

Towards Robust Semantic Segmentation of Accident Scenes via Multi-Source Mixed Sampling and Meta-Learning

Xinyu Luo*, Jiaming Zhang*, Kailun Yang†, Alina Roitberg, Kunyu Peng, Rainer Stiefelhagen
CV:HCI Lab, Karlsruhe Institute of Technology

Abstract

Autonomous vehicles utilize urban scene segmentation to understand the real world like a human and react accordingly. Semantic segmentation of normal scenes has experienced a remarkable rise in accuracy on conventional benchmarks. However, a significant portion of real-life accidents features abnormal scenes, such as those with object deformations, overturns, and unexpected traffic behaviors. Since even small mis-segmentation of driving scenes can lead to serious threats to human lives, the robustness of such models in accident scenarios is an extremely important factor in ensuring safety of intelligent transportation systems.

In this paper, we propose a Multi-source Meta-learning Unsupervised Domain Adaptation (MMUDA) framework, to improve the generalization of segmentation transformers to extreme accident scenes. In MMUDA, we make use of Multi-Domain Mixed Sampling to augment the images of multiple-source domains (normal scenes) with the target data appearances (abnormal scenes). To train our model, we intertwine and study a meta-learning strategy in the multi-source setting for robustifying the segmentation results. We further enhance the segmentation backbone (SegFormer) with a HybridASPP decoder design, featuring large window attention spatial pyramid pooling and strip pooling, to efficiently aggregate long-range contextual dependencies. Our approach achieves a mIoU score of 46.97% on the DADA-seg benchmark, surpassing the previous state-of-the-art model by more than 7.50%.¹

1. Introduction

With the rapid development of computer vision algorithms in Intelligent Transportation Systems (ITS), road safety for Intelligent Vehicles (IV) has gradually become one of the most concerning issues in this community. The Advanced Driver Assistance Systems (ADAS) are required

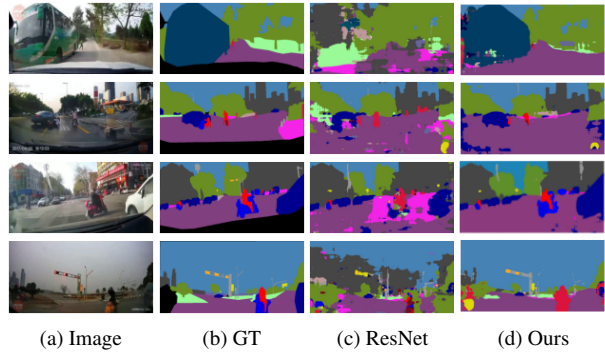


Figure 1. **Semantic segmentation of accident scenes.** Compared to the source-only model (e.g., a ResNet), our model generalizes better in the abnormal cases. From top to bottom are dash-crossing pedestrians, overturned motorcycles, collisions, and motion blur.

to correctly handle both, *normal* driving scenes, which are addressed by most of the published datasets, and *abnormal* situations (i.e., edge cases) that may unexpectedly appear in real-world scenes. Fueled by rapid improvements in general semantic segmentation research, great progress has been made in the field of autonomous driving [28, 31, 33] in recent years. However, these segmentation models are mainly designed for normal driving scenes, while real-life accidents often encounter abnormalities and critical situations, such as overturned vehicles in front of the ego-vehicle or distorted shots caused by motion blur. Several examples of such abnormal cases taken from accident scenes are presented in Fig. 1. If a standard semantic segmentation model, which does not see any abnormal samples during training, is deployed in real world, it can hardly obtain correct results when encountering an unusual accident- or near-accident scene, resulting in a failure of driving assistance. The large domain gap between the normal- and accident scenes negatively impacts segmentation performance [60], greatly limiting applications of autonomous driving in practice.

Despite its high relevance for applications, only a few works address the task of accident scenes segmentation, which aims to adapt models trained in normal scenarios to the abnormal ones. [60, 61] introduced DADA-seg – a new traffic accidents dataset covering labelled and unla-

*The first two authors contribute equally to this work.

†Corresponding author (e-mail: kailun.yang@kit.edu).

¹Code will be made publicly available at <https://github.com/xinyu-laura/MMUDA>.

belled images from real-world abnormal driving scenes. The event-aware ISSAFE architecture [60] was proposed to fuse RGB images and event data and therefore to capture the dynamic context. The Trans4Trans model [58] leveraged transformer-based encoder and decoder and was simply transferred, without any adaptation design, from multiple source datasets, *i.e.*, Cityscapes [8] and ACDC [35]. However, mixing data only in a dataset-wise manner is limited in terms of balancing the data distribution and further data diversification remains an important milestone. To the best of our knowledge, only the single-source Unsupervised Domain Adaptation (UDA) [60] and the multi-source transfer-learning method [58] from normal- to abnormal domain were investigated. *Multi-source normal-to-abnormal UDA, yet, remains unexplored.* We believe, that leveraging multiple data origins has strong potential for robustifying accident scene segmentation, as such extreme scenarios often contain diverse abnormal factors and composite scenes, that could be better addressed by exploiting rich ontologies covered in diverse sources.

