

ORIGINAL ARTICLE

Open Access



An uncertainty-aware domain adaptive semantic segmentation framework

Huilin Yin¹, Pengyu Wang¹, Boyu Liu¹ and Jun Yan^{1*}

Abstract

Semantic segmentation is significant to realize the scene understanding of autonomous driving. Due to the lack of annotated real-world data, the technology of domain adaptation is applied so that the model is trained on the synthetic data and inferred on the real data. However, this domain gap leads to aleatoric and epistemic uncertainty. These uncertainties link to the potential safety issue of autonomous driving in normal weather and adverse weather. In this study, we explore the scientific problem that has received sparse attention previously. We postulate that the Dual Attention module can mitigate the uncertainty in the task of semantic segmentation and provide some empirical study to validate it. Furthermore, the utilization of Kullback-Leibler divergence (KL divergence) helps the estimation of aleatoric uncertainty and boosts the robustness of the segmentation model. Our empirical study on the diverse datasets of semantic segmentation demonstrates the effectiveness of our method in normal and adverse weather. Our code is available at: <https://github.com/liubo629/Seg-Uncertainty-dual-attention>.

Keywords: Deep learning, Semantic segmentation, Domain adaptation, Uncertainty

1 Introduction

Scene understanding is a necessary requisite for autonomous vehicles. Semantic segmentation, a task based on pixel-level image classification [1], is an effective approach for scene understanding. Partitioning images or video frames into multiple segments or objects is important to the operation of autonomous vehicles to search the navigable surface and detect pedestrians. With the help of high-level semantic segmentation technology, accurate and reliable road conditions can be caught to support the safe operation of autonomous vehicles. In the past decade, deep learning technologies have been applied to semantic segmentation with remarkable performance [2–10]. The release of datasets and the maintenance of benchmarks also boost the research of semantic segmentation in the scenario of autonomous driving [11, 12].

However, the achievement of semantic segmentation is at the price of a large amount of dense pixel-level an-

notations obtained by expensive labor cost [13]. To mitigate such a problem, domain adaptation techniques have been proposed to bridge the gap between source and target domains [14]. The source domain has a large number of annotations, and the target domain has no (or very few) annotations. Domain adaptation aims at utilizing the source domain data and the target domain data to adapt the segmentation model to the target environment. It is done by using maximum mean discrepancies (MMD) or adversarial training [15] to minimize the difference between the distributions of the intermediate features. Other approaches include the data reconstruction of source and target samples [16] or self-training to provide pseudo labels [17].

Although current domain adaptation methods have achieved certain results, they are dominated by some level of uncertainty [18, 19]. Previous work focused on the uncertainty of semantic segmentation that mainly solves the problem existed in the scenario without domain shift [20–23]. Therefore, it is exploratory to utilize the uncertainty measurement to improve the generalization ability of cross-domain semantic segmentation. Furthermore,

*Correspondence: yanjun@tongji.edu.cn

¹School of Electronic and Information Engineering, Tongji University, Shanghai, 201804, China

the mainstream semantic segmentation methods are designed and evaluated for the operation at daytime. However, in adverse weather conditions such as snow, rain, fog, or night, the uncertainty measurement of the domain adaptation method and the correction based on uncertainty are still rarely studied. To realize the Safety of the Intended Functionality (SOTIF) in autonomous driving [24], it is essential to study the recognition performance of the perception model in both normal weather and adverse weather. In such a process, it is essential to quantify the uncertainty for the amelioration of SOTIF.

In this paper, we propose a segmentation network with a Dual Attention mechanism [25] to realize the considerable performance of pixel recognition. In this way, we can accurately capture the uncertain parts of the segmentation results and ignore them. Empirical study has shown our postulation that such an attention mechanism helps mitigate the uncertainty of the segmentation model. In addition, we propose an uncertainty indicator based on KL divergence derived from information theory [26] to estimate the uncertainty of the prediction results more accurately. For the remission of epistemic uncertainty of the semantic segmentation model based on domain adaptation, the method of Monte Carlo Dropout (MC-Dropout) [27] is applied. The uncertainty-based domain adaptation framework is illustrated in Fig. 1. Our framework of semantic segmentation model based on domain adaptation realizes considerable generalization ability in normal weather and adverse weather.

The main contributions are summarized as follows:

- We propose a novel segmentation network with the Dual Attention mechanism to catch the uncertain parts of the model with the promotion of the performance of semantic segmentation.
- We design an uncertainty indicator based on KL divergence to estimate the uncertainty of the prediction results.
- We verify the proposed methods on two transfer learning tasks, i.e., GTA5 (Grand Theft Auto 5) dataset [28] → Cityscapes dataset [11] and GTA5 dataset → ACDC (the Adverse Conditions Dataset with Correspondences) dataset [29]. The proposed method has achieved competitive performance to the existing methods in the domain adaptation task of GTA5 → Cityscapes. Moreover, it is validated in the domain adaptation task GTA5 → ACDC that the consideration of uncertainty can reduce the negative impact of adverse weather on the segmentation results.

This paper is organized as follows: Sect. 2 gives a literature review. Section 3 presents the details of our proposed method. Training details and experiment results are presented in Sect. 4. Section 5 gives a conclusion and outlook on our research.

2 Related works

2.1 Semantic segmentation adaptation

The goal of traditional machine learning is to learn a model on a set of training samples to find the most appropriate

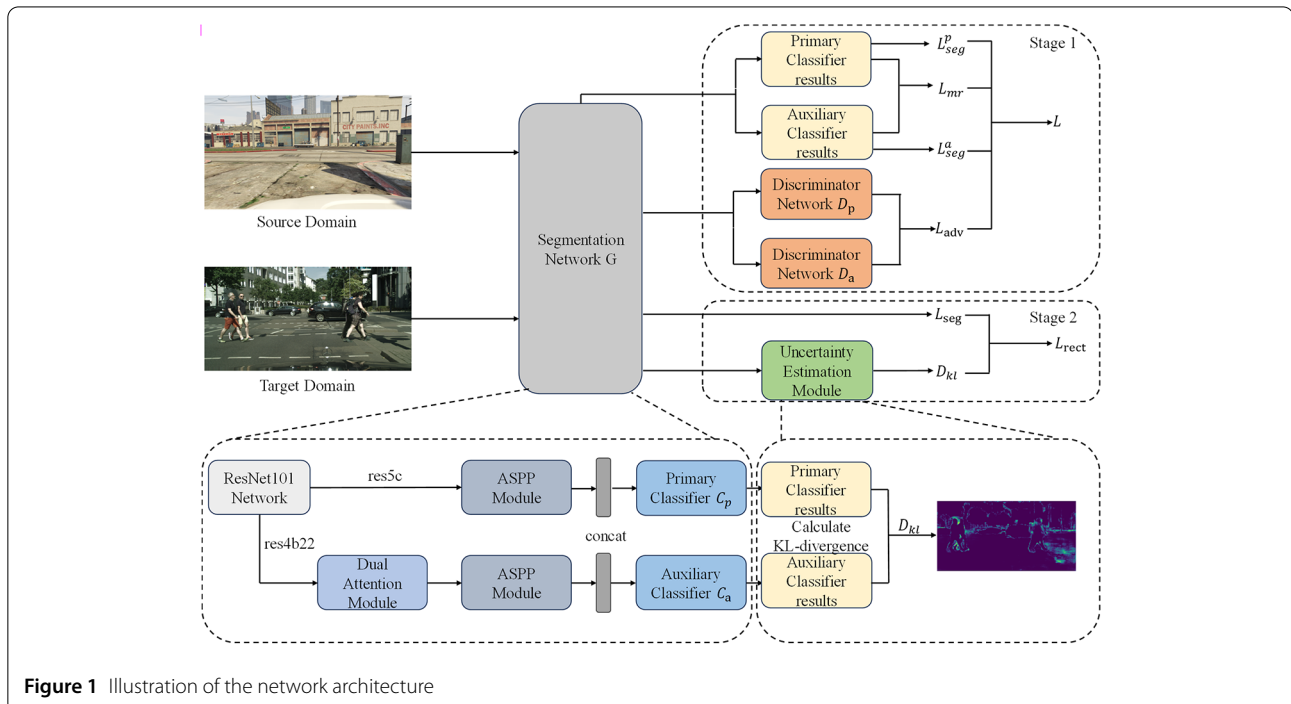


Figure 1 Illustration of the network architecture

objective function for unseen test data. However, to train such a model, it assumes that both the training and test data come from the same distribution and share a similar joint probability distribution. This condition is difficult to achieve in real-world applications because training and test sets often come from different feature spaces or distributions [14]. It is a laborious and time-consuming task to generally create new labeled datasets as training datasets, rather than using relevant publicly available annotated data.

Domain Adaptation is a transfer learning method in which the source and target tasks are the same, but the data distributions in the source and target domains are different. It has two key elements: source domain and target domain. The source domain has a large number of labeled samples, while the target domain has no (or very few) labeled samples. Domain adaptation typically learns a model from one source domain data and generalizes the model to the target domain by minimizing the difference between domain distributions [14]. The domain adaptation problem can be resolved with the generative model to align the distribution [30, 31], disentangled information reconstruction [32–34], or statistical method to measure the domain distribution gap [35, 36].

