Cross-Modality Domain Adaptation for Freespace Detection: A Simple yet Effective Baseline

Yuanbin Wang
Institute of Artificial Intelligence, Beihang University
Hangzhou Innovation Institute, Beihang University
SenseTime Research

Leyan Zhu
Institute of Artificial Intelligence, Beihang University
Hangzhou Innovation Institute, Beihang University
SenseTime Research

Shaofei Huang
Institute of Information Engineering, Chinese Academy of Sciences
School of Cyber Security, University of Chinese Academy of Sciences

Tianrui Hui
Institute of Information Engineering, Chinese Academy of Sciences
School of Cyber Security, University of Chinese Academy of Sciences

Xiaojie Li
SenseTime Research

Fei Wang
SenseTime Research

Si Liu
Institute of Artificial Intelligence, Beihang University
Hangzhou Innovation Institute, Beihang University

ABSTRACT
As one of the fundamental functions of autonomous driving system, freespace detection aims at classifying each pixel of the image captured by the camera as drivable or non-drivable. Current works of freespace detection heavily rely on large amount of densely labeled training data for accuracy and robustness, which is time-consuming and laborious to collect and annotate. To the best of our knowledge, we are the first work to explore unsupervised domain adaptation for freespace detection to alleviate the data limitation problem with synthetic data. We develop a cross-modality domain adaptation framework which exploits both RGB images and surface normal maps generated from depth images. A Collaborative Cross Guidance (CCG) module is proposed to leverage the context information of one modality to guide the other modality in a cross manner, thus realizing inter-modality intra-domain complement. To better bridge the domain gap between source domain (synthetic data) and target domain (real-world data), we also propose a Selective Feature Alignment (SFA) module which only aligns the features of consistent foreground area between the two domains, thus realizing inter-domain intra-modality adaptation. Extensive experiments are conducted by adapting three different synthetic datasets to one real-world dataset for freespace detection respectively. Our method performs closely to fully supervised freespace detection methods (93.08% vs. 97.50% F1 score) and outperforms other general unsupervised domain adaptation methods for semantic segmentation with large margins, which shows the promising potential of domain adaptation for freespace detection.

CCS CONCEPTS
• Computing methodologies → Computer vision; Transfer learning; Unsupervised learning; Image segmentation.

KEYWORDS
Freespace Detection, Domain Adaptation, Cross-Modality

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM Reference Format:
Yuanbin Wang, Leyan Zhu, Shaofei Huang, Tianrui Hui, Xiaojie Li, Fei Wang, and Si Liu. 2022. Cross-Modality Domain Adaptation for Freespace Detection: A Simple yet Effective Baseline. In Proceedings of the 30th ACM International Conference on Multimedia (MM ’22), October 10–14, 2022, Lisbon, Portugal. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3503161.3547752

1 INTRODUCTION
Autonomous driving [3, 37, 54, 62] has drawn remarkable attention in recent years from both industry and academia for pursuing intelligent transportation. As one of the essential functions of autonomous driving system, freespace detection aims to perform binary classification on each pixel of the image to label it as drivable or non-drivable, so as to realize safety-ensured vehicle navigation. Current freespace detection methods [7, 10, 42, 46, 50, 53] usually depend on massive labeled training data which is costly to collect and annotate, thus restricting their practical applications. Concretely, vehicles equipped with high-definition cameras are required to drive under various scenarios (e.g., weather conditions,
Figure 1: Illustration of our motivation. (a) Inter-modality intra-domain complement. RGB model either mis-classifies road with shadow into non-road area (box ①) or fails to distinguish between road and sidewalks (box ②), while surface normal model falsely categorizes grassland into road area (box ③). These two modalities are complementary to each other. (b) Inter-domain intra-modality adaptation. Foreground areas of RGB images from both domains belong to the same semantic category which is easy to adapt, while background areas are diverse in semantic categories and styles which are difficult to adapt (left column). Surface normal maps from both domains are domain-invariant in style, thus they are not adapted in our method (right column). The top half of surface normal map is visualized in black because of lacking depth information of this region.

seasons, cities and traffics), which is time-consuming and expensive. It is also laborious to annotate captured road images with pixel-level dense masks. In order to alleviate this data limitation, synthetic images can actually serve as a large-scale data source with much-reduced collection and annotation costs. However, there is a huge gap between synthetic and real-world domains [21, 49, 55], thus leading to performance drop when directly applying a model trained on synthetic images to real ones.

To bridge the domain gap, we propose to investigate unsupervised domain adaptation for freespace detection (UDA-FD), where only labels of source domain are available. To the best of our knowledge, the UDA-FD task has not been explored by previous works. Since RGB data is vulnerable to extreme visual conditions like low illumination, depth data, which could be obtained with stereo cameras, becomes a promising supplement [6, 16, 59] due to its good trade-off between reliability and cost. Since freespace can be hypothesized as a plane, on which the points have similar surface normals, we follow [16] to extract surface normal maps from depth images as a more robust additional input. Based on the above analysis, we propose a cross-modality domain adaptation framework for freespace detection which exploits both RGB data and surface normal data from source and target domains. In order to sufficiently employ the multi-modality multi-domain information, our framework considers two essential perspectives for tackling the UDA-FD task, namely inter-modality intra-domain complement and inter-domain intra-modality adaptation.