To improve the robustness of semantic segmentation in accident scenarios, we propose a novel *Multi-source Meta-learning UDA* framework to help the transformer models in generalization to the unusual target scenes (*MMUDA* for short). Our framework learns from the label-rich datasets of conventional driving scenes (*source*), and then automatically adapts to the target domain of abnormal accident scenes with only unlabelled training data (*target*). To effectively learn from the entire unlabelled target domain dataset, we put forward a *Multi-Domain Mixed Sampling (MDMS)* strategy, which is inspired by the cross-domain mixing approach in DACS [41] and augments the training samples of multiple source domains. Two major differences compared to the *normal-to-normal* DACS are that *i)* we adapt the single-source method to a multi-source setup, and *ii)* we further investigate the multi-source mixing technique in our *normal-to-abnormal* setting. More precisely, in the case of single-domain mixed sampling, the augmented sample is formed by mixing the source normal image and the target abnormal image. In the case of the multi-domain mixing, some marks from each source domain image are extracted and then pasted onto the target domain image. The pseudo-labels for the augmented image are mixed by the source ground-truth labels and the target pseudo labels.

In the training phase, we use the meta-learning for domain generalization (MLDG) strategy [22], which was adapted in [57] to model the domain transfer problem with an episodic training paradigm, leading to superior performance in image classification. MLDG can be viewed as a regularization mechanism that prevents the model from overfitting. Different from the original MLDG, our MMUDA framework performs meta-learning across multiple source domains and jointly with the target domain, after

which we apply it to the normal-to-abnormal UDA setting. In addition, we enhance the segmentation backbone (SegFormer) with a *HybridASPP* decoder design, which leverages large window attention spatial pyramid pooling [46] and strip pooling [18] with a long but narrow kernel. The HybridASPP decoder replaces the vanilla MLP-based decoder of SegFormer, and this helps to efficiently extract large regions of global context and long-range dependencies. Comprehensive experiments demonstrate the effectiveness of our proposed methods. On the challenging DADA-seg benchmark [60], our approach achieves a mIoU score of 46.97%, surpassing the previous state-of-the-art transformer model [58] by more than +7.50%.

Our contributions are summarized as follows:

- We propose a novel *Multi-source Meta-learning UDA (MMUDA)* framework for better adaptation from multi-source domains of normal driving scenes to the domain of abnormal accident scenes.
- We develop a *Multi-Domain Mixed Sampling (MDMS)* approach to augment the training data from multiple labelled source domains with data appearances from the unlabelled target domain data.
- We employ meta-learning and analyze its effects under different combinations of multiple source datasets.
- We introduce an enhanced *HybridASPP* to replace the vanilla MLP-based decoder of SegFormer, which makes the framework more efficient and effective.

2. Related Work

Semantic segmentation. Semantic segmentation has experienced a great breakthrough since the emergence of FCN [25] classifying pixels end-to-end. Subsequent networks, *e.g.*, [2, 12, 18, 52, 64] improved FCN in different aspects, significantly pushing segmentation performance, but also raising computational cost. To alleviate this issue, compact segmentation models [28, 32, 33] are designed to hold a better accuracy-efficiency trade-off. Disentangled non-local blocks [48, 49, 67] have also been explored to efficiently collect omni-range dependencies. Recently, the semantic segmentation field has been driven by the newly emerged and highly effective transformer-based architectures [6, 38, 44, 46, 65]. Compared to FCN-based approaches, such models are able to handle long-range dependencies by design and have quickly climbed to the top of segmentation benchmarks. Furthermore, MLP-like architectures [4, 17, 40, 51, 59] alternate token- and channel-mixing to enhance global reasoning. The recently proposed SegFormer architecture [45] leverages a hierarchical transformer encoder with a lightweight All-MLP decoder, generating powerful representations without complex and computationally demanding modules. In this work, we build on SegFormer and introduce multiple building blocks for its improvement specifically for accident scene segmentation.