In the task of semantic segmentation, some pioneering works [15, 37, 38] are proposed to reduce the domain gap through adversarial learning. In this way, the model could be trained on the labeled data with the target style. Besides, some recent studies leverage geometric information [39, 40] to enhance the results of domain adaptation. Furthermore, some methods [13, 41] further improve results by enhancing the alignment of categorical features.

The above methods are all based on normal weather images as target domain data. However, the system in autonomous driving needs stable and considerable performance even in adverse weather conditions [42, 43]. Therefore, the domain adaptation task should guarantee accurate segmentation performance not only in normal weather but also in abnormal weather. Diverse methods could be a feasible method to alleviate this problem including the exploitation of a visual inductive prior derived from physics-based reflection models for domain adaptation [44] or leveraging the labeled images from the target domain to improve the accuracy by model distillation [45]. The curriculum model adaptation method is utilized to learn from both bad weather and normal condition images [46, 47]. Access to ground-truth annotations under adverse conditions is indispensable for achieving high performance [29]. There is still huge space for improvement in the performance of the current method for abnormal weather. In such a scenario, uncertainty is linked to the reliability of the results, which deserves further study.

2.2 Visual attention method

Conceptually, the attention mechanism can be understood as selecting the key parts of the visual field. The attention mechanisms are usually categorized into two mainstream types: spatial and channel attention. Spatial attention mechanisms focus on critical parts of the task. For example, Spatial Transformer Networks [48] (STN) is a classical spatial attention model that localizes the target location and performs the feature representation operation at the spatial level. Channel attention mechanisms focus on the interactions of features between different channels. Typically, Squeeze-and-Excitation Networks [49] (SENet) improve the quality of feature extraction by explicitly modeling inter-channel relationships.

In addition, the research community has proposed the self-attention mechanism that enhances the perception field and context-awareness of the network by capturing the global information, i.e., the interrelationships among different parts of the entire input feature graph. For example, Dual Attention Network [25] (DANet) combines spatial and channel attention to map the global contextual information by extending the local features generated by residual networks to extract more relevant features.

With such a mechanism, the visual system can allocate visual processing resources to these visually significant areas [50]. The neural computation on the salient part of the visual object helps guide the high-level vision tasks including image classification, object detection [51], semantic segmentation [25], image generator [52], video understanding [53], and medical vision task [54].

In this study, we will explore the relationship between the attention mechanism and the cross-domain semantic segmentation task. We postulate that the model with the attention mechanism can capture the uncertain part of the prediction result more accurately.

2.3 Uncertainty estimation

For the evaluation of unexpected behavior of SOTIF risk [24], it is essential to study the triggering conditions and performance limitations of the autonomous driving environment perception system. The trigger condition corresponds to the environmental sensor part of the perception system, which can be specific to the driving scenarios that can cause autonomous driving safety accidents. The performance limitations include the limitations of the perception sensor and the limitations of the perception algorithm. In this study, we would mainly focus on the limitations of the perception algorithm so that the uncertainty of neural visual systems in autonomous driving should be quantified [55].

The uncertainty of neural networks can be divided into aleatoric uncertainty which links to the uncertainty of the data distributions and epistemic uncertainty which links to the uncertainty of the model parameters [20]. For the

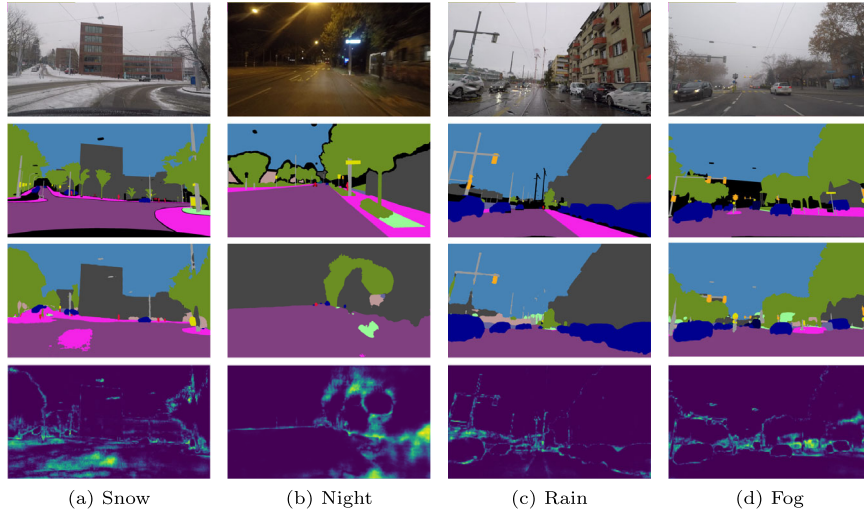


Figure 2 Uncertainty under abnormal conditions

domain-adaptation task, there exists a cross-domain uncertainty [56].

The Bayesian theorem [57] can measure the subjective confidence in the statics [58]. It can be utilized in the posterior inference, parameter estimation, probabilistic latent variable model, and other machine learning tasks [59]. After the renaissance of deep learning, the Bayesian method is utilized to evaluate the uncertainty that exists in the training and inference of deep neural networks. Diverse methods are proposed and applied in the concrete applications including MC-Dropout [20, 60–62], Deep Ensemble [21, 63], Direct Modeling [64–66], and Error Propagation [67].

The uncertainty estimation and calibration technologies have been applied in the tasks of semantic segmentation of autonomous driving [20–23]. However, the previous study does not pay attention to the scenarios of abnormal weathers in which the probability of misclassification would greatly increase [29]. As illustrated in Fig. 2, the first row presents images captured under snow, night, fog, and rain conditions, while the second and third rows depict their corresponding labeled and predicted graphs. It is evident that both feature extraction and model accuracy are significantly impacted. Subsequently, heatmap visualization is employed to depict the model's uncertainty. Notably, heatmaps enable the spatial distribution of data to be visualized through color changes, facilitating clear identification and comparison of uncertainty levels across different regions. The heatmap corresponding to the image is displayed in the fourth row. Typically, uncertainty concentrates at the edges of segmented objects as they demarcate different regions. However, examination of the fourth-row heatmap reveals that, in abnormal weather conditions, uncertainty is not solely concentrated at object edges; rather,

many objects exhibit fuzzy and indistinct boundaries. Additionally, areas devoid of objects sometimes feature concentrated uncertainty, greatly diminishing model recognition accuracy. Hence, mitigating the impact of uncertainty on autonomous perception systems constitutes a crucial challenge warranting further investigation.

The research community terms such a phenomenon as a challenge for the task of semantic segmentation [46]. Moreover, the uncertainty would be more critical in the cross-domain task than in the inside-domain task. Thus, this study will explore the uncertainty of cross-domain semantic segmentation in both normal weather and adverse weather.

3 Method

3.1 Problem settings

We focus on the uncertainty estimation problem of unsupervised domain adaptation in semantic segmentation, where both unlabeled target domain data and labeled source domain data would be handled:

$$D_T = \{x_t^j\}_{j=1}^N, \quad (1)$$

$$D_S = \{x_s^i | y_s^i\}_{i=1}^M. \quad (2)$$

In Eq. (1) and Eq. (2), the symbols M and N denote the numbers of samples in the source and target domains, respectively. The symbol x_s^i represents the i -th sample data in the source domain, and y_s^i is its label. Correspondingly, the symbol x_t^j represents the j -th sample data in the target domain. The aim is to generalize a model for the target domain by minimizing the difference between domain distributions [14]. However, it is difficult for current methods to eliminate the gap between the two domains.

The uncertainty of the generated results is salient. Such prediction results are unreliable, which can lead to safety incidents. Therefore, it is necessary to quantify the uncertainty. Then, we utilize a discriminator to distinguish whether the input is from the source or target domain and gain the adversarial loss.