For the first perspective, it is known that RGB data and surface normal data hold different characteristics. RGB data can well capture color and texture changes in appearance, while surface normal data can well capture height change in 3D space. In Figure 1(a), we present the freespace predictions of model trained with RGB images only and surface normal maps only respectively. It is shown that RGB model falsely recognizes shadow areas as non-road regions (box ① in RGB Prediction of Figure 1(a)). Furthermore, due to the similarity in appearance between roads and sidewalks, it is hard for RGB model to distinguish non-drivable sidewalks from roads as illustrated in box ② in RGB Prediction of Figure 1(a). The surface normal model, by contrast, classifies these areas into correct labels because it is insensitive to illumination changes and is able to capture height changes in 3D space. As shown in Surface Normal Map of Figure 1(a), a clear boundary can be observed between road and sidewalk areas in the surface normal map, so that it is easy for the model to distinguish them. On the other hand, as shown in box ③ in Surface Normal Prediction of Figure 1(a), for the grassland area where the height changes are minor, surface normal model wrongly labels grassland as road. However, the RGB model makes correct prediction given to the obvious appearance contrast between road and grassland. To take advantage of the mutual complement of these two modalities, we devise a Collaborative Cross Guidance (CCG) module which leverages context information from each modality to crossly guide the feature learning of the other modality in both source and target domains. For each modality, channel-wise global context and spatial-wise response map which is supervised by an auxiliary loss are exploited sequentially to guide the feature learning of the other modality. In this way, multi-modality information can be sufficiently utilized to correct the flawed predictions collaboratively.
For the second perspective, foreground regions in images from source and target domains contain the same semantic category (i.e., drivable road) with similar style, while background regions usually contain different semantic categories with diverse styles. For example, in the first column of Figure 1(b), the foreground areas of source and target images are both three marked roads with similar flat ground and gray appearance. However, the background areas show considerable diversity in both semantic categories and appearances. The image from source domain contains building and grassland, while the image from target domain contains railway and sidewalks, which has large discrepancy in appearance. Considering the different characteristics of foreground and background, it is easier and more stable to mainly adapt features of foreground regions between source and target domains. We accordingly devise a Selective Feature Adaptation (SFA) module which exploits the final predicted logits as attention maps to selectively highlight the foreground region features that are then aligned between source and target domains. It is noticed that our SFA module only works with RGB images, since there is no obvious difference in style between surface normal maps from the two domains as shown in the second column of Figure 1(b).

Our contributions can be summarized as follows:

- To the best of our knowledge, we are the first study to explore unsupervised domain adaptation for freespace detection which assists in alleviating the problem of lacking training data and high annotation cost.
- We propose a cross-modality domain adaptation framework utilizing both RGB and surface normal data. A Collaborative Cross Guidance (CCG) module is developed to realize inter-modality intra-domain complement and a Selective Feature Adaptation (SFA) module is developed to realize inter-domain intra-modality adaptation.
- Extensive experiments show that our method outperforms other general UDA methods for semantic segmentation and achieves close performance to fully supervised freespace detection methods by adapting the model trained on the source domain to target domain without labels (93.08% v.s. 97.50% F1 score on KITTI-Road dataset [18]).

2 RELATED WORK

2.1 Freespace Detection

Freespace detection is a critical task in autonomous driving, and is beneficial to many other tasks [7, 10, 50]. Traditional methods typically adopt hand-crafted features (e.g., color, edge, texture) of the road and detect the road region with pixel-wise classification [1, 2, 51]. However, the generalization ability of these methods is still far from satisfactory due to their strong reliance on road geometry or appearances.

With the development of deep learning [30–32] and proposal of Fully Convolutional Network (FCN) [41], methods that regard freespace detection as a semantic segmentation problem have arisen [24, 42, 46, 53]. For example, Lu et al. [42] proposes an encoder-decoder architecture to segment road regions on a bird’s-eye-viewed image. Han et al. [24] proposes a semi-supervised learning freespace detection method together with a weakly supervised learning method to better leverage the feature distribution of labeled and unlabeled data. The above RGB-based methods are vulnerable to visual noises like illumination changes and overexposure, which leads to inadequate robustness of these methods.

To overcome these issues, some methods introduce other modalities of data as supplement of information, such as Lidar [4, 11, 23] and depth [6, 16, 60, 61]. Gu et al. [23] progressively adapts Lidar data space into RGB data space to effectively fuse features of the two modalities. Chen et al. [11] proposes to fuse RGB and depth features with a dynamic fusion module (DFM). Given the assumption that road regions of an image share the same plane, surface normal information estimated from depth images is further incorporated to help with segmenting road regions [16, 60]. By including mixture of multi-modality inputs apart from RGB data, the reliability of freespace detection is greatly improved.

