We Learn Better Road Pothole Detection: From Attention Aggregation to Adversarial Domain Adaptation

Rui Fan[1], Hengli Wang[2], Mohammud J. Bocus[3], and Ming Liu[2](✉)

[1] UC San Diego, La Jolla, USA
rui.fan@ieee.org
[2] HKUST Robotics Institute, Kowloon, Hong Kong
{hwangdf,eelium}@ust.hk
[3] University of Bristol, Bristol, UK
junaid.bocus@bristol.ac.uk

Abstract. Manual visual inspection performed by certified inspectors is still the main form of road pothole detection. This process is, however, not only tedious, time-consuming and costly, but also dangerous for the inspectors. Furthermore, the road pothole detection results are always subjective, because they depend entirely on the individual experience. Our recently introduced disparity (or inverse depth) transformation algorithm allows better discrimination between damaged and undamaged road areas, and it can be easily deployed to any semantic segmentation network for better road pothole detection results. To boost the performance, we propose a novel attention aggregation (AA) framework, which takes the advantages of different types of attention modules. In addition, we develop an effective training set augmentation technique based on adversarial domain adaptation, where the synthetic road RGB images and transformed road disparity (or inverse depth) images are generated to enhance the training of semantic segmentation networks. The experimental results demonstrate that, firstly, the transformed disparity (or inverse depth) images become more informative; secondly, AA-UNet and AA-RTFNet, our best performing implementations, respectively outperform all other state-of-the-art single-modal and data-fusion networks for road pothole detection; and finally, the training set augmentation technique based on adversarial domain adaptation not only improves the accuracy of the state-of-the-art semantic segmentation networks, but also accelerates their convergence.

Source Code and Dataset:
http://sites.google.com/view/pothole-600

R. Fan and H. Wang—These authors contributed equally to this work and are therefore joint first authors.

A. Bartoli and A. Fusiello (Eds.): ECCV 2020 Workshops, LNCS 12538, pp. 1–16, 2020.
https://doi.org/10.1007/978-3-030-66823-5_17

1 Introduction

Potholes are small concave depressions on the road surface [1]. They arise due to a number of environmental factors, such as water permeating into the ground under the asphalt road surface [2]. The affected road areas are further deteriorated due to the vibration of tires, making the road surface impracticable for driving. Furthermore, vehicular traffic can cause the subsurface materials to move, and this generates a weak spot under the street. With time, the road damage worsens due to the frequent movement of vehicles over the surface and this causes new road potholes to emerge [3].

Road pothole is not just an inconvenience, but also poses a safety risk, because it can severely affect vehicle condition, driving comfort, and traffic safety [2]. It was reported in 2015 that Danielle Rowe, an Olympic gold medalist as well as three-time world champion, had eight fractured ribs resulting in a punctured lung, after hitting a pothole during a race [4]. Therefore, it is crucial and necessary to regularly inspect road potholes and repair them in time.

Currently, manual visual inspection performed by certified inspectors is still the main form of road pothole detection [5]. However, this process is not only time-consuming, exhausting and expensive, but also hazardous for the inspectors [3]. For example, the city of San Diego repairs more than 30K potholes per year using hot patches compound and bagged asphalt, and they have been requesting residents to report potholes so as to relieve the burden on the local road maintenance department [6]. Elsewhere, the UK government is set to pledge billions of pounds for filling potholes across the country [7]. Additionally, the pothole detection results are always subjective, as the decisions depend entirely on the inspector's experience and judgment [8]. Hence, there has been a strong demand for automated road condition assessment systems, which can not only acquire 2D/3D road data, but also detect and predict road potholes accurately, robustly and objectively [9].

Specifically, automated road pothole detection has been considered as more than an infrastructure maintenance problem in recent years, as many self-driving car companies have included road pothole detection into their autonomous car perception modules. For instance, Jaguar Land Rover announced their recent research achievements on road pothole detection/prediction [10], where the vehicles can not only gather the location and severity data of the road potholes, but also send driver warnings to slow down the car. Ford also claimed that they were experimenting with data-driven technologies to warn drivers of the pothole locations [11]. Furthermore, during the Consumer Electronics Show (CES) 2020, Mobileye demonstrated their solutions[1] for road pothole detection, which are based on machine vision and intelligence. With recent advances in image analysis and deep learning, especially for 3D vision data, depth/disparity image analysis and convolutional neural networks (CNNs) have become the mainstream techniques for road pothole detection [8].

[1] http://s21.q4cdn.com/600692695/files/doc_presentations/2020/1/Mobileye-CES-2020-presentation.pdf.

Given the 3D road data, image segmentation algorithms are typically performed to detect potholes. For example, Jahanshahi *et al.* [12] employed Otsu's thresholding method [13] to segment depth images for road pothole detection. In [2], we proposed a disparity image transformation algorithm, which can better distinguish between damaged and undamaged road areas. The road potholes were then detected using a surface modeling approach. Subsequently, we minimized the computational complexity of our algorithm and successfully embedded it in a drone for real-time road inspection [8]. Recently, the aforementioned algorithm was proved to have a numeric solution [5], which allows it to be easily deployed to any existing semantic segmentation networks for end-to-end road pothole detection.

In this paper, we first briefly introduce the disparity (or inverse depth, as disparity is in inverse proportion to depth) transformation (DT) algorithm proposed in [5]. We then exploit the aggregation of different types of attention modules (AMs) so as to improve the semantic segmentation networks for better road pothole detection. Furthermore, we develop a novel adversarial domain adaptation framework for training set augmentation. Moreover, we publish our road pothole detection dataset, named *Pothole-600*, at http://sites.google.com/view/pothole-600 for research purposes. According to our experimental results presented in Sect. 6, training CNNs with augmented road data yields better semantic segmentation results, where convergence is achieved with fewer iterations at the same time.