3.2 Network architecture

Our network architecture is illustrated in Fig. 1. It is composed of one generator G and two discriminators D_p and D_a . The baseline model of G is Deeplabv3+ [8], which adopts the ResNet101 [68] as the backbone model. The generator has one feature extractor and two classifier modules, each of which contains one Atrous Spatial Pyramid Pooling (ASPP) module [6]. The auxiliary classifier is connected after the res4b22 layer, and there is a Dual Attention module [25] before the ASPP module. The Dual Attention module, including a spatial attention part and a channel attention part, can enhance the discriminant ability of uncertain feature representations for scene segmentation. Besides, we use pseudo-label learning to make the model learn the dataset more adequately, making the results better. The model training is divided into two stages:

Stage 1. Firstly, we input a source domain sample x_s and its corresponding labels y_s into the segmentation network to extract the feature map of the image, and then the feature map is recovered to the same size as the input image by up-sampling through the inverse convolutional layer. Secondly, we input it into two classifiers C_p and C_a , and predict the category of each pixel point to get the pixel-level label $F_p(x_s)$ and $F_a(x_s)$. In Eq. (3) and Eq. (4), F_p and F_a are the segmentation functions, where F_p maps the input samples to the predictions of the primary classifier and F_a maps the input samples to the predictions of the auxiliary classifier. Then, we input a target domain sample x_t . After this step, the network gets the pixel-level labels $F_p(x_t)$ and $F_a(x_t)$. Thirdly, we learn source domain knowledge from labeled source domain data by calculating the cross-entropy loss of the primary classifier and auxiliary classifier separately as described by Eq. (3) and Eq. (4):

$$L_{\text{seg}}^p = - \sum_{h=1}^H \sum_{w=1}^W \sum_{c=1}^C y_s^i \log(F_p(x_s^i)), \quad (3)$$

$$L_{\text{seg}}^a = - \sum_{h=1}^H \sum_{w=1}^W \sum_{c=1}^C y_s^i \log(F_a(x_s^i)), \quad (4)$$

where H , W , and C represent the height, width, and the number of categories of the image, respectively. Since the segmentation loss does not involve data from the target domain, it is still necessary to introduce the adversarial loss in order to minimize the difference between the source and target domain distributions. Thus, we input the results

into two discriminators D_p and D_a to get the adversarial loss defined in Eq. (5) and Eq. (6):

$$L_{\text{adv}}^p = E[\log(D_p(F_p(x_t^i))) + \log(1 - D_p(F_p(x_t^i)))], \quad (5)$$

$$L_{\text{adv}}^a = E[\log(D_a(F_a(x_t^i))) + \log(1 - D_a(F_a(x_t^i)))]. \quad (6)$$

Finally, we deduce the variance sum of $F_p(x_t)$ and $F_a(x_t)$ as the memory regularization loss [69] defined in Eq. (7):

$$L_{\text{mr}} = - \sum_{h=1}^H \sum_{w=1}^W \sum_{c=1}^C F_a(x_t^i) \log(F_p(x_t^i)) - \sum_{h=1}^H \sum_{w=1}^W \sum_{c=1}^C F_p(x_t^i) \log(F_a(x_t^i)). \quad (7)$$

Calculating L_{mr} is to compute the KL divergence on the predictions of the two classifiers and its purpose is to reduce the gap between the predictions of the two classifiers. Considering that the auxiliary classifiers use shallower features as inputs, so in this study, we give the segmentation loss obtained by the auxiliary classifiers L_{seg}^a and the adversarial loss L_{adv}^a multiplied by a smaller weight λ_{seg} and λ_{adv} . A weight λ_{mr} is also given to L_{mr} , the final L_{seg}^a and L_{adv}^a are defined in Eq. (8) and Eq. (9), and Stage 1 final loss function is defined in Eq. (10):

$$L_{\text{seg}} = L_{\text{seg}}^p + \lambda_{\text{seg}} L_{\text{seg}}^a, \quad (8)$$

$$L_{\text{adv}} = \lambda_{\text{adv}}^p L_{\text{adv}}^p + \lambda_{\text{adv}}^a L_{\text{adv}}^a, \quad (9)$$

$$L = L_{\text{seg}} + L_{\text{adv}} + \lambda_{\text{mr}} L_{\text{mr}}, \quad (10)$$

where $\lambda_{\text{seg}} = 0.5$, $\lambda_{\text{adv}}^p = 0.001$, $\lambda_{\text{adv}}^a = 0.0002$, $\lambda_{\text{mr}} = 0.1$ [69].

Stage 2. Stage 1 trained model is used as a pre-trained model in Stage 2 while pseudo labels of the target dataset are generated using this model, and we introduced uncertainty estimation in the loss function in Stage 2. In addition, we used the uncertainty of the obtained prediction results to modify the segmentation loss. So that the impact of poor quality pseudo labels on the results can be reduced effectively. Therefore, we have the following segmentation loss L defined in Eq. (11) to get the rectified loss L_{rect} :

$$L_{\text{rect}} = E[\exp(-D_{\text{kl}}) L_{\text{seg}}^p + D_{\text{kl}}]. \quad (11)$$

Since the computation of the loss function for Stage 2 involves uncertainty estimation, the derivation of this loss function will be demonstrated in detail in Subsect. 3.3. The final prediction $\hat{y}_t^j$ of the model is the combined output of the two classifiers, which is defined in Eq. (12), and the $\hat{y}_t^j$ is the pseudo label:

$$\hat{y}_t^j = \text{argmax}(F_p(x_t^i) + 0.5F_a(x_t^i)). \quad (12)$$

3.3 Uncertainty estimation

Bayesian neural networks and deep ensembles are commonly used methods for estimating the uncertainty of a semantic segmentation model. However, these methods require excessive memory and time, prompting the need for more lightweight alternatives. In this study, we propose the adoption of a segmentation model with two different classifiers, which allows for the acquisition of multi-level information. When the generated pseudo-label data closely resembles the target domain data, the classified regions with differences between the two domains can be considered uncertain, and larger differences indicate higher uncertainty.

Additionally, the inclusion of an auxiliary classifier serves two primary purposes. Firstly, it mitigates the issue of gradient vanishing by involving hidden units and intermediate layers of the segmentation network during training, thus preventing overfitting. Secondly, the auxiliary classifier has fewer layers compared to the primary classifier and incorporates the Dual Attention Module so that features containing more spatial, positional, and detailed information can be extracted. Consequently, integrating an auxiliary classifier provides a comprehensive representation of the obtained features, offering additional regularization and amplifying the gradient signal during backpropagation.

In our study, we utilize the output results from “res5c” and “res4b22” for uncertainty estimation. The “res4b22” layer, situated within the fourth residual block of ResNet-101, represents a deeper layer that strikes a balance between spatial resolution and semantic information. The “res5c” layer typically serves as the final layer in the fifth residual block of ResNet-101, operating at the highest level of abstraction and integrating the most comprehensive and discriminative features extracted by the network, this layer’s outputs are often used for making the final predictions. The selection of the outputs from “res4b22” and “res5c” for uncertainty analysis ensures that the model can handle and integrate rich features while providing different perspectives for model learning—one closer to the output and the other slightly earlier in the network. Discrepancies between the predictions of these two stages indicate regions of low confidence or significant divergence in the network’s predictions, reflecting prediction uncertainty. Such disparities are crucial for identifying unreliable pseudo-labels or erroneous predictions during training, thereby enhancing the model’s reliability. In other words, this approach by exploiting the different properties of the “res5c” and “res4b22” layers in ResNet-101, the segmentation model can effectively utilize the internal differences as signals of uncertainty, thus enhancing its ability to learn from new domains of unlabelled data and improving its overall robustness and accuracy.

For a given input sample, the main classifier and auxiliary classifier produce distinct prediction results. The prediction variance $\text{Var}(p_t)$ when training with pseudo labels

can be obtained can be approximated as Eq. (13):

$$\text{Var}(p_t) \approx E[(F(x_t^j|\theta_t) - \hat{p}_t^j)^2], \quad (13)$$

where $F(x_t^j|\theta_t)$ is the prediction of the input target domain sample x_t^j , $\hat{p}_t^j$ is the truth class probability of the corresponding pseudo label. Because when applying the pseudo-labeling method to the task of domain adaptive semantic segmentation, the goal of training is to reduce the deviation of the prediction results from the pseudo-labeling, so that the prediction variance will also become smaller and smaller, it is not suitable for estimating the uncertainty of the pseudo-labeling in the training. Therefore, this study utilizes the auxiliary classifier in the segmentation model to estimate the uncertainty of pseudo-labels, the prediction variance $\text{Var}(p_t)$ is defined in Eq. (14):

$$\text{Var}(p_t) \approx E[(F_p(x_t^j|\theta_t) - F_a(x_t^j|\theta_t))^2], \quad (14)$$

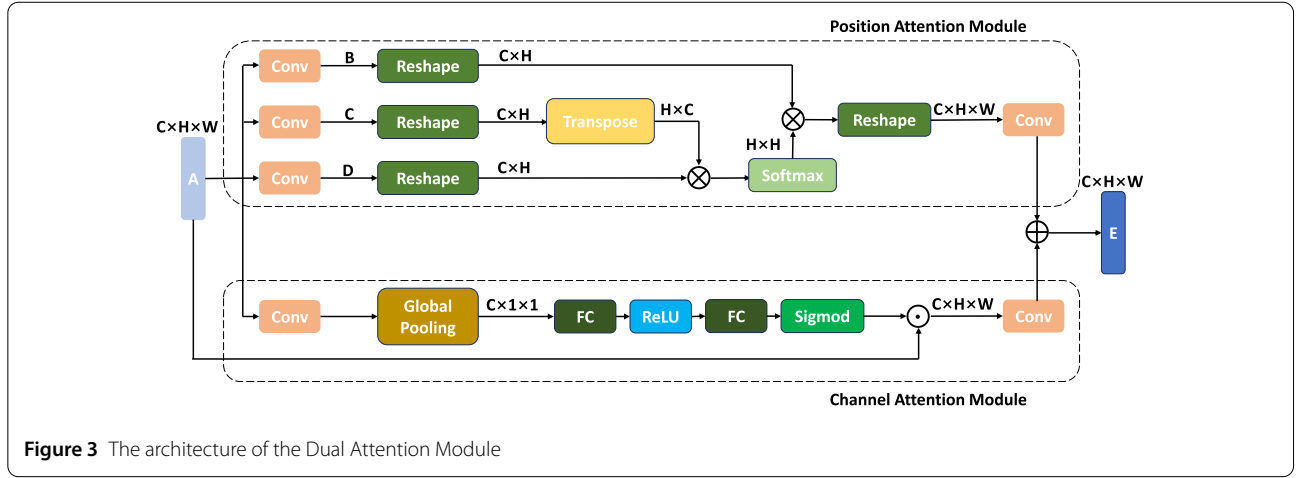
where $F_a(x_t^j|\theta_t)$ denotes the auxiliary classifier output of the segmentation model. In this study, instead of returning the variance of the prediction results of the two classifiers directly to the network as a result of the loss function, we involve it as a regularization term of the segmentation loss L_{seg} in the training of the network. Whereas, the KL distance between the prediction results of the two classifiers is used as a measure of uncertainty. The KL divergence D_{kl} can be written as Eq. (15):

$$D_{\text{kl}} = E\left[F_p(x_t^j|\theta_t) \log\left(\frac{F_p(x_t^j|\theta_t)}{F_a(x_t^j|\theta_t)}\right)\right]. \quad (15)$$