2.2 Domain Adaptation

Since directly evaluating on one domain with a model trained on another domain usually causes performance drops, domain adaptation methods are developed to help address the domain shift between two distinct domains. Domain adaptation is widely used in classification [14, 17, 21, 64], object detection [6, 33, 52, 65, 66], semantic segmentation [20, 27, 57, 71, 78] and many other fields [39, 72, 74, 75]. In particular, unsupervised domain adaptation (UDA) methods have attracted substantial attention since they are free from manually-annotated labels in the target domain. UDA methods can be roughly divided into three categories:

(1) Feature-level domain adaptation methods. Inspired by the success of generative adversarial network (GAN) [22], this kind of methods [29, 38, 56, 76] employs adversarial training scheme to minimize the distance between the feature distribution of source and target domains. A discriminator is trained to estimate the domain of latent features generated from the network to make the main prediction. With the competition, domain-agnostic features will be generated, thereby reducing the domain shift between source and target domains. Feature-level domain adaptation methods is broadly applied in segmentation tasks [9, 28, 43, 44, 55, 57] with the assistance with segmentation loss functions. For example, Adapt-Seg [55] applies multi-level adversarial network to align features from different levels, which leads to improved performance.

(2) Image-level domain adaptation methods. This type of methods performs pixel-level alignment between source and target domains via image-to-image translation strategies such as GAN [12, 27, 35, 40, 67] and style transfer [5, 68, 70]. The representative GAN-based methods include CyCADA [27], which uses CycleGAN [77] to transfer images from source domain into target domain while maintaining semantic consistency with the original images. The synthesized target-styled images share the annotations with original images and can be used as extra training data.

(3) Self-training domain adaptation methods. Self-training is another widely explored type of methods to train the network with the pseudo labels on the target domain to reinforce the network beliefs [29, 45, 47, 73]. Typically, the network is first trained in the source domain, and then used to generate pseudo labels in the target domain, which are subsequently used as supervision to train the target network. After several repeated iterations, the network can also achieve impressive results on the target domain. In [40, 78, 79],
pseudo labels are selected with a confidence threshold calculated based on the whole dataset in an offline manner, which requires storing all predictions for the whole dataset. To alleviate this issue, Fabio et al. [49] propose Weighted Pseudo-Labels (WPL) to estimate a global threshold during optimizing process, reducing the cost of threshold calculation.

3 METHODS

3.1 Overview

In this work, we focus on the task of unsupervised domain adaptation for freespace detection, with labels only available for source domain. Figure 2(a) illustrates the overall architecture of our cross-modality domain adaptation framework. We take a pair of RGB image and surface normal (SN) map from source or target domain as inputs. The surface normal map is calculated with the corresponding depth image and intrinsic matrix of the camera following [16]. RGB image \(X_{RGB} \in \mathbb{R}^{H \times W \times 3}\) and SN map \(X_{SN} \in \mathbb{R}^{H \times W \times 3}\) are processed by an RGB encoder and an SN encoder respectively, where we adopt typical CNN backbones. The RGB encoder and SN encoder have the same architecture but with different parameters. The encoded features of the two modalities are denoted as \(F_{RGB} \in \mathbb{R}^{H' \times W' \times C}\) and \(F_{SN} \in \mathbb{R}^{H' \times W' \times C}\), where \(H'\) and \(W'\) denote the spatial size of feature map and \(C\) denotes the channel number. Then, \(F_{RGB}\) and \(F_{SN}\) are fed into the Collaborative Cross Guidance (CCG) module to complement information between the two modalities. In order to suppress the mis-response of SN feature with RGB information, we propose to leverage the foreground response map of RGB feature to filter out the false positive responses of SN feature. To this end, we extract the foreground logit, which is generated with a segmentation head applied to the enhanced RGB feature \(F_{RGB}'\), as the foreground attention map of RGB modality. The RGB foreground attention map \(A_{RGB} \in \mathbb{R}^{H \times W}\) is obtained as follows:

\[
P = \text{softmax}([P_0, P_1]),
\]

\[
A_{RGB} = \hat{P}_1,
\]

where \(P \in \mathbb{R}^{2 \times H' \times W'}\) is the output of segmentation head, with channel-0 denoting the probability of being background and channel-1 denoting the probability of being foreground. The larger value of \(A_{i,j}^{RGB}\) indicates the higher confidence of position \((i, j)\) being recognized as foreground. To ensure that the foreground areas are precisely predicted on \(A_{RGB}\), auxiliary segmentation loss \(L_{seg}^{RGB}\) is introduced as supervision. Similarly, auxiliary segmentation loss \(L_{seg}^{SN}\) is applied in SN→RGB process.