2 Related Works

2.1 Semantic Segmentation

Fully convolutional network (FCN) [14] was the first end-to-end single-modal CNN designed for semantic segmentation. Based on FCN, U-Net [15] adopts an encoder-decoder architecture. It also adds skip connections between the encoder and decoder to help smooth the gradient flow and restore the locations of objects. Additionally, PSPNet [16], DeepLabv3+ [17] and DenseASPP [18] leverage a pyramid pooling module to extract context information for better segmentation performance. Furthermore, GSCNN [19] employs a two-branch framework consisting of a shape branch and a regular branch, which can effectively improve the semantic predictions on the boundaries.

Different from the above-mentioned single-modal networks, many data-fusion networks have also been proposed to improve semantic segmentation accuracy by extracting and fusing the features from multi-modalities of visual information [20,21]. For instance, FuseNet [22] and depth-aware CNN [23] adopt the popular encoder-decoder architecture, but employ different operations to fuse the feature maps obtained from the RGB and depth branches. Moreover, RTFNet [24] was developed to improve semantic segmentation performance by fusing the features extracted from RGB images and thermal images. It also adopts an encoder-decoder architecture and an element-wise addition fusion strategy.

2.2 Attention Module

Due to their simplicity and effectiveness, AMs have been widely used in various computer vision tasks. AMs typically learn the weight distribution (WD) of an input feature map and output an updated feature map based on the learned WD [25]. Specifically, Squeeze-and-Excitation Network (SENet) [26] employs a channel-wise AM to improve image classification accuracy. Furthermore, Wang *et al.* [27] presented a non-local module to capture long-range dependencies for video classification. OCNet [28] and DANet [29] proposed different self-attention modules that are capable of using contextual information for semantic segmentation. Moreover, CCNet [30] adopts a criss-cross AM to obtain dense contextual information in a more efficient way. Different from the aforementioned studies, we propose an attention aggregation (AA) framework that focuses on the combination of different AMs. Based on this idea, our proposed AA-UNet and AA-RTFNet can take advantage of different AMs and yield accurate results for road pothole detection.

2.3 Adversarial Domain Adaptation

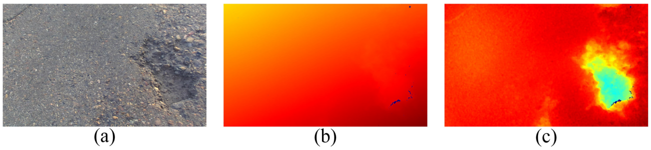
Since the concept of "generative adversarial network (GAN)" [31] was first introduced in 2014, great efforts have been made in this research area to improve the existing computer vision algorithms. The recipe for their success is the use of an adversarial loss, which makes the generated synthetic images become indistinguishable from the real images when minimized [32].

Recent image-to-image translation approaches typically utilize a dataset, which contains paired source and target images, to learn a parametric translation using CNNs. One of the most well-known work is the "pix2pix" framework [33] proposed by Isola *et al.*, which employs a conditional GAN to learn the mapping from source images to target images.

In addition to the paired image-to-image translation approaches mentioned above, many unsupervised approaches have also been proposed in recent years to tackle unpaired image-to-image translation problem, where the primary goal is to learn a mapping $G : \mathcal{S} \rightarrow \mathcal{T}$ from source domain $\mathcal{S}$ to target domain $\mathcal{T}$, so that the distribution of images from $G(\mathcal{S})$ is indistinguishable from the distribution $\mathcal{T}$. CycleGAN [32] is a representative work handling unpaired image-to-image translation, where an inverse mapping $F : \mathcal{T} \rightarrow \mathcal{S}$ and a cycle-consistency loss (aiming at forcing $F(G(\mathcal{S})) \cong \mathcal{S}$) were coupled with $G : \mathcal{S} \rightarrow \mathcal{T}$. Our proposed training set augmentation technique is developed based on CycleGAN [32], but it performs paired image-to-image translation.

3 Disparity (or Inverse Depth) Transformation

DT aims at transforming a disparity or inverse depth image $\mathbf{G}$ into a quasi bird's eye view, whereby the pixels in the undamaged road areas possess similar values, while they differ significantly from those of the pothole pixels.



(a) (b) (c)

Fig. 1. Disparity transformation: (a) RGB image; (b) disparity image produced by PT-SRP [34]; and (c) transformed disparity image.

Since the concept of "v-disparity domain" was introduced in [35], disparity image analysis has become a common technique used for 3D driving scene understanding [8]. The projections of the on-road disparity (or inverse depth) pixels in the v-disparity domain can be represented by a non-linear model as follows:

$$\tilde{\mathbf{q}} = \mathbf{M}\tilde{\mathbf{p}} = \varkappa \begin{bmatrix} -\sin\Phi & \cos\Phi & \kappa \\ 0 & 1/\varkappa & 0 \\ 0 & 0 & 1/\varkappa \end{bmatrix} \tilde{\mathbf{p}}, \tag{1}$$

where $\tilde{\mathbf{p}} = [u, v, 1]^\top$ is the homogeneous coordinates of a pixel in the disparity (or inverse depth) image, and $\tilde{\mathbf{q}} = [g, v, 1]^\top$ is the homogeneous coordinates of its projection in the v-disparity domain. $\mathbf{\Phi}$ can be estimated via [8]:

$$\arg\min_{\Phi} \mathbf{g}^\top\mathbf{g} - \mathbf{g}^\top\mathbf{T}(\Phi)\big(\mathbf{T}(\Phi)^\top\mathbf{T}(\Phi)\big)^{-1}\mathbf{T}(\Phi)^\top\mathbf{g}, \tag{2}$$

where $\mathbf{g}$ is a k-entry vector of disparity (or inverse depth) values, $\mathbf{1}_k$ is a k-entry vector of ones, $\mathbf{u}$ and $\mathbf{v}$ are two k-entry vectors storing the horizontal and vertical coordinates of the observed pixels, respectively, and $\mathbf{T}(\Phi) = [\mathbf{1}_k, \cos\Phi\mathbf{v} - \sin\Phi\mathbf{u}]$. (2) has a closed-form solution as follows [5]:

$$\Phi = \arctan\frac{\omega_4\omega_0 - \omega_3\omega_1 + q\sqrt{\Delta}}{\omega_3\omega_2 + \omega_5\omega_1 - \omega_5\omega_0 - \omega_4\omega_2} \quad \text{s.t. } q \in \{-1, 1\}, \tag{3}$$

where

$$\Delta = (\omega_4\omega_0 - \omega_3\omega_1)^2 + (\omega_3\omega_2 - \omega_5\omega_0)^2 - (\omega_4\omega_2 - \omega_5\omega_1)^2. \tag{4}$$