We estimate the uncertainty of the pseudo labels using the predicted variance of the two classifiers, which corresponds to a regularization term used to correct for the noise of the pseudo labels and the loss function of the objective can be expressed as Eq. (16):

$$L_{\text{rect}} = E\left[\frac{1}{\text{Var}(p_t)} \text{Bias}(p_t) + \text{Var}(p_t)\right], \quad (16)$$

where $\text{Var}(p_t)$ is the pseudo-label training prediction variance, and $\text{Bias}(p_t)$ is the deviation of the labeling obtained from the network prediction from the true value of the corresponding target domain. When the prediction variance is large it means that the difference between the results obtained by the two classifiers is large, which means that this pseudo label has a large uncertainty and is unreliable, so to avoid over-penalizing the prediction $\text{Bias}(p_t)$ because of these unreliable samples, the loss function is designed to return a very small value, in this case, to reduce the negative impact on the model training. However, by this way the model becomes inclined to predict samples that have large $\text{Var}(p_t)$, so $\text{Var}(p_t)$ is added to the loss function to introduce



a regular term. In addition, because the prediction variance $\text{Var}(p_t)$ may be close to 0, to prevent the problem of dividing by 0 and to stabilize the training, $1/\text{Var}(p_t)$ in the loss function was replaced as $\exp(-\text{Var}(p_t))$, and then we replace $\text{Var}(p_t)$ as D_{kl} , which is finally combined as a regularization term with the segmentation loss L_{seg} of the network, and the final loss function is expressed as Eq. (11).

3.4 Dual attention module

We introduce a Dual Attention module into the segmentation network, which can capture the global correlation between features in the spatial and channel dimensions of the features, respectively, and enhance the expression ability of the features. The method consists of two modules, the Spatial Attention Module, and the Channel Attention Module. Both of them can be flexibly added to the convolutional layer of ResNet. The structure of the dual-channel attention module is shown in Fig. 3.

The structure of the spatial attention module is shown in Fig. 3. The feature map A is the input of the module, which is sent to the three convolutional layers to obtain the feature maps B, C, D , these three convolutional layers have the same number of channels of the input and output feature maps, and the convolution kernel (kernel size) is set to 3×3 , the step size (stride) is set to 1, and the image padding (padding) is set to 1, and finally the dilated convolution parameter (dilation) is also set to 1, which means no dilated convolution is used. According to the output image size of the convolution layer computation mechanism, we deduce Eq. (17) and Eq. (18). the feature maps with size $C \times H \times W$ will still have the same size $C \times H \times W$ after such a convolution layer. After that, a BN layer and a ReLU layer are connected after the convolutional layer, and a nonlinear activation function layer is connected after the two convolutional layers to avoid the degradation of the two convolutional layers into one layer and to increase the nonlinear

expression of the network.

$$H_{\text{out}} = \left\lceil \frac{H_{\text{in}} + 2 \times \text{padding} - \text{dilation} \times (\text{kernelsize}[0] - 1) - 1}{\text{stride}} + 1 \right\rceil, \quad (17)$$

$$W_{\text{out}} = \left\lceil \frac{W_{\text{in}} + 2 \times \text{padding} - \text{dilation} \times (\text{kernelsize}[1] - 1) - 1}{\text{stride}} + 1 \right\rceil. \quad (18)$$

In Eq. (17) and Eq. (18), the symbols H_{in} , W_{in} , H_{out} and W_{out} are the feature maps of the input convolutional layers and the output respectively. The width and height of the feature map. Then, the matrices C and D are transformed into $C \times N$, $N = H \times W$ and the matrix obtained after multiplying the transpose of C and D is softmax by rows to obtain the spatial attention map S , whose dimension is $N \times N$, and the formula for S is defined in Eq. (19):

$$s_{ji} = \frac{\exp(B_i \cdot C_j)}{\sum_{i=1}^N \exp(B_i \cdot C_j)}, \quad (19)$$

where the symbol s_{ji} represents the element of row j , column i in S , representing the influence of the feature at position i on position j . If the features at both positions belong to the same category, then the value of s_{ji} will be relatively large. After that, the dimension of B is resized to $C \times N$, which is a matrix multiplied by the transpose of S . The result is multiplied by a Dual Attention parameter α , which is used to control the effect of the spatial attention module. The final obtained output E is given by the following equation defined in Eq. (20):

$$E_j = \alpha \sum_{i=1}^N (s_{ji} D_i) + A_j. \quad (20)$$

Our channel attention module uses SENet, which is a feature map of size $[C, 1, 1]$ obtained by pooling feature maps of size $[C, H, W]$ extracted from a single image by the network. The resulting feature map is fed into the MLP network composed of a fully connected layer, ReLU

activation function layer, fully connected layer, and Sigmoid activation function layer to get the feature of size $[C, 1, 1]$. It is equivalent to giving each of the C channels a weight to multiply the feature with the input pooling. The large weight denotes the important feature map of the corresponding channel, which makes the network focus on the interpretable feature map. Compared with the traditional channel attention module, SENet performs well and is more memory-efficient.

Compared with the Dual Attention mechanism [25], we replace the channel attention module in the Dual Attention module with a lighter Attention with Squeezing and Excitation modules of the same kind [49]. Our approach employs fewer parameters and achieves better accuracy. Simultaneously, we introduce a spatial attention module into our attention mechanism, allowing for a more focused consideration of important regions within the image, thereby enhancing processing effectiveness and efficiency.

In our model, the Dual Attention Module mainly plays the role of auxiliary training, and we place the Dual Attention Module before the auxiliary classifier. We have designed further empirical experiments to address this point, and the results of the experiments as well as the analysis of the results are shown in Subsect. 4.5.2.

4 Experiment

4.1 Datasets and evaluation metric

4.1.1 Datasets

We evaluate the proposed method on two adaptation tasks: i.e., GTA5→Cityscapes and GTA5→ACDC. The source dataset GTA5 is the synthetic dataset. GTA5 contains 24,966 high-resolution images, which are collected from a video game. The target dataset Cityscapes collects street-view scenes from 50 different cities, which contains 2975 images for training and 500 images for validation. Specifically, we use the ACDC dataset as the target domain. ACDC consists of a large set of 4006 images which are equally distributed between four common adverse conditions: fog, night, rain, and snow. These images have high-quality pixel-level semantic annotation and are compatible with Cityscapes annotation style.

4.1.2 Evaluation metric

We report pre-class IoU and mean IoU over all classes. For GTA5→Cityscapes, we focus on 19 classes of pre-class IoU and mean IoU. For GTA5→ACDC, in addition to the pre-class IoU for all classes and mean IoU, we also generalize the mean IoU of the image under four adverse conditions.

4.2 Training details

We train the model in two stages. In the first stage, we perform unsupervised domain adaptive training for GTA5→Cityscapes and GTA5→ACDC. Then, we use the trained model to generate pseudo labels for unlabeled

Cityscapes and ACDC datasets. In the second stage, we use pseudo labels to fine-tune the model from the first stage for supervised learning.

Stage 1. We use GTA5 as the source-domain dataset and Cityscapes and ACDC as the target-domain dataset. Firstly, we train the model with a batch size of 4. We resize the input source-domain images to 1280×720 and the target-domain images to 1024×512 , and randomly crop the input images to 780×512 . Secondly, we adopt the SGD (Stochastic Gradient Descent) optimizer with a batch size of 2 for the segmentation model and set its initial learning rate to 0.0002. The optimizer for the two discriminators D_p and D_a is Adam [70], and the initial learning rate is set to 0.0001. We also find that the model performs best when the dropout value is 0.1, and both the segmentation model G and the discriminators D_p and D_a use a multi-learning rate adjustment strategy.

Stage 2. The best model trained in Stage 1 is firstly used as a pre-trained model to generate pseudo labels for Cityscapes and ACDC datasets. The unlabeled target domain images, the ground truth labels, and the pseudo labels are shown in Fig. 4. The generated pseudo labels can be more accurately classified into cars, pedestrians, sidewalks, traffic signs, and buildings, although the edges of different objects are relatively rough. However, it is still more difficult to learn the small category data with a small number of samples, in which there is a large amount of uncertainty.