Finally, the SN feature map \(F_{SN}\) is multiplied with \(A_{RGB}\) in an element-wise manner to enhance response of foreground areas and suppress response of background areas, thus eliminating the false positive areas on SN features:

\[
F_{SN} = p_{SN} \odot A_{RGB},
\]

where \(\odot\) denotes the Hadamard product. Through the above interactions between RGB and SN features, we can leverage useful information of one modality to guide the other and produce better feature representations of both modalities for the final prediction.

3.2 Collaborative Cross Guidance Module

RGB data and SN data have different characteristics. RGB data works well with the obvious texture and color features (e.g., roads v.s. grasslands) but fails to distinguish between areas with similar appearance (e.g., roads v.s. sidewalks), while SN data works well in capturing height changes but fails to capture texture changes. Motivated by the above differences, we introduce a Collaborative Cross Guidance (CCG) module to complement information between the two modalities.

The detail of CCG module is illustrated in Figure 2(c). Feature map \(F_{RGB} \in \mathbb{R}^{H' \times W' \times C}\) from RGB encoder and feature map \(F_{SN} \in \mathbb{R}^{H' \times W' \times C}\) from SN encoder are taken as inputs of CCG module. To clearly elaborate the process of cross guidance, we take the operations of RGB feature guiding SN feature (denoted as RGB→SN) as an example. The operations of SN→RGB are the same.

First of all, the global context vector of SN feature map \(g_{SN} \in \mathbb{R}^{C}\) is extracted to integrate the global information of SN data. Concretely, we first conduct global average pooling on \(F_{SN}\) to obtain the global feature of SN data as \(\sigma_{SN} \in \mathbb{R}^{C}\). A fully-connected layer together with sigmoid activation is applied to \(\sigma_{SN}\) as follow:

\[
g_{SN} = \sigma(W(\sigma_{SN}))\tag{1},
\]

where \(\sigma(\cdot)\) denotes sigmoid function and \(W \in \mathbb{R}^{C \times C}\) is the projection weight. The combination of global average pooling, fully-connected layer and sigmoid activation is abbreviated as GCE in Figure 2(c). The global context vector \(g_{SN}\) is then multiplied to the RGB feature map \(F_{RGB}\) to pre-inform the RGB branch with SN information for generating better representations. The generated feature map is denoted as \(F_{RGB}' \in \mathbb{R}^{H \times W' \times C}\).

In order to suppress the mis-response of SN feature with RGB information, we propose to leverage the foreground response map of RGB feature to filter out the false positive responses of SN feature. To this end, we extract the foreground logit, which is generated with a segmentation head applied to the enhanced RGB feature \(F_{RGB}'\), as the foreground attention map of RGB modality. The RGB foreground attention map \(A_{RGB} \in \mathbb{R}^{H \times W}\) is obtained as follows:

\[
P = \text{softmax}([P_0, P_1]),
\]

\[
A_{RGB} = \hat{P}_1,
\]

where \(P \in \mathbb{R}^{2 \times H' \times W'}\) is the output of segmentation head, with channel-0 denoting the probability of being background and channel-1 denoting the probability of being foreground. The larger value of \(A_{i,j}^{RGB}\) indicates the higher confidence of position \((i, j)\) being recognized as foreground. To ensure that the foreground areas are precisely predicted on \(A_{RGB}\), auxiliary segmentation loss \(L_{seg}^{RGB}\) is introduced as supervision. Similarly, auxiliary segmentation loss \(L_{seg}^{SN}\) is applied in SN→RGB process.

Finally, the SN feature map \(F_{SN}\) is multiplied with \(A_{RGB}\) in an element-wise manner to enhance response of foreground areas and suppress response of background areas, thus eliminating the false positive areas on SN features:

\[
F_{SN} = p_{SN} \odot A_{RGB},
\]

where \(\odot\) denotes the Hadamard product. Through the above interactions between RGB and SN features, we can leverage useful information of one modality to guide the other and produce better feature representations of both modalities for the final prediction.

3.3 Selective Feature Alignment Module

In order to adapt the model trained on the source domain to the target domain, we develop a Selective Feature Alignment (SFA) module for feature-level domain adaptation. We adopt adversarial strategy to align the features from source and target domains, with a discriminator that distinguishes which domain the feature belongs to. Furthermore, as described in Figure 1(b), it is easier to adapt features of only foreground area with the same category than to adapt features of background area with diverse semantics. To this end, before feature maps are fed into the discriminator, we first
Figure 2: Overall architecture of our proposed method. (a) Pipeline for cross-modality domain adaptation framework. An RGB encoder and a surface normal (SN) encoder are adopted to extract features from source and target domains, followed by CCG module to complement between two modalities. SFA module is leveraged to reduce domain gap of RGB features between source and target domains. (b) Illustration of Selective Feature Alignment (SFA) module. (c) Illustration of Collaborative Cross Guidance (CCG) module. (Subscripts $S$ and $T$ denoting source and target domains are omitted for simplification in (a) and (c).)

apply foreground attention maps to selectively suppress responses of background areas, so as to mainly align foreground features from two domains.

