture the possible reason is that it is more difficult to recover the lost information caused by downsampling at the end of the network than at the beginning. The results show that our proposed CPKD framework achieves the best performance.

*2) Ablation Study on the Structure of the CPD Module:* We conduct an ablation study to illustrate the benefits of the channel-wise map and the position-wise map in the CPD module. We remove the channel-wise map and the position-wise map to design variants, respectively. The results of each variant are displayed in Tab. III. From the results, we can find that a network with CPD modules containing both channel-wise map and position-wise map achieves the best

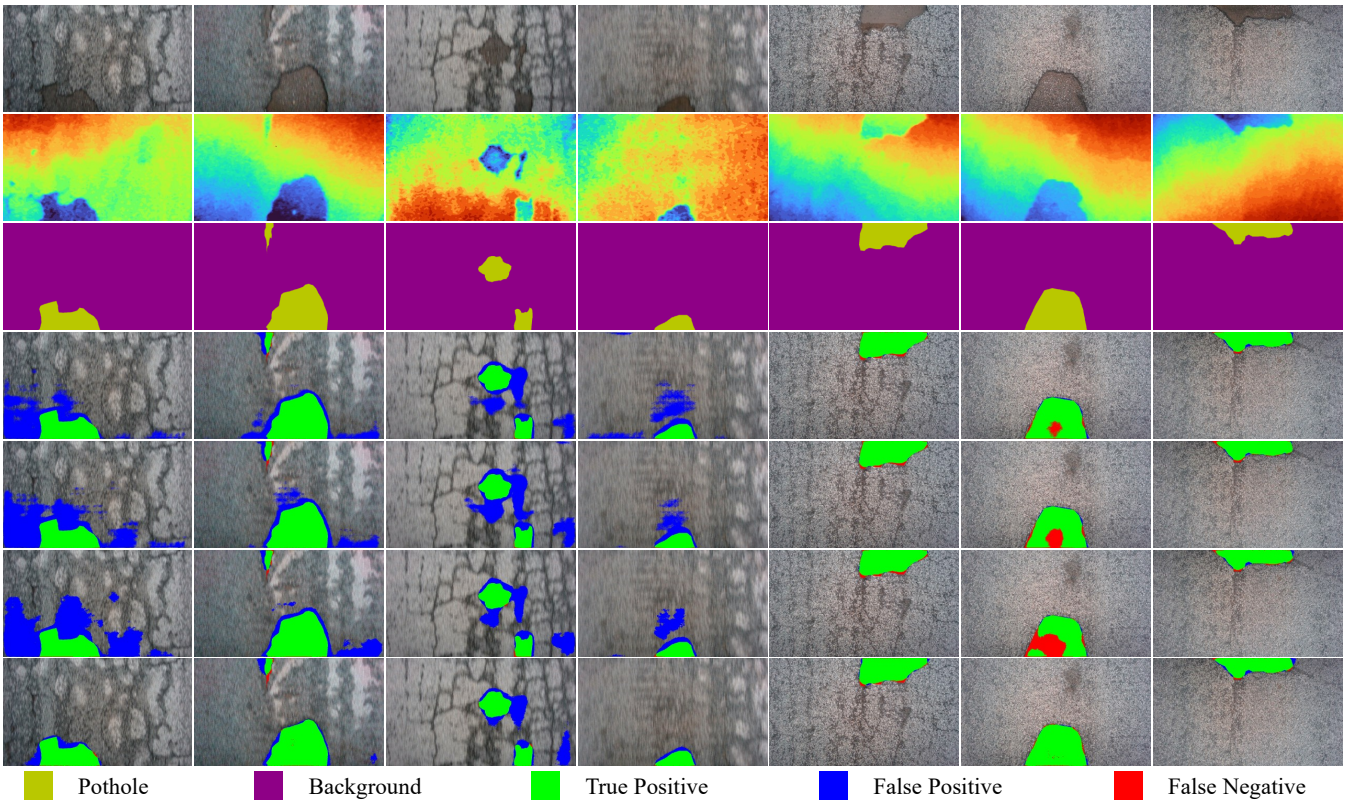

Fig. 4: Sample qualitative demonstrations of the four networks in Tab. I. The 4-th row to the 7-th row are respectively demonstrations of teacher, student-u, student-s, and student-c. The figure is best viewed in color.

performance. The results illustrate the effectiveness of each part of our proposed CPD module.