Then, the generated pseudo labels are used as annotations for the Cityscapes and ACDC dataset, and uncertainty estimation is added to automatically learn a pseudo-labeling threshold for further fine-tuning of the model. In the process of fine-tuning the model using pseudo labels and uncertainty estimation, the training parameters are similar to the Stage 1, with the difference that the model is trained with a batch size of 11 and a crop size of 512×256 for the input image, which effectively prevents the model from overfitting.

4.3 Comparisons with other methods

Synthetic-to-real. We present the adaptation results on task GTA5→Cityscapes in Table 1 with comparisons to several domain adaptation methods: Adaptsegnet [37], Advent [15], BDL [16], CLAN [13], MRNet [71], MRNet+Pseudo [72]. For these methods, the baseline model is trained using ResNet101-based architectures. The competitive methods could be roughly categorized according to the usage of pseudo labels. For example, AdaptSegNet, CLAN, and ADVENT focus on aligning the distribution between the source domain and the target domain without the utilization of the pseudo labels. BDL and MRNet+Pseudo are based on pseudo labels learning to fully exploit the unlabeled target-domain data. The

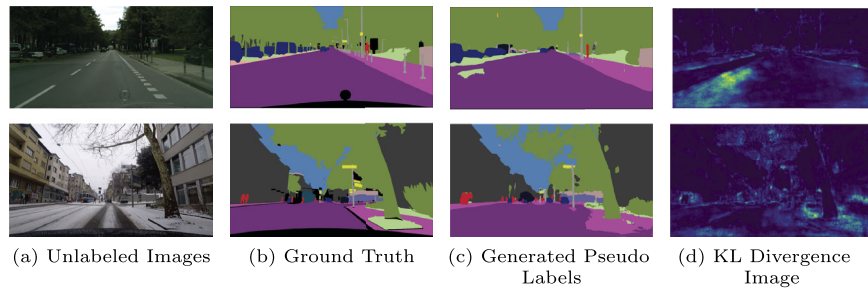


Figure 4 Generated pseudo labels on Cityscapes and ACDC datasets and their KL divergence

proposed method achieves the 51.7% mIoU result, which surpasses other methods. Furthermore, our method performs better IoU on most pre-class, especially on small objects such as rider, bike, sign, pole, and light.

First of all, we analyze the performance of the proposed method based on the widely used GTA5→Cityscapes benchmark. Table 1 shows that: 1) We have achieved better performance in domain adaptation from synthetic to real datasets, and the proposed method realizes the recognition performance of 51.7% mIoU, which exceeds the mIoU of other methods. In addition, our method performs best in most scenarios including road, sidewalk, building, wall, fence, pole, sign, veg, sky, person, rider, car, bus, and bike, and also performs competitively in other pre-classes. 2) Compared to the model with pseudo labels, such as MRNet+Pseudo (48.3% mIoU) and BDL(48.5% mIoU), our proposed method achieves a performance of 51.7% with a significant improvement of +3.4% and +3.2% in terms of mIoU, respectively. 3) Compared to the model without pseudo labels, we used the attention module method (46.72% mIoU), which performs better than most traditional model, such as the MRNet (45.5% mIoU), CLAN (43.2% mIoU), ADVENT (45.5% mIoU) and AdaptSegNet (42.4% mIoU), it also verifies the effectiveness of the attention module.

Normal-to-adverse. We present a new benchmark for unsupervised domain adaptation of semantic segmentation: GTA5→ACDC. It is easy to see the phenomenon that all mIoU metrics have decreased to some extent compared to normal weather. Besides, we change the target domain and test data, to analyze the impact of different adverse weather on the results of segmentation. The results are shown in Table 2 and Table 3. Table 1 shows that: 1) We achieve good performance in domain adaptation from normal dataset to adverse dataset, the proposed method with Attention module and the uncertainty method arrives the results 37.44% mIoU. Although GTA5→ACDC is a new task and there are no other benchmarks to compare, it somewhat proves the validity of our method by comparing the four methods in Tables 2 and 3, and

further proof will be provided by subsequent ablation experiments. 2) The study experimented with and analyzed four different weather conditions: fog, rain, snow, and night. The findings revealed that these conditions have an impact on the model performance in a sequential manner. It was observed that the fog condition had the least effect on the performance, whereas the effect increased with rain, snow, and night. Rain and snow for the lens blocking, and the coverage of the object will lead to a decline in the performance of the model. The night conditions may lead to very bad results due to the lack of light, and the darkness in an unlit area can make it difficult for the model to distinguish objects, requiring a unique approach different from other weather conditions to solve the problem.

Visualization. We provide qualitative results of semantic segmentation for two transfer learning tasks. The results of the GTA5→Cityscapes task are shown in Fig. 5, and the results of the GTA5→ACDC task are shown in Fig. 6.

4.4 Effect of uncertainty estimation

We performed the ablation experiment to further investigate the effect of the uncertainty method in the proposed model. As shown in Table 4, we can find that identifying and exploiting these uncertainties helps us to achieve better results, regardless of whether the target domain is a normal dataset or an adverse dataset. When the target domain is the Cityscapes dataset, the mIoU of the model with the uncertainty method improves by +2.35% compared to the model without the uncertainty method. Similarly, when the target domain is the ACDC dataset, the mIoU of the model with the uncertainty method improves by +1.29% compared to the model without the uncertainty method. It proves the validity of the uncertainty module. Based on the visualization results in Figs. 5 and 6, it is observed that incorporating uncertainty estimation into the optimization objective concentrates the uncertainty distribution in the segmentation results from the edges of the object to the more challenging regions to segment. Specifically, when the target dataset is Cityscapes, the model used

Table 1 Quantitative results on GTA5→Cityscapes. We present pre-class IoU and mIoU. The best accuracy in every column is in bold

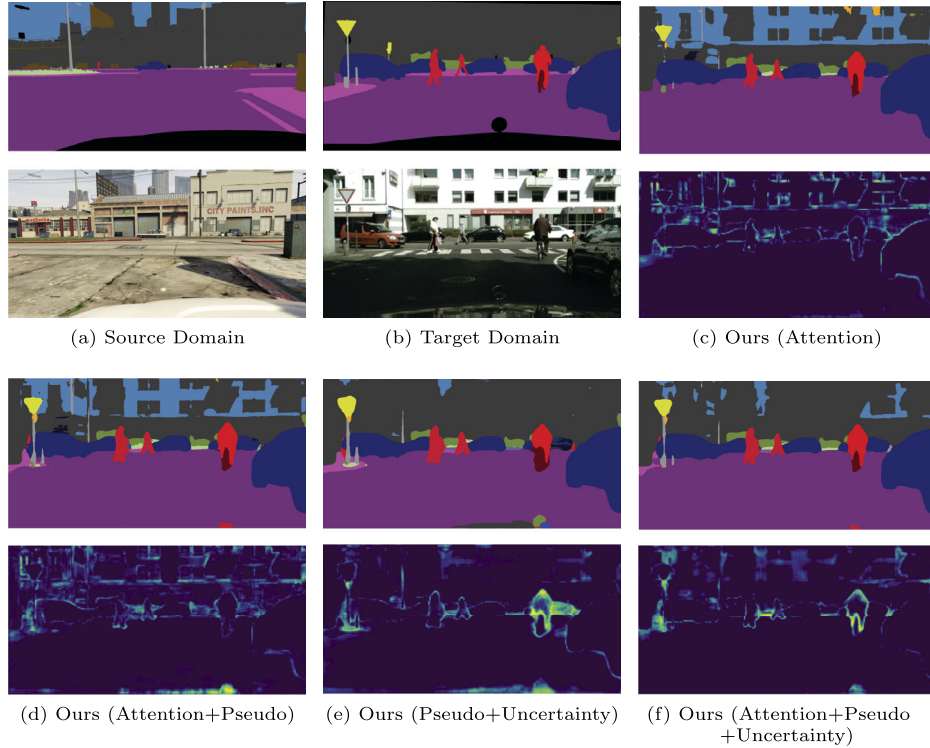
Method	Road	Sidewalk	Building	Wall	Fence	Pole	Light	Sign	Veg	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Mbike	Bike	mIoU
GTA5→Cityscapes																				
Source model	75.8	16.8	77.2	12.5	21	25.5	30.1	20.1	81.3	24.6	70.3	53.8	26.4	49.9	17.2	25.9	6.5	25.3	36	36.6
AdaptSegNet [37]	86.5	36	79.9	23.4	23.3	23.9	35.2	14.8	83.4	33.3	75.6	58.5	27.6	73.7	32.5	35.4	3.9	30.1	28.1	42.4
ADVENT [15]	89.4	33.1	81	26.6	26.8	27.2	33.5	24.7	83.9	36.7	78.8	58.7	30.5	84.8	38.5	44.5	1.7	31.6	32.4	45.5
BDL [16]	91	44.7	84.2	34.6	27.6	30.2	36	36	85	43.6	83	58.6	31.6	83.3	35.3	49.7	3.3	28.8	35.6	48.5
CLAN [13]	87	27.1	79.6	27.3	23.3	28.3	35.5	24.2	83.6	27.4	74.2	58.6	28	76.2	33.1	36.7	6.7	31.9	31.4	43.2
MRNet [71]	89.1	23.9	82.2	19.5	20.1	33.5	42.2	39.1	85.3	33.7	76.4	60.2	33.7	86	36.1	43.3	5.9	22.8	30.8	45.5
Ours (Attention)	91.94	43.32	82.42	25.72	26.03	35.63	41.24	23.58	75.94	32.85	79.64	64.21	36.62	86.67	37.29	36.42	0.01	28.89	30.21	46.72
MRNet+Pseudo	90.5	35	84.6	34.3	24	36.8	44.1	42.7	84.5	33.6	82.5	63.1	34.4	85.8	32.9	38.2	2	27.1	41.8	48.3
[72]																				
Ours (Attention+Pseudo)	93.21	49.34	84.13	28.7	28.65	39.86	44.03	43.18	85.7	38.58	82.04	62.8	31.73	87.01	26.95	30.95	0	27.3	48.92	49.11
Ours (Pseudo+Uncertainty)	91	48.32	85.3	29.68	22.65	39.42	46.84	42.25	86.27	37.67	87.69	60.65	31.69	84.39	33.1	35.64	0	22.25	40.69	48.71
Ours (Attention+Pseudo+Uncertainty)	93.46	50.26	84.63	28.31	30.09	40.02	48.01	36.31	85.88	41.38	86.46	64.14	36.77	88.21	34.11	39.44	0	31.56	50.04	51.07