As illustrated in Figure 2(a), to obtain precise estimation of the foreground area, the foreground segmentation logit produced by the segmentation head serves as the foreground attention map $A_K \in \mathbb{R}^{H \times W}$, $K \in \{S, T\}$. The value at pixel $(i, j)$ means its probability of being road area. The generation of $A_K$ is similar to Equation 2 and 3.

The detail of SFA module is presented in Figure 2(b). Given feature map $F_{K,n}^{\text{RGB}}$, $K \in \{S, T\}$ extracted from the $n$-th stage of RGB encoder, the foreground attention map $A_K$ is firstly downsampled to meet the spatial size of the two features to produce $A_{K,n}$. Afterwards, the two feature maps are multiplied with $A_{S,n}$ and $A_{T,n}$ respectively to relatively suppress responses at background areas, which conversely enhance the foreground response as follows:

$$F_{S,n}^{\text{RGB}} = F_{S,n}^{\text{RGB}} \odot A_{S,n},$$

$$F_{T,n}^{\text{RGB}} = F_{T,n}^{\text{RGB}} \odot A_{T,n}. \quad (5)$$

The enhanced feature maps $F_{S,n}^{\text{RGB}}$ and $F_{T,n}^{\text{RGB}}$ are then fed to a fully convolutional discriminator $D$ that performs binary classification to determine which domain the feature is from. In our work, we define label 0 for source domain and label 1 for target domain. The corresponding loss function is as follows:

$$L_{\text{adv}} = \sum_{i,j} \log(D(F_{T,n}^{\text{RGB}})_{i,j}) + \log(1 - D(F_{S,n}^{\text{RGB}})_{i,j}), \quad (6)$$

where $i, j$ denote the row index and column index respectively.

In the mean time, the RGB-encoder is trained to fool the discriminator by producing feature maps from target domain which are highly similar to feature maps from source domain. By maximizing the probability of the target prediction being considered as the source prediction, the feature distribution shift between two domains is eliminated in the RGB encoder, thus greatly alleviating the performance degradation on target domain.

The SFA module is only applied to RGB features since surface normal data is domain invariant as shown in Figure 1(b). Experiments in Section 4.4 further demonstrate this hypothesis, so that it is unnecessary to align the feature distribution of surface normal.

3.4 Training Scheme

We use self-training method to train our model in an iterative manner. The training procedure of our method contains 3 rounds. In the first round, since the labels of target domain are not available, we only calculate the segmentation loss for the source domain. Thus, the complete loss function for round 1 is designed as:

$$L = \lambda_3 L_{\text{seg}}^{\text{RGB}} + \lambda_2 L_{\text{seg}}^{\text{SN}} + \lambda_4 L_{\text{seg}}^{\text{SFA}} + \lambda_4 L_{\text{adv}}. \quad (7)$$

where $\lambda_{(\cdot)}$ denotes the weight of each term. For the last 2 rounds of training, we generate pseudo labels for target domain with the network trained in the previous round. Concretely, we set a threshold $\alpha$ to ignore low-confidence pixels with classification probability between $1 - \alpha$ and $\alpha$ when calculating segmentation losses. Thus,
the complete loss function for the last 2 rounds is defined as:
\[
\mathcal{L} = \lambda_1 L_{seg}^{RGB} + \lambda_2 L_{seg}^{SN} + \lambda_3 L_{seg}^{seg} + \lambda_4 L_{adv}
\]
where \( G \) denotes the segmentation network. By alternately training the segmentation network and the discriminator, the segmentation losses of both domains are minimized, while the probability of target feature being regarded as source feature is maximized.

## 4 EXPERIMENTS

### 4.1 Datasets and Evaluation Metrics

In the experiment part, we adopt the setting of adaptation from synthetic datasets to real-world datasets. Three synthetic datasets including R2D [16], SYNTHIA-SF [26] and Virtual KITTI [19] are used as source domains and one real-world dataset KITTI-Road [18] is used as target domain. Detailed description on these datasets can be found in Appendix B. Since multiple submission on the test set of KITTI-Road dataset is not supported, we follow most of the previous works [6, 16] to divide the original training set into 173 for training and 116 for validation. Unless otherwise specified, the experiments following are conducted on these splits.

We adopt the most commonly used metrics including precision (PRE), recall (REC), F1 score and the intersection over union (IoU) as our evaluation metrics.