The expressions of ω_0-ω_5 are given in [5]. κ and $\varkappa$ can then be obtained using:

$$\mathbf{x} = \varkappa \begin{bmatrix} \kappa \\ 1 \end{bmatrix} = \big(\mathbf{T}(\Phi)^\top\mathbf{T}(\Phi)\big)^{-1}\mathbf{T}(\Phi)^\top\mathbf{g}. \tag{5}$$

DT can therefore be realized using [5]:

$$\mathbf{G}'(\mathbf{p}) = \mathbf{G}(\mathbf{p}) - \varkappa\big(\cos\Phi v - \sin\Phi u\big) - \varkappa\kappa + \Lambda, \tag{6}$$

where Λ is a constant used to ensure that the values in the transformed disparity (or depth inverse) image $\mathbf{G}'$ are non-negative. An example of the transformed disparity (or inverse depth) image is shown in Fig. 1, where it can be observed that the damaged road area becomes highly distinguishable. The effectiveness of DT on improving semantic segmentation is discussed in Sect. 6.4.

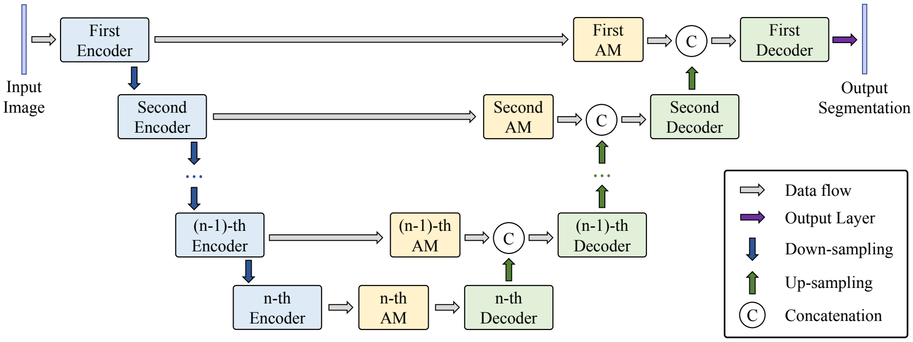


Fig. 2. The architecture of the proposed attention aggregation framework for our AA-UNet and AA-RTFNet.

4 Attention Aggregation Framework

The architecture of our proposed attention aggregation framework is illustrated in Fig. 2. We add different AMs into the existing CNNs that adopt the popular encoder-decoder architecture. Firstly, U-Net [15] has demonstrated the effectiveness of employing skip connections, which concatenate the same-scale feature maps produced by the encoder and decoder. However, these two feature maps can present large difference because of the different numbers of transformations undergone, which can result in significant performance degradation. To alleviate this drawback, we add an AM for the encoder feature map before the concatenation in each skip connection, as shown in Fig. 2 (from the 1st to (n−1)-th AMs), where n denotes the number of network levels. These AMs enable the encoder feature maps to focus on the potholes, which can shorten the gap between the same-scale feature maps produced by the encoder and decoder. This further improves pothole detection performance. Secondly, many studies [29,30] have already demonstrated that adding an AM for a high-level feature map can significantly improve the overall performance. Therefore, we follow this paradigm and add an AM at the highest level, as shown in Fig. 2 (n-th AM).

We use three AMs in our attention aggregation framework: 1) Channel Attention Module (CAM), 2) Position Attention Module (PAM) and 3) Dual Attention Module (DAM) [29], as illustrated in Fig. 3. Similar to SENet [26], our CAM is designed to assign each channel with a weight since some channels are more important. It first employs a global average pooling layer to squeeze spatial information, and then utilizes fully connected (FC) layers to generate the WD, which is finally combined with the input feature map by element-wise multiplication operation to generate the output feature map. Different from CAM, our PAM focuses on spatial information. It first generates the spatial WD and applies it on the input feature map to generate the output feature map. DAM [29] is composed of a channel attention submodule and a position attention submodule. Different from our CAM and PAM, these two submodules adopt the self-attention scheme to generate the WD, which can achieve better performance at the expense of

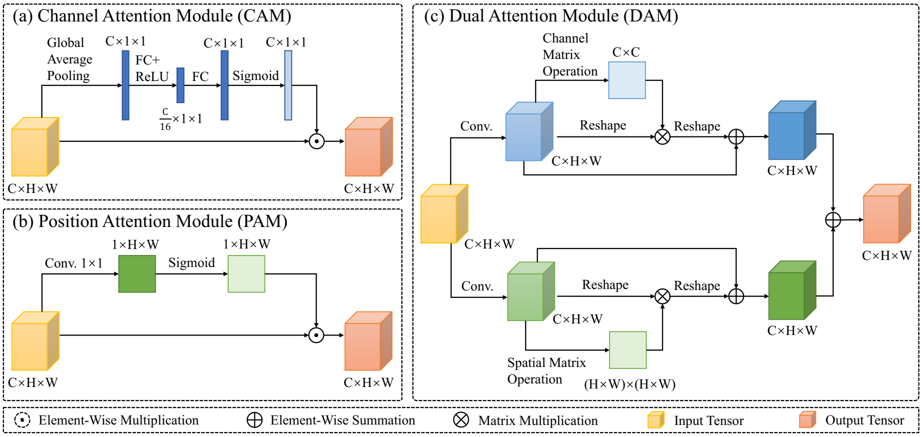


Fig. 3. The illustrations of the three AMs used in our attention aggregation framework.

a higher computational complexity. Since the memory consumed by DAM will grow significantly with the increase of feature map size, we only use it at the highest level (n-th AM) so as to ensure computational efficiency.