Table 2 Quantitative results on GTA5→ACDC. We present pre-class IoU and mIoU. The best accuracy in every column is in bold

Method	Road	Sidewalk	Building	Wall	Fence	Pole	Light	Sign	Veg	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Mbike	Bike	mIoU
GTA5→ACDC																				
Ours (Attention)	71.34	7.49	51.64	24.56	14.79	25.04	52.03	19.25	66.93	28.78	81.1	42.76	14.16	75.1	32.08	30.3	0.03	25.18	5.6	35.17
Ours (Attention+Pseudo)	74.62	27.03	51.82	25.74	16.54	31.73	53.09	34.36	63.35	26.21	80.58	42.08	7.18	72.77	24.94	28.07	0.25	19.06	7.41	36.15
Ours (Pseudo+Uncertainty)	69.07	8.97	54.73	27.94	19.06	34.2	57.99	30.05	65.24	25.06	80.78	46.53	12.39	75.01	34.2	27.73	0	12.03	11.97	36.5
Ours (Attention+Pseudo+Uncertainty)	72.83	16.29	53.03	27.44	17.94	34.1	55.06	34.02	66.18	29.01	81.08	47.13	12.19	75.97	27.89	33.8	0.44	14.01	12.89	37.44

Table 3 Quantitative results on GTA5→ACDC. We present the mIoU of adverse conditions. The best accuracy in every column is in bold

Method	ACDC	ACDC_fog	ACDC_snow	ACDC_rain	ACDC_night
Ours (Attention)	35.17	46.15	40.03	42.29	9.17
Ours (Attention+Pseudo)	36.15	46.58	41.47	41.05	14.53
Ours (Pseudo+Uncertainty)	36.5	48.11	41.21	41.99	13.72
Ours (Attention+Pseudo+Uncertainty)	37.44	47.53	42.74	42.17	14.57

**Figure 5** The visualization of some GTA5→Cityscapes results

in Fig. 5 (f) can more accurately distinguish the boundary between the person and the bicycle, as well as the background of the image, i.e., the background building and the sky, when compared to the model used in Fig. 5 (d), which does not incorporate uncertainty estimation and mistakes the buildings in the background for the sky. Similarly, when the target dataset is ACDC, as shown in Fig. 6, compared to the model used in Fig. 6 (d) without uncertainty estimation, the model used in Fig. 6 (f) with uncertainty estimation effectively reduces misidentification and overidentification, and the overall segmentation accuracy is slightly improved, it proves the effectiveness of the uncertainty method in the presence of disturbances from adverse conditions. According to the experimental results in Table 3, we can see that the model with the uncertainty estimation module achieves considerable performance in weather conditions such as snow and night conditions. The potential reason is that the low-quality image under the abnor-

mal weather may lead to large uncertainty. It highlights the importance of the uncertainty estimation module.

Uncertainty estimation is an important tool for domain adaptive models using pseudo labels. It not only improves the quality of pseudo-labels, but also helps the model to more effectively perform sample selection, balance exploration and exploitation, automatically adjust the weights of the loss function, and evaluate the generalization ability of the model. The correct use of uncertainty information can significantly improve the prediction accuracy and robustness of the model on the target domain. However, the uncertainty method here is to ignore the features with greater uncertainty, which obviously will have a negative impact on the network. Experimental results show that accounting for uncertainty does improve results, especially in the normal domain, although the uncertainty is higher under adverse conditions, the results under adverse conditions are not more significant than those under normal conditions. The potential reason is that the uncertainty is

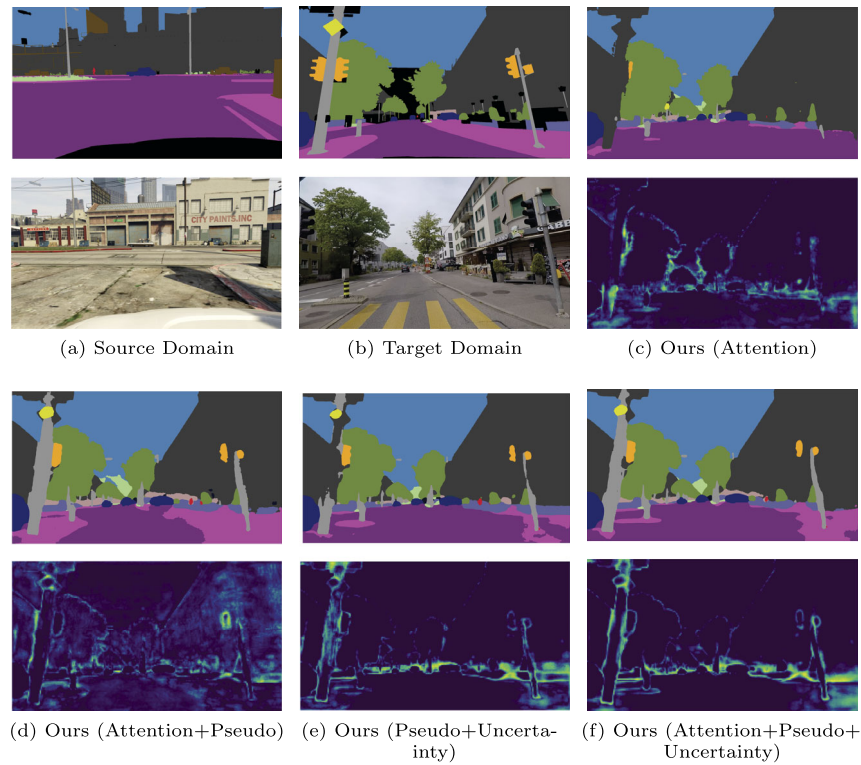


Figure 6 The visualization of some GTA5→ACDC results

Table 4 Ablation study on Uncertainty module

Method	Target domain	mIoU
Ours (without Uncertainty module)	Cityscapes	49.11
Ours (with Uncertainty module)	Cityscapes	51.06
Ours (without Uncertainty module)	ACDC	36.15
Ours (with Uncertainty module)	ACDC	37.44

Table 5 Ablation study on Dual Attention Module

Method	target domain	mIoU
Ours (without Attention module)	Cityscapes	48.71
Ours (with Attention module)	Cityscapes	51.06
Ours (without Attention module)	ACDC	36.5
Ours (with Attention module)	ACDC	37.44

too high under adverse weather, and the model learns too little valid information, which leads to the underfitting of the data in the target domain.

4.5 Effect of dual attention module

4.5.1 Ablation experiment

As shown in Table 1 and Table 2, some small objects such as poles and sidewalks get better results because selective fusion over local features enhances the discrimination of details. Meanwhile, the selective integration among channel maps can capture context information. The semantic consistency has been improved obviously. Therefore, we believe that adding an attention module can capture uncertainty more accurately and improve the adaptation result. As shown in Table 5, when the target domain is the Cityscapes dataset, the model with the Attention module improves mIoU by +2.35% compared to the model without the Attention module. Similarly, when the target do-

main is the ACDC dataset, the model with the Attention module improves mIoU by +0.94% compared to the model without the Attention module. In Fig. 5, compared to the model used in Fig. 5 (e) without Dual Attention Module, the model used in Fig. 5 (f) better recognizes the boundary between the person and the boundary of the bike and avoids mixing the boundary of the person and the bicycle with the boundary of the car located behind them, and we find that the uncertainty is clearly focused on the boundary of the small object in Fig. 5 (f), as opposed to the high uncertainty shown in Fig. 5 (e), which leads to a much finer level of semantic segmentation, it proves that the Dual Attention Module enhances the predictions of the auxiliary classifier and reduces the uncertainty on critical information. Similarly, In Fig. 5, compared to the model used in Fig. 6 (e) without the Dual Attention Module, the model used in Fig. 6 (f) with the Dual Attention Module recognizes the boundary of the pole more clearly, and the seg-

mentation accuracy of the other things also has a slight improvement, it proves the effectiveness of the Dual Attention Module in the presence of disturbances from adverse conditions.