### C. Comparative Study

In order to demonstrate the effectiveness of our proposed CPKD framework, we compare the results of the teacher network, student network trained with a supervised method (abbreviated as student-s), and student network trained with the CPKD method (abbreviated as student-c). Moreover, we also compare the results of the student network that removes the downsampling layer and upsampling layer trained with a supervised method (abbreviated as student-u) to show the influence of both layers in a network. We evaluate the performance of the teacher network and the student network in terms of inference speed and segmentation accuracy. We test the runtime of each pair of images for each network on NVIDIA RTX 3060 and RTX 3090, respectively.

*1) Quantitative Results:* The quantitative results of the above networks (i.e., the teacher network, the variants student-s, student-c, and student-u) are shown in Tab. I. Comparing the results of the student-u and the student-s, we can find that the downsampling layer and upsampling layer increase the inference speed. However, they also lead to performance degradation due to the loss of spatial information during downsampling. Comparing the results of student-c and student-s, we can find that the student network learns more knowledge from the teacher network with our CPKD method than the student network learns on its own with the supervised method. Comparing the results of student-c and student-u, we can find that our proposed CPKD module can learn the lost information from the teacher network. The comparative results also show that our proposed CPKD framework can greatly increase the inference speed of the network while improving the accuracy of the student network. Comparing the results of the teacher and student-c, we can find that the inference speed of student-c is much faster than that of the teacher. All the results demonstrate that the CPKD module can transfer knowledge in two feature maps with different resolutions and numbers of channels.

It should be noted that the accuracy of student-c is better than that of the teacher. One possible influencing factor is the labels of wet areas in the dataset. As analyzed in section IV, the labels of the wet areas generated by MAFNet are inaccurate. The teacher network is affected by these inaccurate labels. However, the student network does not learn the relevant knowledge and reduce the influence of inaccurate labels, thus making the accuracy of the student network higher than that of the teacher network. In addition, another reason may be that the training labels of the student network are outputs of the teacher network, thus avoiding the influence of the inaccurate labels of the wet areas.

*2) Qualitative Demonstrations:* Some sample qualitative results of the four networks are shown in Fig. 4. From the first 4 columns of the results, we can see that the network trained by the supervised method presents a large area of inaccurate segmentation in wet areas. However, the student

network trained with our proposed CPKD method achieves the best results in the wet area, which is able to reduce the interference of the wet areas on the negative-obstacle segmentation. This result supports our conjecture on the reason for the higher accuracy of the student network than the teacher network. From the other results, we can see that our proposed CPKD method enables the student network to learn the main knowledge from the teacher network. Overall, these results illustrate the effectiveness of our proposed CPKD method.

## VI. Conclusion and Future Work

We proposed here a novel CPKD framework to improve the accuracy and efficiency of the student network. We proposed the CPKD framework to transfer knowledge between two feature maps with different resolutions and channel numbers. The experimental results show that our proposed CPKD module can learn the lost information caused by a downsampling layer from the teacher network. The student network trained with the Channel and Position-wise Knowledge Distillation framework achieves higher accuracy and efficiency. We also release a novel RGB-D dataset NO-4K for the segmentation of negative obstacles. However, there are also some limitations in our work. In future work, we would like to design a novel network with better performance for negative-obstacle segmentation as a teacher network.