To demonstrate the effectiveness of our framework, we employ it in a single-modal network (U-Net) and a data-fusion network (RTFNet), and dub them as AA-UNet and AA-RTFNet, respectively. The specific architectures (the selection of each AM) of our AA-UNet and AA-RTFNet are discussed in Sect. 6.3.

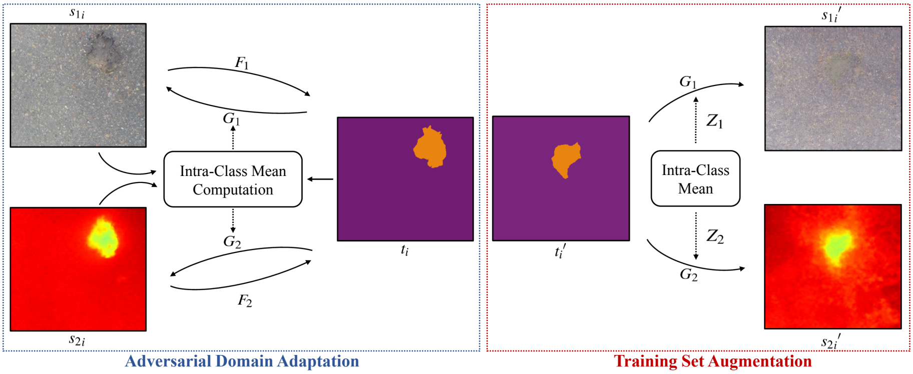
5 Adversarial Domain Adaptation for Training Set Augmentation

In this paper, adversarial domain adaptation is utilized to augment training set so that the semantic segmentation networks can perform more robustly. Our proposed training set augmentation framework is illustrated in Fig. 4, where $F_1 : \mathcal{S}_1 \rightarrow \mathcal{T}$ translates RGB images $s_{1i} \in \mathcal{S}_1$ to pothole detection ground truth $t_i \in \mathcal{T}$; $G_1 : \mathcal{T} \rightarrow \mathcal{S}_1$ translates pothole detection ground truth $t_i \in \mathcal{T}$ back to RGB images $s_{1i} \in \mathcal{S}_1$; $F_2 : \mathcal{S}_2 \rightarrow \mathcal{T}$ translates our transformed disparity images $s_{2i} \in \mathcal{S}_2$ to pothole detection ground truth $t_i \in \mathcal{T}$; and $G_2 : \mathcal{T} \rightarrow \mathcal{S}_2$ translates pothole detection ground truth $t_i \in \mathcal{T}$ back to our transformed disparity images $s_{2i} \in \mathcal{S}_2$. The learning of G_1 and G_2 is guided by the intra-class means. Our full objective is:

$$\begin{aligned}\mathcal{L}(G_1, G_2, F_1, F_2, D_{\mathcal{S}_1}, D_{\mathcal{S}_2}, D_{\mathcal{T}}) = {} & \mathcal{L}_{\mathrm{GAN}}(G_1, D_{\mathcal{S}_1}, \mathcal{T}, \mathcal{S}_1) + \mathcal{L}_{\mathrm{GAN}}(F_1, D_{\mathcal{T}}, \mathcal{S}_1, \mathcal{T}) \\ & + \mathcal{L}_{\mathrm{GAN}}(G_2, D_{\mathcal{S}_2}, \mathcal{T}, \mathcal{S}_2) + \mathcal{L}_{\mathrm{GAN}}(F_2, D_{\mathcal{T}}, \mathcal{S}_2, \mathcal{T}) \\ & + \mathcal{L}_{\mathrm{cyc}}(G_1, F_1) + \mathcal{L}_{\mathrm{cyc}}(G_2, F_2),\end{aligned} \tag{7}$$

where

$$\mathcal{L}_{GAN}(G, D_{\mathcal{S}}, \mathcal{T}, \mathcal{S}) = \mathbb{E}_{s \sim p_{\mathrm{data}}(s)}[\log D_{\mathcal{S}}(s)] + \mathbb{E}_{t \sim p_{\mathrm{data}}(t)}[\log(1 - D_{\mathcal{S}}(G(t)))], \tag{8}$$



Adversarial Domain Adaptation

Training Set Augmentation

Fig. 4. Adversarial domain adaptation for training set augmentation.

$$\mathcal{L}_{GAN}(F, D_{\mathcal{T}}, \mathcal{S}, \mathcal{T}) = \mathbb{E}_{t\sim p_{\mathrm{data}}(t)}[\log D_{\mathcal{T}}(t)] + \mathbb{E}_{s\sim p_{\mathrm{data}}(s)}[\log(1 - D_{\mathcal{T}}(F(s)))], \quad (9)$$

$$\mathcal{L}_{cyc}(G, F) = \mathbb{E}_{s\sim p_{\mathrm{data}}(s)}[G(F(s)) - s] + \mathbb{E}_{t\sim p_{\mathrm{data}}(t)}[F(G(t)) - t], \quad (10)$$

$D_{\mathcal{S}}$ and $D_{\mathcal{T}}$ are two adversarial discriminators: $D_{\mathcal{S}}$ aims to distinguish between images $\{s\}$ and the translated images $\{G(t)\}$, while $D_{\mathcal{T}}$ aims to distinguish between images $\{t\}$ and the translated images $\{F(s)\}$; $s \sim p_{\mathrm{data}}(s)$ and $t \sim p_{\mathrm{data}}(t)$ denote the data distributions of the source and target domains, respectively.