Furthermore, as shown in Table 1, the method which uses the attention module and pseudo method (49.11% mIoU) improves +0.81% mIoU compared to the MRNet+Pseudo method, also our proposed model without the uncertainty method and pseudo labels (46.72%) performs improves mIoU by +1.22% than MRNet, it verifies the effectiveness of the proposed Attention module in reducing the noise from the pseudo labels and demonstrates that the Dual Attention module can effectively improve the performance of the model, which is related to its ability to enhance the representation of features by capturing the global correlation of the features between the temporal and spatial dimensions. At the same time, the Dual Attention Module can capture the image context information by multi-scale feature fusion to enhance the ability of the model to extract multi-scale features, to reduce the optimal joint generalization error and inter-domain variation. The addition of a Dual Attention Module before the auxiliary classifier plays a role in reducing the pseudo-label noise and assisting the model training, which achieves the purpose of improving the model accuracy.

4.5.2 Effect of dual attention module's position

We have conducted further experiments on the effect of the position of the dual attention module on the model's performance. The results of the experiments are shown in Table 6. It is shown that the dual attention module performs better at the low-level features branch than at the high-level features branch in both Stage 1 and Stage 2. There are several potential reasons: 1) Improving computational efficiency: Applying the Dual Attention Module to low-level features is feasible to inspire the network to learn features more efficiently, thus saving computational resources and improving the computational efficiency of the model. 2) Reducing pseudo-label uncertainty: Adding a Dual Attention Module before the auxiliary classifier can be exploited to narrow the gap in feature representation capabilities between the primary and auxiliary classifiers, which reduces pseudo-labeling uncertainty and also reduces the risk of critical information being ignored due to high uncertainty between the predictions of the primary and auxiliary classifiers. 3) Preventing overfitting: This mechanism works to prevent over-reliance on high-level features. The integration of the Dual Attention Module prior to the primary classifier may induce the model to excessively depend on the elevated high-level features enhanced by the attention mechanism, potentially overlooking the nuanced and intricate low-level feature details.

Table 6 Effect of Dual Attention Module's Position

Method	Stage	mIoU
low-level features branch	Stage 1	46.72
low-level features branch	Stage 2	51.07
high-level features branch	Stage 1	45.19
high-level features branch	Stage 2	50.19

Table 7 Comparisons with DANet and SENet

Method	Stage	mIoU	Params (M)
SENet	Stage 1	43.43	103.36
SENet	Stage 2	48.56	103.36
DANet	Stage 1	45.29	136.25
DANet	Stage 2	49.09	136.25
Ours	Stage 1	46.72	136.16
Ours	Stage 2	51.06	136.16

4.5.3 Comparisons with DANet and SENet

We have conducted the experiment to demonstrate the performance boosting of our proposed attention module, and the results of the experiments are shown in Table 7.

The experimental result shows that our method achieves optimal performance. The overall number of parameters of the model is less compared to DANet, since the structure of Segmentation Network G is the same in Stage 1 and Stage 2, the number of parameters is also the same. Based on experimental results, it can be deduced that our model achieves a balance between accuracy and number of parameters.

5 Conclusion

This study presents a novel framework for uncertainty estimation in domain adaptive semantic segmentation. The framework incorporates a Dual Attention Module and an uncertainty indicator based on KL divergence. We evaluate the proposed methods on two transfer learning tasks: GTA5→Cityscapes and GTA5→ACDC. Our results demonstrate that the generalization ability of the cross-domain semantic segmentation model is improved in both normal and adverse weather conditions. Additionally, we assess the model's robustness and generalization ability under anomalous conditions on the task of GTA5→ACDC. The experimental results show that our method achieves considerable performance in handling diverse datasets and effectively mitigating uncertainty issues in domain adaptation. Furthermore, this research confirms the effectiveness of the attention mechanism in reducing model uncertainty. These findings are crucial for ensuring the safety and reliability of autonomous driving systems, particularly in complex environmental conditions.

This work mainly focuses on the paradigm that the specific model corresponds to the specific tasks. Currently, one unified model with multiple tasks is becoming dominant in autonomous driving research. There are several

salient research perspectives. First, the meaningful incorporation of uncertainty and the unified foundation model would be a valuable scientific problem in autonomous driving. Second, the development of generative models like diffusion models [73] and 3D Gaussian splatting models [74] would further inspire cross-domain adaptation research. Last but not least, the connections between uncertainty and robustness deserve further mining. Our proposed method improves feature representation through explicit inter-channel interactions, while the up-to-date visual backbones, such as Swin Transformer [75], can improve the computation efficiency and the feature representation through the local sense fields and multi-scale modeling. We leave the scaling of our proposed uncertainty-aware domain adaptation method on the new backbones like Swin Transformer [75] in future work.

Acknowledgements

The authors are grateful for the efforts of our colleagues at the Sino-German Center of Intelligent Systems, Tongji University.

Author contributions

HY led the project and provided financial support. The main idea of this paper was proposed by JY. PW and BL contributed to the design and implementation of the research, JY and PW prepared the manuscript initially and performed all the steps of the proofs in this research. All authors read and approved the final manuscript.

Funding

This study is supported by the Chinese Natural Science Foundation Key Program under Grant No. 62133011.

Availability of data and materials

Not applicable.

Code availability

<https://github.com/liubo629/Seg-Uncertainty-dual-attention>.

Declarations

Competing interests

There are no conflicts of interest for this paper.

Received: 18 March 2024 Revised: 8 May 2024 Accepted: 31 May 2024

Published online: 08 July 2024

References

1. R. Szeliski, *Computer Vision - Algorithms and Applications*, 2nd edn. Texts in Computer Science. (Springer, Berlin, 2022)
2. J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(4), 640–651 (2015)
3. H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in *CVPR* (2017), pp. 6230–6239
4. V. Badrinarayanan, A. Kendall, R.C. Segnet, A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(12), 2481–2495 (2017)
5. L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected crfs, in *ICLR* (2015)
6. L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L.Y. Deeplab, Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**(4), 834–848 (2018)
7. L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation (2017). CoRR, [arXiv:1706.05587](https://arxiv.org/abs/1706.05587)
8. L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in *ECCV* (Springer, Berlin, 2018), pp. 833–851
9. Y. Yuan, X. Chen, J. Wang, Object-contextual representations for semantic segmentation, in *ECCV* (2020), pp. 173–190
10. Y. Yuan, L. Huang, J. Guo, C. Zhang, X. Chen, J.W. Ocnct, Object context for semantic segmentation. *Int. J. Comput. Vis.* **129**(8), 2375–2398 (2021)
11. M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in *CVPR* (2016), pp. 3213–3223
12. G. Neuhold, T. Ollmann, S.R. Bulò, P. Kotschieder, The mapillary vistas dataset for semantic understanding of street scenes, in *ICCV* (2017), pp. 5000–5009
13. Y. Luo, L. Zheng, T. Guan, J. Yu, Y. Yang, Taking a closer look at domain shift: category-level adversaries for semantics consistent domain adaptation, in *CVPR* (2019), pp. 2507–2516
14. V.M. Patel, R. Gopalan, R. Li, R. Chellappa, Visual domain adaptation: a survey of recent advances. *IEEE Signal Process. Mag.* **32**(3), 53–69 (2015)
15. T.-H. Vu, H. Jain, M. Bucher, M. Cord, P. Pérez, ADVENT: adversarial entropy minimization for domain adaptation in semantic segmentation, in *CVPR* (2019), pp. 2517–2526
16. Y. Li, L. Yuan, N. Vasconcelos, Bidirectional learning for domain adaptation of semantic segmentation, in *CVPR* (2019), pp. 6936–6945
17. Z. Zheng, Y. Yang, Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *Int. J. Comput. Vis.* **129**(4), 1106–1120 (2021)
18. M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P.W. Fieguth, X. Cao, A. Khosravi, U.R. Acharya, V. Makarek, S. Nahavandi, A review of uncertainty quantification in deep learning: techniques, applications and challenges. *Inf. Fusion* **76**, 243–297 (2021)
19. D. Feng, A. Harakeh, S.L. Waslander, K. Dietmayer, A review and comparative study on probabilistic object detection in autonomous driving. *IEEE Trans. Intell. Transp. Syst.* **23**, 9961–9980 (2020)
20. A. Kendall, Y. Gal, What uncertainties do we need in Bayesian deep learning for computer vision?, in *NeurIPS* (2017), pp. 5574–5584
21. F.K. Gustafsson, M. Danelljan, T.B. Schön, Evaluating scalable Bayesian deep learning methods for robust computer vision, in *CVPR Workshops* (2020), pp. 1289–1298
22. A. Kendall, V. Badrinarayanan, R. Cipolla, Bayesian segnet: model uncertainty in deep convolutional encoder-decoder architectures for scene understanding, in *BMVC* (2017)
23. J. Postels, F. Ferroni, H. Coskun, N. Navab, F. Tombari, Sampling-free epistemic uncertainty estimation using approximated variance propagation, in *ICCV* (2019), pp. 2931–2940
24. ISO Central Secretary Organization, ISO/PAS 21448: road vehicles — safety of the intended functionality. Standard, International Organization for Standardization, Geneva, Switzerland, 2019
25. J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, in *CVPR* (2019), pp. 3146–3154
26. T.M. Cover, J.A. Thomas, *Elements of Information Theory* (Wiley, New York, 2001)
27. Y. Gal, Z. Ghahramani, Dropout as a Bayesian approximation: representing model uncertainty in deep learning, in *ICML* (2016), pp. 1050–1059
28. S.R. Richter, V. Vineet, S. Roth, V. Koltun, Playing for data: ground truth from computer games, in *ECCV* (2016), pp. 102–118
29. C. Sakaridis, D. Dai, L.V. Gool, ACDC: the adverse conditions dataset with correspondences for semantic driving scene understanding, in *ICCV* (2021), pp. 10745–10755
30. E. Tzeng, J. Hoffman, K. Saenko, T. Darrell, Adversarial discriminative domain adaptation, in *CVPR* (2017), pp. 2962–2971
31. Z. Pei, Z. Cao, M. Long, J. Wang, Multi-adversarial domain adaptation. *Proc. AAAI Conf. Artif. Intell.* **32**, 3934–3941 (2018)
32. M. Chen, Z.E. Xu, K.Q. Weinberger, F. Sha, Marginalized denoising autoencoders for domain adaptation, in *ICML* (2012)
33. K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, D. Erhan, Domain separation networks, in *NeurIPS* (2016), pp. 343–351
34. R. Sun, X. Zhu, C. Wu, C. Huang, J. Shi, L. Ma, Not all areas are equal: transfer learning for semantic segmentation via hierarchical region selection, in *CVPR* (2019), pp. 4360–4369
35. M. Ghifary, W.B. Kleijn, M. Zhang, Domain adaptive neural networks for object recognition, in *PRICAI* (2014), pp. 898–904