### 4.2 Implementation Details

Two candidates for RGB/SN encoders are adopted in our work including ResNet-50 [25] and HRNetV2-W48 [63]. Both networks are pretrained on ImageNet dataset [13]. The segmentation heads of the main decoder and auxiliary decoders in CCG module all contain two 1x1 convolutions. The discriminator consists of four strided convolutional layers for feature map downsampling and LeakyReLU [69] is adopted as the activation function with negative slope set to 0.2. For experiments with ResNet-50, only the last stage of feature maps produced by the two encoders are included for following operations. 1 x 1 convolutions are applied first to reduce the channel number from 2048 to 512. For experiments with HRNet, 4 feature maps with hierarchial resolutions are produced by the encoder for each modality. In CCG module, only the largest RGB/SN feature maps among them are used to generate foreground attention maps \( A^{RGB} \) and \( A^{SN} \). Afterwards, all the encoder feature maps of both modalities are multiplied with the corresponding foreground attention maps for cross-modality guidance and then stacked together before feeding to segmentation head for final prediction. As for the SFA module, feature-level alignment is only performed on the largest feature map.

Our method is implemented using PyTorch [48] and experiments are conducted on 4 NVIDIA GTX 1080Ti GPUs with 11GB memory. To train the segmentation network, we apply an SGD optimizer with a learning rate of \( 2.5 \times 10^{-4} \), momentum of 0.9 and weight decay of \( 5 \times 10^{-4} \). Adam [36] with default parameters is utilized as the optimizer of the discriminator. As for the weight of each loss, we set \( \lambda_1 = 0.5, \lambda_2 = 0.5, \lambda_3 = 0.5, \lambda_4 = 0.5 \) for experiments with HRNetV2-W48 and \( \lambda_1 = 10^{-4} \) for experiments with ResNet-50. The threshold \( \alpha \) for generating pseudo labels is set to 0.99.

### 4.3 Comparison with State-of-the-art Methods

To demonstrate the effectiveness of our proposed framework, we carry out experiments to adapt from 3 different synthetic datasets to the real-world dataset KITTI-Road respectively. The results on the validation set are reported in Table 1. Since we are the first to explore the UDA-FD task, there is no previous works in this area to compare with. Thus we pick out several methods of UDA for semantic segmentation which could be adapted to our task and reproduce these methods based on their released codes. The selected methods includes AdaptSeg [55], which is based on feature-level adaptation, ADVENT [57] and IntraDA [47], which seek to minimize the entropy map, and DADA [58], which incorporates depth images as extra supervision. We also conduct more experiments using both RGB and depth information to demonstrate the superiority of our method in Appendix C. As shown in the first 3 rows of Table 1, we train our segmentation network with labeled images on the source domains, which can further improve the performance.
KitTI-Road to provide an upper bound of KitTI-Road validation set.

Take Virtual KitTI→KitTI-Road as an example, our method with ResNet-50 as backbone achieves high performances on the most important metrics including F1 score and IoU compared with the other UDA segmentation methods such as DADA, which also leverages depth images. 1.18% and 2.06% performance gains of F1 score and IoU are obtained by training with pseudo labels (denoted as w/ ST). Replacing the backbone with HRNetV2-W48 further boosts the performance to 94.37% F1 score and 89.34% IoU, which is only 2.18% lower than full-supervision baseline on F1 score, demonstrating the potential of unsupervised domain adaptation framework for freespace detection. It is interesting that other UDA methods like ADVENT [57] and IntraDA [47] typically achieve high recall but much lower precision and IoU, implying that these methods tend to falsely include confusing background areas into foreground areas for lacking complement between RGB and SN information. The visualization results shown in Figure 3 further prove it. Similar performance gap could be observed on R2D→KITTI-Road and SYNTHIA-SF→KITTI-Road except that the overall performances are lower due to the larger domain gap of these two settings, which indicates that the gap magnitude between source and target domains is one of the key factors to affect the adaptation performance.

In addition, we compare with other published fully-supervised and semi-supervised methods on KitTI-Road Benchmark and the results are presented in Table 2. We use Virtual KitTI as source dataset and the original training set of KitTI-Road as target dataset. It is shown that our method outperforms all the UDA methods and one of the semi-supervised methods RGB36-super [4]. Besides, our method achieves close AP and MaxF to the fully supervised methods, further indicating the large potential of our method.

### 4.4 Ablation Study

To evaluate different design of our model, we conduct ablation study on R2D→KITTI-Road with ResNet-50 as backbone.

#### Component analysis.

Table 3 presents the ablation results of input modalities and our proposed modules. The 1-st row denotes the RGB-only results without any adaptation to the target domain, which are far from satisfactory. By inserting the SFA module to reduce domain gap (2-nd row), notable gain with 18.25% on F1 score and 22.10% on IoU is obtained, demonstrating the necessity for feature alignment in domain adaptation. In the 3-rd row, surface normal information is introduced as supplement of RGB information and brings 5.03% F1 score and 7.28% IoU improvements, showing the effectiveness of multi-modal learning. Finally, together with CCG module to complement between RGB and SN data, the performances are further boosted, resulting in the highest scores on both F1 score and IoU.