With well-learned mapping functions G_1 and G_2, we can generate an infinite number of synthetic RGB images ${s_1}_i' \in \mathcal{S}_1'$ and their corresponding synthetic transformed disparity images ${s_2}_i' \in \mathcal{S}_2'$ from a randomly generated pothole detection ground truth $t_i' \in \mathcal{T}'$. In order to expand the distributions of the two domains $s_1' \sim p_{\mathrm{data}}(s_1')$ and $s_2' \sim p_{\mathrm{data}}(s_2')$, we add random Gaussian noises Z_1 and Z_2 into G_1 and G_2 when generating ${s_1}_i'$ and ${s_2}_i'$, as shown in Fig. 4. Some examples in the augmented training set are shown in Fig. 5. The benefits of our proposed training set augmentation technique for semantic segmentation are discussed in Sect. 6.4.

6 Experiments

6.1 Datasets

Pothole-600. In our experiments, we utilized a stereo camera to capture stereo road images. These images are then split into a training set, a validation set and a testing set, which contains 240, 180 and 180 pairs of RGB images and transformed disparity images, respectively.

Augmented Training Set. We use adversarial domain adaptation to produce an augmented training set, which contains 2,400 pairs of RGB images and transformed disparity images. The performance comparison between using the original training set and using the augmented training set is presented in Sect. 6.4.

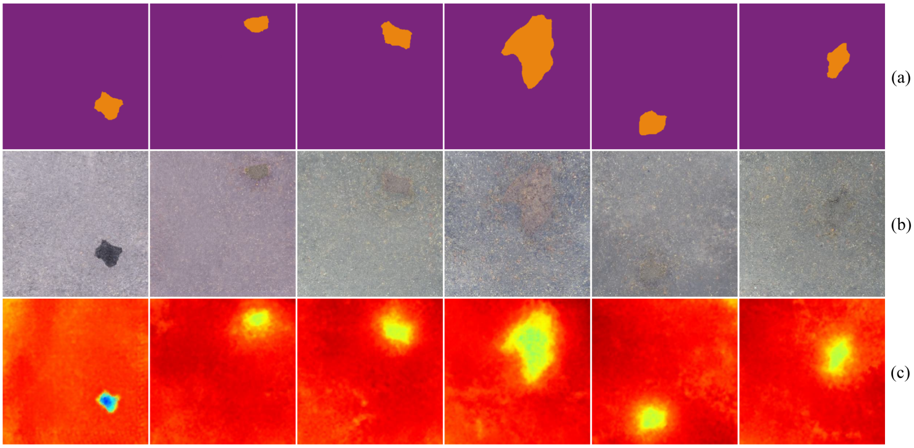


Fig. 5. Examples of training set augmentation results: (a) randomly created pothole detection ground truth; (b) generated RGB images; and (c) generated transformed disparity images.

6.2 Experimental Setup

In our experiments, we first select the architecture of our AA-UNet and AA-RTFNet, as presented in Sect. 6.3. Then, we compare our AA-UNet and AA-RTFNet with eight state-of-the-art (SoA) CNNs (five single-modal ones and three data-fusion ones) for road pothole detection. Each single-modal CNN is trained using RGB images (RGB) and transformed disparity images (T-Disp), respectively; while each data-fusion CNN is trained using RGB and transformed disparity images (RGB+T-Disp). Furthermore, we also select different numbers of RGB images and transformed disparity images from our augmented training set to train the CNNs. The experimental results are presented in Sect. 6.4.

To quantify the performance of these CNNs, we adopt the commonly used F-score (Fsc) and intersection over union (IoU) metrics, and compute their mean values across the testing set, denoted as mFsc and mIoU, respectively. Moreover, the stochastic gradient descent with momentum (SGDM) [36] is used to optimize the CNNs.

6.3 Architecture Selection of AA-UNet and AA-RTFNet

In this subsection, we conduct experiments to select the best architecture for our AA-UNet and AA-RTFNet. All the AA-UNet variants use the same training setups, so do all the AA-RTFNet variants. It should be noted here that $n = 5$ is for both AA-UNet and AA-RTFNet. We also record the inference time of each variant on an NVIDIA GTX 1080Ti graphics card for comparison. (B)–(L) in Table 1 present the effects of a single AM at different network levels. We can see that an AM can bring in better performance improvement when it is added at a higher level, as this can influence the subsequent processes. Moreover, DAM

Table 1. Performances of different AA-UNet variants on the Pothole-600 validation set, where (A) is the U-Net baseline; and (B)–(T) are different variants. Best Results are shown in bold type.

No.	Attention aggregation scheme					Evaluation metrics		
	1st	2nd	3rd	4th	5th	mFsc (%)	mIoU (%)	Runtime (ms)
(A)	–	–	–	–	–	75.9	61.2	**31.3**
(B)	–	–	–	–	DAM	79.7	66.3	33.7
(C)	–	–	–	–	CAM	79.5	66.0	31.5
(D)	–	–	–	–	PAM	79.4	65.9	31.7
(E)	–	–	–	CAM	–	78.7	64.9	31.6
(F)	–	–	–	PAM	–	78.5	64.6	31.9
(G)	–	–	CAM	–	–	78.0	63.9	31.8
(H)	–	–	PAM	–	–	77.7	63.5	32.0
(I)	–	CAM	–	–	–	77.8	63.6	32.1
(J)	–	PAM	–	–	–	77.5	63.2	32.6
(K)	CAM	–	–	–	–	77.6	63.4	32.3
(L)	PAM	–	–	–	–	77.8	63.7	33.5
(M)	–	–	–	CAM	DAM	80.2	66.9	33.8
(N)	–	–	–	PAM	DAM	77.1	62.7	34.0
(O)	–	–	CAM	CAM	DAM	80.7	67.6	33.9
(P)	–	–	PAM	CAM	DAM	77.8	63.6	34.2
(Q)	–	CAM	CAM	CAM	DAM	81.0	68.0	34.1
(R)	–	PAM	CAM	CAM	DAM	79.7	66.2	34.5
(S)	CAM	CAM	CAM	CAM	DAM	81.3	68.5	34.3
(T)	PAM	CAM	CAM	CAM	DAM	**82.6**	**70.3**	34.7

outperforms CAM and PAM at the highest level, since DAM adopts the self-attention scheme, which can achieve better performance, as mentioned above. Furthermore, our CAM performs better than our PAM at higher levels, since feature maps at higher levels have more channels but limited spatial sizes and it is more useful to apply weights on channels. Conversely, feature maps at lower levels have larger spatial sizes but limited channels, and thus it is more useful to adopt our PAM.