## Acknowledgement

This work was supported in part by National Natural Science Foundation of China under Grant 62003286, in part by Zhejiang Lab under grant 2021NL0AB01, in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515010116, and in part by CCF-Baidu Open Fund under Grant 182215PCK04183.

## References

[1] T. Verster and E. Fourie, "The good, the bad and the ugly of south african fatal road accidents," *South African Journal of Science*, vol. 114, no. 7-8, pp. 63–69, 2018.

[2] G. Cheng and J. Y. Zheng, "Sequential semantic segmentation of road profiles for path and speed planning," *IEEE Trans. on Intell. Transp. Sys.*, vol. 23, no. 12, pp. 23 869–23 882, 2022.

[3] R. Fan, H. Wang, M. J. Bocus, and M. Liu, "We learn better road pothole detection: from attention aggregation to adversarial domain adaptation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 285–300.

[4] Z. Feng, Y. Guo, Q. Liang, M. U. M. Bhutta, H. Wang, M. Liu, and Y. Sun, "Mafnet: Segmentation of road potholes with multimodal attention fusion network for autonomous vehicles," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022.

[5] D. Qin, J.-J. Bu, Z. Liu, X. Shen, S. Zhou, J.-J. Gu, Z.-H. Wang, L. Wu, and H.-F. Dai, "Efficient medical image segmentation based on knowledge distillation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3820–3831, 2021.

[6] S. Xu, A. Huang, L. Chen, and B. Zhang, "Convolutional neural network pruning: A survey," in *2020 39th Chinese Control Conference (CCC)*, 2020, pp. 7458–7463.

[7] X. Han, C. Nguyen, S. You, and J. Lu, "Single image water hazard detection using fcn with reflection attention units," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 105–120.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[9] Y. Bhatia, R. Rai, V. Gupta, N. Aggarwal, A. Akula *et al.*, "Convolutional neural networks based potholes detection using thermal imaging," *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 3, pp. 578–588, 2022.

[10] R. Wu, J. Fan, L. Guo, L. Qiao, M. U. M. Bhutta, B. Hosking, S. Vityazev, and R. Fan, "Scale-adaptive pothole detection and tracking from 3-d road point clouds," in *2021 IEEE International Conference on Imaging Systems and Techniques (IST)*, 2021, pp. 1–5.

[11] Y. Pan, X. Zhang, G. Cervone, and L. Yang, "Detection of asphalt pavement potholes and cracks based on the unmanned aerial vehicle multispectral imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 10, pp. 3701–3712, 2018.

[12] Y. Sun, W. Zuo, and M. Liu, "Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes," *IEEE Robot. Autom. Lett*, vol. 4, no. 3, pp. 2576–2583, 2019.

[13] R. Fan, H. Wang, Y. Wang, M. Liu, and I. Pitas, "Graph attention layer evolves semantic segmentation for road pothole detection: A benchmark and algorithms," *IEEE transactions on image processing*, vol. 30, pp. 8144–8154, 2021.

[14] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[15] L. Wang and K.-J. Yoon, "Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 3048–3068, 2022.

[16] Y. Liu, C. Shu, J. Wang, and C. Shen, "Structured knowledge distillation for dense prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7035–7049, 2023.

[17] N. Komodakis and S. Zagoruyko, "Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer," in *ICLR*, 2017.

[18] C. Shu, Y. Liu, J. Gao, Z. Yan, and C. Shen, "Channel-wise knowledge distillation for dense prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5311–5320.

[19] Z. Zheng, R. Ye, Q. Hou, D. Ren, P. Wang, W. Zuo, and M.-M. Cheng, "Localization distillation for object detection," *arXiv preprint arXiv:2204.05957*, 2022.

[20] Z. Feng, Y. Guo, and Y. Sun, "Cekd: Cross-modal edge-privileged knowledge distillation for semantic scene understanding using only thermal images," *IEEE Robot. Autom. Lett*, vol. 8, no. 4, pp. 2205–2212, 2023.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.