#### Feature modalities to adapt in SFA.

Table 4 evaluates different modalities for feature alignment in SFA module. The 1-st row shows the results of simply involving two modalities without adaptation. By aligning only RGB features from both domains, notable performance gain (5.53% improvement in IoU) is obtained in the 2-nd row. However, when applying feature alignment to SN features in the 3-rd row, obvious performance drop could be observed comparing with the 1-st row. Aligning both RGB and SN features achieves sub-optimal performances comparing with aligning RGB features only as shown in the 4-th row. The above results indicate that SN data is domain-invariant to some extent. In our work, we only adapt the feature distribution of RGB data.

### 4.5 Qualitative Analysis

As shown in Figure 3, we visualize the predictions of our method and others on the R2D→KITTI-Road setting. For example, in the 4-th column, AdaptSeg [55] and DADA [58] predict sidewalk on the right as road area, while our method is able to differentiate these two categories. Although depth information is also utilized in DADA [58], it is not sufficient to suppress the false predictions on sidewalks. In addition, by training with the selected high-confidence pixels as supervision, our method generates predictions with cleaner boundaries as shown in the 4-th row. Our method shows superiority in distinguishing road area and other road-like areas compared with other UDA methods. Qualitative results on the other two settings are shown in Appendix D.

We also visualize the segmentation results predicted with different modalities in Figure 4. Mutual complement between RGB and SN data can be observed from their predictions. For example, in the 1-st column, SN model can roughly distinguish the sidewalk and...
5 CONCLUSION

In this paper, we explore unsupervised domain adaptation for free-space detection (UDA-FD) between synthetic and real-world data to resolve the data limitation problem and alleviate annotation costs. To the best of our knowledge, we are the first to investigate the UDA-FD task. We propose a cross-modality domain adaptation framework which simultaneously considers both inter-modality intra-domain complement and inter-domain intra-modality adaptation. Extensive experimental results show that our proposed method achieves close performance to fully supervised freespace detection methods, demonstrating the potential of UDA for freespace detection. In the future, we hope to explore more domain adaptation techniques for freespace detection, and generalize our cross-modality domain adaptation framework to other settings, such as freespace detection under extreme weather conditions.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 62122010 and Grant 61876177, in part by the Fundamental Research Funds for the Central Universities, and in part by the Key Research and Development Program of Zhejiang Province under Grant 2022C01082.
adaptation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision.

[50] Liang Qi, Mengchu Zhou, and Wenjing Luan. 2018. A dynamic road incident information delivery strategy to reduce urban traffic congestion. IEEE/CAA Journal of Automatica Sinica (2018).

[51] Paul Sturgess, Karteek Alahari, Lubor Ladicky, and Philip HS Torr. 2009. Combining appearance and structure from motion features for road scene understanding. In BMVC-British Machine Vision Conference.

[52] Peng Su, Kun Wang, Xingyu Zeng, Xiuxiang Tang, Dapeng Chen, Di Qiu, and Xiaogang Wang. 2020. Adapting object detectors with conditional domain normalization. In European Conference on Computer Vision.

[53] Jie-Young Sun, Seung-wook Kim, Sang-Won Lee, Ye-Won Kim, and Sung-Jea Ko. 2019. Reverse and boundary attention network for road segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops.

[54] Lei Su, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijayshai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. 2020. Scalability in perception for autonomous driving: Waymo open dataset. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.

[55] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Mannohman Chandraker. 2018. Learning to adapt structured output space for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition.

[56] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Mannohman Chandraker. 2019. Domain adaptation for structured output via discriminative patch representations. In Proceedings of the IEEE conference on computer vision and pattern recognition.

[57] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. 2019. Adven: Adversarial entropy minimization for domain adaptation in semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

[58] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. 2019. Dada: Depth-aware domain adaptation in semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

[59] Hengli Wang, Rui Fan, Peide Cai, and Ming Liu. 2021. Snr-roadseg+: Rethinking appearance and structure from motion features for road scene understanding. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision.

[60] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. 2015. Exploring sequence feature alignment for domain adaptation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision.

[61] Feifei Zhang, Tianzhu Zhang, Qirong Mao, Lingyu Duan, and Changsheng Xu. 2019. Prototype-based domain adaptation for 3D human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

[62] Xudong Wang, Zhaowei Cai, Dacheng Tao, and Hannu Teneketzis. 2019. Towards universal object detection by domain attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

[63] Zhedong Zheng and Yi Yang. 2019. Unsupervised scene adaptation with memory networks for unsupervised scene adaptation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

[64] Juanzhong Zhao, Zhidong Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. 2019. Scalability in perception for autonomous driving: Waymo open dataset. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision.
A UNIQUE POINTS OF THE UDA-FD TASK
It is true that UDA-FD shares some similarity with UDA for semantic segmentation (UDA-SS), with both two tasks aim at pixel-level classification. However, there are still difference between them, which lies mainly in two aspects:

The UDA-SS task labels almost each pixel with one of a closed set of known categories. However, UDA-FD task separates foreground pixels (i.e., drivable area) from background pixels which are composed of an open set of multiple unknown categories. Therefore, the uncertainty of background semantics leads to the difficulty to adapt between source and target domains. To alleviate this issue, we propose the SFA module which aligns only foreground features to diminish the disturbance of complex background feature. Besides, the uncertainty of background semantics also prevents UDA-SS methods based on minimizing entropy to perform well on the UDA-FD task.