Based on these observations, we test the performance of different attention aggregation schemes for our AA-UNet and AA-RTFNet on the validation set, as shown on (M)–(T) in Table 1 and (B)–(J) in Table 2, respectively. We can see that adopting PAM at the lowest network level, adopting DAM at the highest network level, and adopting CAM at other network levels can achieve the best performance for both AA-UNet and AA-RTFNet. Compared with the baseline models, our AA-UNet and AA-RTFNet can increase the mIoU by 9.1% and 5.4%,

Table 2. Performances of different AA-RTFNet variants on the Pothole-600 validation set, where (A) is the RTFNet baseline; and (B)–(J) are different variants. Best Results are shown in bold type.

No.	Attention aggregation scheme					Evaluation metrics		
	1st	2nd	3rd	4th	5th	mFsc (%)	mIoU (%)	Runtime (ms)
(A)	–	–	–	–	–	81.3	68.5	**46.7**
(B)	–	–	–	–	DAM	82.5	70.2	49.1
(C)	–	–	–	CAM	DAM	82.6	70.4	49.2
(D)	–	–	–	PAM	DAM	81.7	69.0	49.4
(E)	–	–	CAM	CAM	DAM	82.8	70.7	49.3
(F)	–	–	PAM	CAM	DAM	81.9	69.3	49.7
(G)	–	CAM	CAM	CAM	DAM	83.4	71.6	49.6
(H)	–	PAM	CAM	CAM	DAM	83.1	71.1	49.9
(I)	CAM	CAM	CAM	CAM	DAM	84.1	72.5	50.0
(J)	PAM	CAM	CAM	CAM	DAM	**85.0**	**73.9**	50.2

respectively, with acceptable extra runtime, which demonstrates the effectiveness and efficiency of our attention aggregation framework.

6.4 Performance Evaluation of Road Pothole Detection

In this subsection, we evaluate the performance of our AA-UNet and AA-RTFNet both qualitatively and quantitatively on the testing set. As mentioned previously, we use different numbers of images selected from the augmented training set to train each CNN. λ denotes the number of samples used in the augmented training set versus the number of samples in the original training set. For example, $\lambda = 2$ means that we train the CNN with $240 \times 2 = 480$ samples randomly selected from the augmented training set. In addition, we introduce a new evaluation metric δ for better comparison. For a given training setup, δ is defined as ratio of the number of iterations for the network to converge using the augmented training set to that of the original training set. $\delta < 1$ means that the training setup converges faster than the baseline setup.

The quantitative results are shown in Fig. 6, where we can clearly observe that the single-modal CNNs with our transformed disparity images as inputs generally perform better than they do with RGB images as inputs, and the mIoU increases by about 17–31%. This is because our transformed disparity images can make the road potholes become highly distinguishable, and can thus benefit all CNNs for road pothole detection. Moreover, we can see that when $\lambda \geq 4$, the CNNs trained with the augmented training set generally outperform themselves when trained with the original training set, and $\delta < 1$ holds in most cases, which demonstrates that adversarial domain adaptation can not only significantly improve pothole detection accuracy but can also accelerate the network convergence. Compared