36. M. Long, Y. Cao, J. Wang, M.I. Jordan, Learning transferable features with deep adaptation networks, in *ICML*, ed. by F.R. Bach, D.M. Blei (2015), pp. 97–105
37. Y.-H. Tsai, W.-C. Hung, S. Schuster, K. Sohn, M.-H. Yang, M. Chandraker, Learning to adapt structured output space for semantic segmentation, in *CVPR* (2018), pp. 7472–7481
38. S. Lee, D. Kim, N. Kim, S.-G. Jeong, Drop to adapt: learning discriminative features for unsupervised domain adaptation, in *ICCV* (2019), pp. 91–100
39. Y. Chen, W. Li, X. Chen, L. Van Gool, Learning semantic segmentation from synthetic data: a geometrically guided input-output adaptation approach, in *CVPR* (2019), pp. 1841–1850
40. W.-L. Chang, H.-P. Wang, W.-H. Peng, W.-C. Chiu, All about structure: adapting structural information across domains for boosting semantic segmentation, in *CVPR* (2019), pp. 1900–1909
41. J. Dong, Y. Cong, G. Sun, Y. Liu, X. Xu, CSCL: critical semantic-consistent learning for unsupervised domain adaptation, in *ECCV* (2020), pp. 745–762
42. C. Michaelis, B. Mitzkus, R. Geirhos, E. Rusak, O. Bringmann, A.S. Ecker, M. Bethge, W. Brendel, Benchmarking robustness in object detection: autonomous driving when winter is coming (2019). CoRR, [arXiv:1907.07484](https://arxiv.org/abs/1907.07484)
43. H. Yin, R. Wang, B. Liu, J. Yan, On adversarial robustness of semantic segmentation models for automated driving, in *IV Symposium* (2022)
44. A. Lengyel, S. Garg, M. Milford, J.C. van Gemert, Zero-shot day-night domain adaptation with a physics prior, in *ICCV* (2021), pp. 4379–4389
45. D. Kothandaraman, R. Chandra, D. Manocha, SS-SFDA: self-supervised source-free domain adaptation for road segmentation in hazardous environments, in *ICCV Workshop* (2021), pp. 3042–3052
46. C. Sakaridis, D. Dai, L. Van Gool, Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(6), 3139–3153 (2022)
47. C. Sakaridis, D. Dai, S. Hecker, L. Van Gool, Model adaptation with synthetic and real data for semantic dense foggy scene understanding, in *ECCV* (2018), pp. 707–724
48. M. Jaderberg, K. Simonyan, A. Zisserman, K. Kavukcuoglu, Spatial transformer networks (2015). CoRR, [arXiv:1506.02025](https://arxiv.org/abs/1506.02025)
49. J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in *CVPR* (2018), pp. 7132–7141
50. A. Borji, L. Itti, State-of-the-art in visual attention modeling. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(1), 185–207 (2013)
51. S. Qiu, Y. Wu, S. Anwar, C. Li, Investigating attention mechanism in 3d point cloud object detection, in *International Conference on 3D Vision* (2021), pp. 403–412
52. H. Zhang, I.J. Goodfellow, D.N. Metaxas, A. Odena, Self-attention generative adversarial networks, in *ICML*, ed. by K. Chaudhuri, R. Salakhutdinov (2019), pp. 7354–7363
53. X. Wang, R.B. Girshick, A. Gupta, K. He, Non-local neural networks, in *CVPR* (2018), pp. 7794–7803
54. T. Kim, H. Lee, D.K. Ucanet, Uncertainty augmented context attention for polyp segmentation, in *ACM MM* (2021), pp. 2167–2175
55. R. McAllister, Y. Gal, A. Kendall, M. van der Wilk, A. Shah, R. Cipolla, A. Weller, Concrete problems for autonomous vehicle safety: advantages of Bayesian deep learning, in *IJCAI* (2017), pp. 4745–4753
56. J. Gawlikowski, C.R. Njiteutcheu Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A.M. Kruspe, R. Triebel, P. Jung, R. Roscher, M. Shahzad, W. Yang, R. Bamler, X.X. Zhu, A survey of uncertainty in deep neural networks (2021). CoRR, [arXiv:2107.03342](https://arxiv.org/abs/2107.03342)
57. T. Bayes, An essay towards solving a problem in the doctrine of chances. *Philos. Trans. R. Soc. Lond.* **53**, 370–418 (2021)
58. P.M. Lee, *Bayesian Statistics: An Introduction* (Wiley, New York, 2012)
59. D. Barber, *Bayesian Reasoning and Machine Learning* (Cambridge University Press, Cambridge, 2012)
60. P. McClure, N. Kriegeskorte, Representation of uncertainty in deep neural networks through sampling (2016). CoRR, [arXiv:1611.01639](https://arxiv.org/abs/1611.01639)
61. S. Wirges, M. Reith-Braun, M. Lauer, C. Stiller, Capturing object detection uncertainty in multi-layer grid maps, in *IEEE Intelligent Vehicles Symposium (IV)* (2019), pp. 1520–1526
62. A. Loquercio, M. Segù, D. Scaramuzza, A general framework for uncertainty estimation in deep learning. *IEEE Robot. Autom. Lett.* **5**(2), 3153–3160 (2020)
63. B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, in *NeurIPS* (2017), pp. 6402–6413
64. J. Gast, S. Roth, Lightweight probabilistic deep networks, in *CVPR* (2018), pp. 3369–3378
65. S. Choi, K. Lee, S. Lim, S. Oh, Uncertainty-aware learning from demonstration using mixture density networks with sampling-free variance modeling, in *ICRA* (2018), pp. 6915–6922
66. G.P. Meyer, A. Laddha, E. Kee, C. Vallespi-Gonzalez, C.K. Wellington, Lasernet: an efficient probabilistic 3d object detector for autonomous driving, in *CVPR* (2019), pp. 12677–12686
67. J. Postels, F. Ferroni, H. Coskun, N. Navab, F. Tombari, Sampling-free epistemic uncertainty estimation using approximated variance propagation, in *ICCV* (2019), pp. 2931–2940
68. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *CVPR* (2016), pp. 770–778
69. Z. Zheng, Y. Yang, Unsupervised scene adaptation with memory regularization in vivo (2019). CoRR, [arXiv:1912.11164](https://arxiv.org/abs/1912.11164)
70. D.P. Kingma, J. Ba, Adam: a method for stochastic optimization (2017)
71. Z. Zheng, Y. Yang, Unsupervised scene adaptation with memory regularization in vivo (2019). CoRR, [arXiv:1912.11164](https://arxiv.org/abs/1912.11164)
72. Z. Zheng, Y. Yang, Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation (2020). CoRR, [arXiv:2003.03773](https://arxiv.org/abs/2003.03773)
73. F.-A. Croitoru, V. Hondru, R.T. Ionescu, M. Shah, Diffusion models in vision: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* (2023). [arXiv:2209.04747](https://arxiv.org/abs/2209.04747)
74. G. Chen, W. Wang, A survey on 3d gaussian splatting (2024). [arXiv:2401.03890](https://arxiv.org/abs/2401.03890)
75. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: hierarchical vision transformer using shifted windows (2021). [arXiv:2103.14030](https://arxiv.org/abs/2103.14030)

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)