In UDA-FD task, there are strong appearance similarity between foreground and some background categories such as sidewalk and curb, which are hard to distinguish from the foreground drivable area only based on RGB visual appearance. As a result, geometric properties, e.g. the planar properties of the road, are usually important cues in UDA-FD task to distinguish these confusing categories. In our method, we integrate the geometric surface normal information into our framework as a complement with RGB information in CCG module. As for UDA-SS task, such properties are relatively less explored since the confusing categories, such as sidewalk and curb, also have their own pixel-level annotations from which UDA-SS methods can directly learn.

B DATASETS
Ready-to-Drive (R2D) [16] is a large-scale synthetic freespace detection dataset collected with the CARLA [15] simulator. It consists of 6 different scenarios with 11,430 pairs of stereo images, together with their corresponding depth images and segmentation labels.

SYNTHIA-SF [26] is a synthetic semantic segmentation dataset consisting of 6 video sequences under different scenarios and traffic conditions. 2,224 images along with depth and segmentation labels are included in total.

Virtual KITTI [19] is a synthetic dataset generated from five different virtual worlds in urban settings under different imaging and weather conditions. In the experiment, we sample a total number of 2,126 frames for training.

KITTI-Road [18] is a real-world dataset for freespace detection. It contains 579 images in total with 289 for training and 290 for testing.

C EXPERIMENT RESULTS
Since there is only DADA [58] that uses depth modality in UDA-SS task, we integrate depth modality into other RGB-only methods by concatenating the depth feature with RGB feature extracted by separated encoders for comparison with our method. Results on R2D→KITTI-Road dataset are summarized in Table 5. As shown in the table, incorporating depth information to these UDA-SS methods can obviously improve their performance, but our method still outperforms these ones by large margins on F1 Score and IoU.

To further validate the effectiveness of our method, we adapt more recent works including PCLA [34] and CCM [38] to UDA-FD task and report the results on R2D → KITTI-ROAD with ResNet-101 as backbone in Table 6. Our method outperforms the two works in terms of both F1 Score and IoU.

### Table 5: Freespace detection performance on KITTI-Road validation set with R2D utilized as source datasets. We integrate depth information in RGB-only methods by concatenating depth map with RGB image as input to the model.

| Method      | Depth | PRE  | REC  | F1 score | IoU  |
|-------------|-------|------|------|----------|------|
| AdaptSeg [55] | √     | 52.71| 98.83| 68.75    | 52.38|
| AdaptSeg [55] | √     | 59.87| 98.93| 74.60    | 59.48|
| ADVENT [57] | √     | 54.24| 98.04| 69.84    | 53.65|
| ADVENT [57] | √     | 56.70| 98.32| 71.92    | 56.15|
| IntraDA [47] |       | 58.98| 97.53| 73.51    | 58.11|
| IntraDA [47] | √     | 60.38| 96.44| 75.23    | 60.53|
| Ours-ResNet50 | √   | 86.27| 92.84| 89.43    | 80.89|

### Table 6: Comparison with more recent works on R2D→KITTI-Road setting.

| Method      | Publication | PRE  | REC  | F1 Score | IoU  |
|-------------|-------------|------|------|----------|------|
| PCLA [34]   | NeurIPS 2020 | 57.74| 95.88| 72.07    | 56.34|
| CCM [38]    | ECCV 2020   | 63.25| 94.58| 75.80    | 61.04|
| Ours-ResNet50 | -         | 86.27| 92.84| 89.43    | 80.89|

D QUALITATIVE RESULTS
In Figure 5 and Figure 6, we visualize qualitative results of our model and other methods on Virtual KITTI → KITTI-Road and SYNTHIA-SF → KITTI-Road settings respectively. We can observe that AdaptSeg [55] and DADA [58] fail to distinguish sidewalks or grasslands from roads (e.g., 4-th column in Figure 5 and 2-nd column in Figure 6) and cannot completely segment the road area when there are shadows (e.g., 3-rd column in Figure 5). However, in the predictions of our method, there are much less mis-classified pixels in these regions, demonstrating the effectiveness of selective feature alignment and cross-modal guidance. As shown in the last row of Figure 5 and Figure 6, with the help of self-training, most false predictions on sidewalks are suppressed.
Figure 5: Qualitative comparison between our model and other methods on Virtual KITTI → KITTI-Road. The true positive, false negative and false positive pixels are respectively shown in green, blue and red.

Figure 6: Qualitative comparison between our model and other methods on SYNTHIA-SF → KITTI-Road. The true positive, false negative and false positive pixels are respectively shown in green, blue and red.