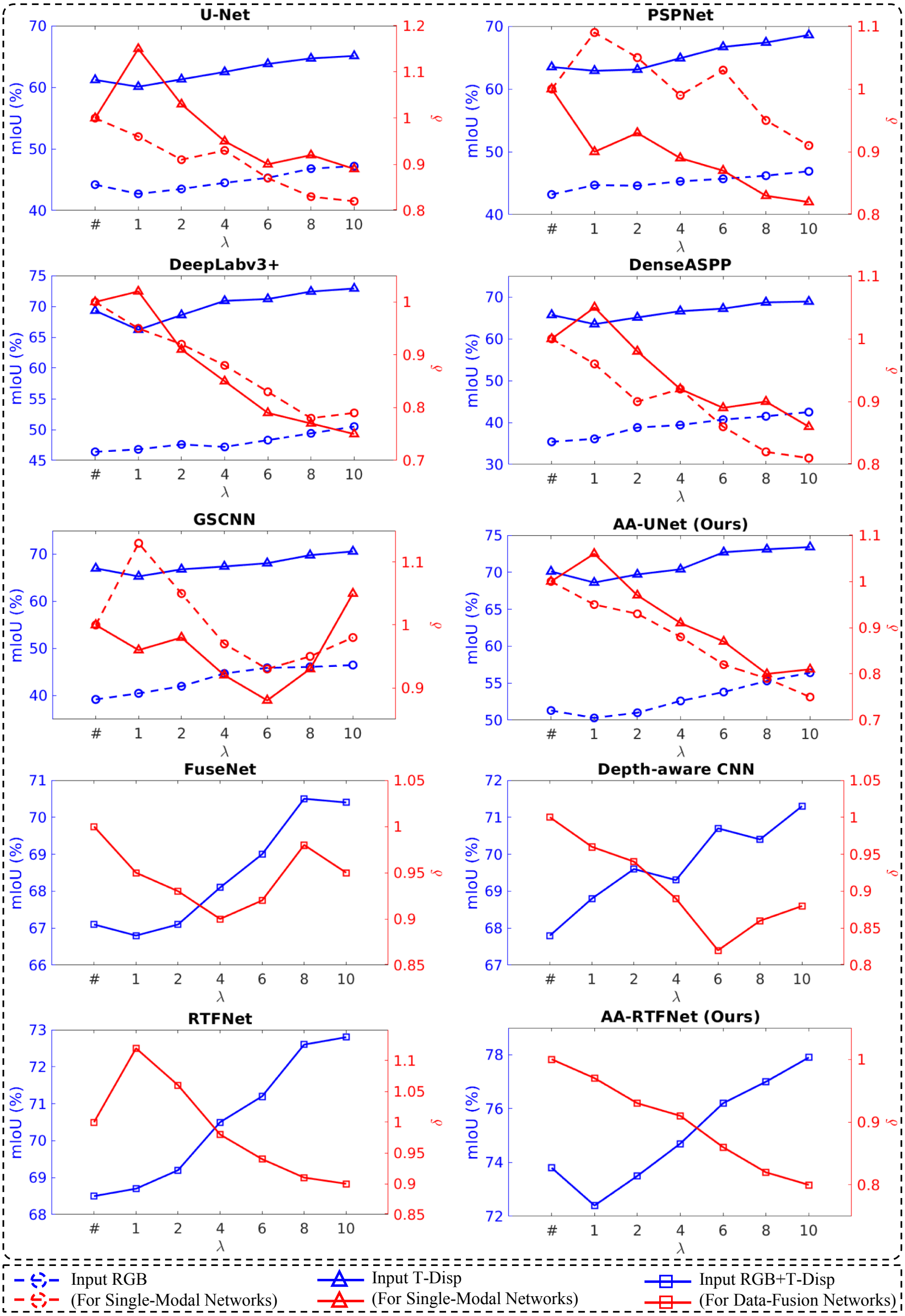


Fig. 6. Performance comparison among eight SoA CNNs, AA-UNet and AA-RTFNet on the Pothole-600 testing set, where the symbol "#" in the λ axis means that we use the original training set in the CNN.

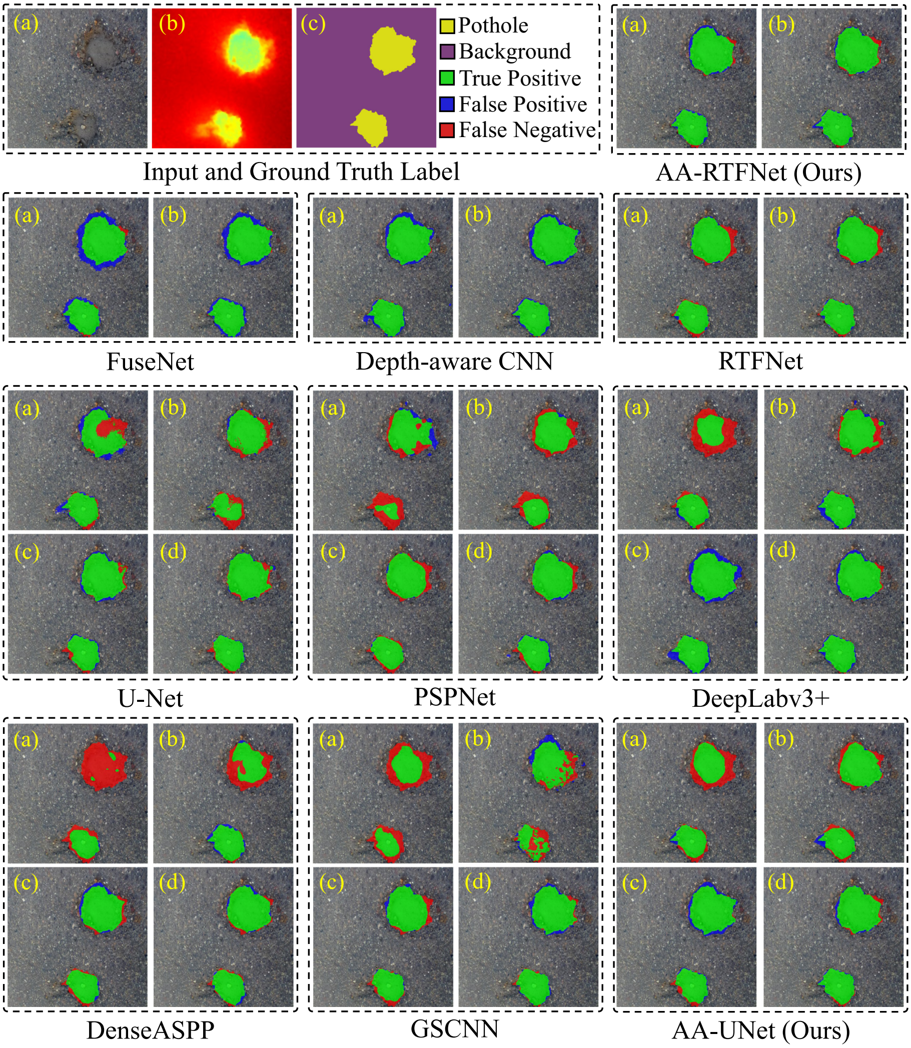


Fig. 7. An example of the experimental results on the Pothole-600 testing set. For the input and ground truth label block: (a) RGB, (b) T-Disp, and (c) ground truth label; For the single-modal network (including U-Net [15], PSPNet [16], DeepLabv3+ [17], DenseASPP [18], GSCNN [19] and our AA-UNet) blocks: (a) input RGB from the original training set, (b) input RGB from the whole augmented training set, (c) input T-Disp from the original training set, and (d) input T-Disp from the whole augmented training set; For the data-fusion network (including FuseNet [22], Depth-aware CNN [23], RTFNet [24] and our AA-RTFNet) blocks: (a) input RGB+T-Disp from the original training set, and (b) input RGB+T-Disp from the whole augmented training set.

with the training setup using the original training set, an increase of around 3–8% is witnessed on the mIoU for the training setup using the whole augmented training set. This is because these two sets share very similar distributions, and our augmented training set possesses an expanded distribution, which can improve road pothole detection performance. In addition, our AA-UNet and AA-RTFNet outperform all other SoA single-modal and data-fusion networks for road pothole detection, respectively, which strongly validates the effectiveness and efficiency of our attention aggregation framework. Readers can see that our AA-UNet can increase the mIoU by approximately 3–14% compared with the SoA single-modal networks, and our AA-RTFNet can increase the mIoU by about 5–8% compared with the SoA data-fusion networks. The qualitative results shown in Fig. 7 can also confirm the superiority of our proposed approaches.

7 Conclusion

The major contributions of this paper include: a) a novel attention aggregation framework, which can help the CNNs focus more on salient objects, such as road potholes, so as to improve semantic segmentation for better pothole detection results; b) a novel training set augmentation technique developed based on adversarial domain adaptation, which can produce more synthetic road RGB images and their corresponding transformed road disparity (or inverse depth) images to improve both the efficiency and accuracy of CNN training; c) a large-scale road pothole detection dataset, publicly available at http://sites.google.com/view/pothole-600 for research purposes. The experimental results validated the effectiveness and feasibility of our proposed attention aggregation framework and the training set augmentation technique for enhancing road pothole detection. Moreover, we believe our proposed techniques can also be used for many other semantic segmentation applications, such as freespace detection.

Acknowledgements. This work was supported by the National Natural Science Foundation of China, under grant No. U1713211, Collaborative Research Fund by Research Grants Council Hong Kong, under Project No. C4063-18G, and the Research Grant Council of Hong Kong SAR Government, China, under Project No. 11210017, awarded to Prof. Ming Liu.

References

1. Mathavan, S., Kamal, K., Rahman, M.: A review of three-dimensional imaging technologies for pavement distress detection and measurements. IEEE Trans. Intell. Transp. Syst. **16**(5), 2353–2362 (2015)
2. Fan, R., Ozgunalp, U., Hosking, B., Liu, M., Pitas, I.: Pothole detection based on disparity transformation and road surface modeling. IEEE Trans. Image Process. **29**, 897–908 (2019)
3. Koch, C., Georgieva, K., Kasireddy, V., Akinci, B., Fieguth, P.: A review on computer vision based defect detection and condition assessment of concrete and asphalt civil infrastructure. Adv. Eng. Inf. **29**(2), 196–210 (2015)

4. Majendie, M.: Dani king: ‘it was just a freak accident but I thought I was going to die’. Technical report, Independent, June 2015
5. Fan, R., Liu, M.: Road damage detection based on unsupervised disparity map segmentation. IEEE Trans. Intell. Transp. Syst. **21**, 4906–4911 (2019)
6. Devine, R.: City of San Diego asking residents to report potholes. Technical report, NBC San Diego, January 2017
7. News, B.: Government to pledge billions for filling potholes. Technical report, BBC News, March 2020
8. Fan, R., Jiao, J., Pan, J., Huang, H., Shen, S., Liu, M.: Real-time dense stereo embedded in a UAV for road inspection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 535–543. IEEE (2019)
9. Leo, M., Furnari, A., Medioni, G.G., Trivedi, M., Farinella, G.M.: Deep learning for assistive computer vision. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 0–0, (2018)
10. Rover, J.L.: Pothole detection technology research announced by jaguar land rover
11. Baraniuk, C.: Ford developing pothole alert system for drivers, February 2017
12. Jahanshahi, M.R., Jazizadeh, F., Masri, S.F., Becerik-Gerber, B.: Unsupervised approach for autonomous pavement-defect detection and quantification using an inexpensive depth sensor. J. Comput. Civ. Eng. **27**(6), 743–754 (2013)
13. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. **9**(1), 62–66 (1979)
14. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
15. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
16. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890 (2017)
17. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818 (2018)
18. Yang, M., Yu, K., Zhang, C., Li, Z., Yang, K.: DenseASPP for semantic segmentation in street scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3684–3692 (2018)
19. Takikawa, T., Acuna, D., Jampani, V., Fidler, S.: Gated-SCNN: gated shape CNNs for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5229–5238 (2019)
20. Wang, H., Fan, R., Sun, Y., Liu, M.: Applying surface normal information in drivable area and road anomaly detection for ground mobile robots. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2020). (To be published)
21. Fan, R., Wang, H., Cai, P., Liu, M.: SNE-RoadSeg: incorporating surface normal information into semantic segmentation for accurate freespace detection. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12375, pp. 340–356. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58577-8_21

22. Hazirbas, C., Ma, L., Domokos, C., Cremers, D.: FuseNet: incorporating depth into semantic segmentation via fusion-based CNN architecture. In: Lai, S.-H., Lepetit, V., Nishino, K., Sato, Y. (eds.) ACCV 2016. LNCS, vol. 10111, pp. 213–228. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-54181-5_14
23. Wang, W., Neumann, U.: Depth-aware CNN for RGB-D segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 135–150 (2018)
24. Sun, Y., Zuo, W., Liu, M.: RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes. IEEE Robot. Autom. Lett. **4**(3), 2576–2583 (2019)
25. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
26. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
27. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018)
28. Yuan, Y., Wang, J.: OCNet: object context network for scene parsing. arXiv preprint arXiv:1809.00916 (2018)
29. Fu, J., et al.: Dual attention network for scene segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3146–3154 (2019)
30. Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W.: CCNet: criss-cross attention for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 603–612 (2019)
31. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
32. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)
33. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
34. Fan, R., Ai, X., Dahnoun, N.: Road surface 3D reconstruction based on dense subpixel disparity map estimation. IEEE Trans. Image Process. **27**(6), 3025–3035 (2018)
35. Labayrade, R., Aubert, D.: A single framework for vehicle roll, pitch, yaw estimation and obstacles detection by stereovision. In: IEEE IV2003 Intelligent Vehicles Symposium. Proceedings (Cat. No. 03TH8683), pp. 31–36. IEEE (2003)
36. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)