Domain adaptation and generalization. While current semantic segmentation models achieve excellent performance on standard benchmarks, the performance drops sharply if the training and test images come from different domains. To counter this effect, multiple methods based on Domain Adaptation (DA) [15, 34, 39, 62, 63] were proposed for automatic adjustment to adverse conditions. In [36], effective usage of synthetic data was explored to better handle domain shifts, with results indicating that the foreground objects should rather be addressed in a detection-based manner. Domain Generalization (DG) is more challenging than DA since DG methods can only access source domain data for training. Target images during the training process cannot be used or observed (while classical DA allows access to unlabelled target domain data). Currently, most DG methods focus on image classification, and only a few [7, 29, 57] are developed to solve semantic segmentation tasks. Another recent group of approaches [19, 53] raised the technique of domain randomization to improve DG. Moreover, open compound domain adaptation approaches [13, 24, 30] have been developed to adapt to a group of unknown heterogeneous domains like scenes in adverse environmental conditions. In this work, we also develop a domain transfer system based on meta-learning and design a multi-source mixed sampling method for enhancing semantic segmentation in accident scenarios.

Meta learning. Meta-learning aims at figuring out *how to learn* and has been effectively applied in different fields, with Model-Agnostic Meta-Learning (MAML) [11] and HyperNetworks [14] being popular approaches for image classification. MAML simulates the domain gap between train and test domains by synthesizing virtual testing domains within each mini-batch during training. As of late, meta-learning has likewise been effectively applied to domain adaptation and domain generalization. Meta-online [21] introduces a strategy to improve the results by learning the underlying circumstances (*e.g.*, model boundary) of the existing domain adaptation strategies. MLDG [22] follows the MAML [11] system and has been effectively used in the DG task. Similar to MAML, the DG task expects that the learned models in seen domains are able to generalize well to novel unseen domains. As a consequence, meta learning has been recently explored for domain adaptation and generalization [1, 5, 10, 13, 57]. In this work, we leverage meta learning to better make use of multi-source data in order to boost generalization of semantic segmentation models in extreme accident scenes.

Mixing. Mixing is a kind of augmentation technique and has successfully been used for both classification and semantic segmentation as suggested in [55]. Mixing has been further developed in the class mixing algorithm [27], where the masks used for mixing are dynamically created based on the predictions of the network. In DACS [41], the concept

of self-training is extended and combined with ClassMix. Their proposed method fine-tunes models with mixed labels by combining ground-truth annotations from the source domain and pseudo-labels from the target domain. Further, context-aware mixing [66] and bi-mix [47] methods are developed to better guide the domain mixup and bridge the distribution gap. In this work, we assemble a multi-source mixed sampling method within our system for robust semantic understanding of accident scenes.

3. Methodology

In this Section, we describe our *Multi-source Meta-learning UDA (MMUDA)* framework in detail. First, we explain the proposed *Multi-Domain Mixed Sampling (MDMS)* in Sec. 3.1. Then, the meta-learning strategy for multi-source UDA in Sec. 3.2 is explained in detail. Finally, Sec. 3.3 provides an overview of the proposed architectural changes and the enhanced *HybridASPP* decoder design.

3.1. MDMS: Multi-Domain Mixed Sampling

Our proposed method builds upon the idea of domain adaptation via cross-domain mixed sampling (DACS) [41]. Unlike DACS, which has only one source domain, our augmented samples are created by mixing pixels from the target domain image with pixels from each source domain image. Before being mixed, the unlabelled target domain images first need to be run through the model to generate pseudo-labels for them. Then, half of the classes in the image from source domain are randomly selected, and the pixels of the corresponding classes are cut from the source domain image and pasted onto the target domain image. For labels, the pseudo-labels of unlabelled target image are mixed with the corresponding ground-truth labels of the source domain image in the same way as the mixed images. The above mixing approach is applied to each source domain. For brevity, the mixed sampling process will be described with only one source domain below.

A source domain is defined by a set of image and label pairs $\{(X_S^i, Y_S^i)\}^{N_S} \in \mathcal{D}_S$, where $X_S^i \in \mathbb{R}^{H \times W \times 3}$ is the image, $Y_S^i \in \mathbb{R}^{H \times W \times C}$ is the C -class label, and N_S is the number of samples in the source domain \mathcal{D}_S . From the target domain \mathcal{D}_T with a number of $N_T = N_L + N_U$ samples, the N_U unlabelled image and pseudo label pairs $\{(X_T^i, \hat{Y}_T^i)\}^{N_U} \in \mathcal{D}_T$ are selected for the mixing approach, where \hat{Y}_T is generated by the segmentation transformer f_{seg} in Fig. 2. The labelled N_L images in the target domain are only used in the testing stage. In the augmented set \mathcal{D}_M with the same N_U samples, an augmented image X_M is generated by mixing a source image X_S and a target image X_T , and the pseudo label \hat{Y}_M by combining the corresponding ground-truth label Y_S and the pseudo label \hat{Y}_T . However, in our case, there is not one but K

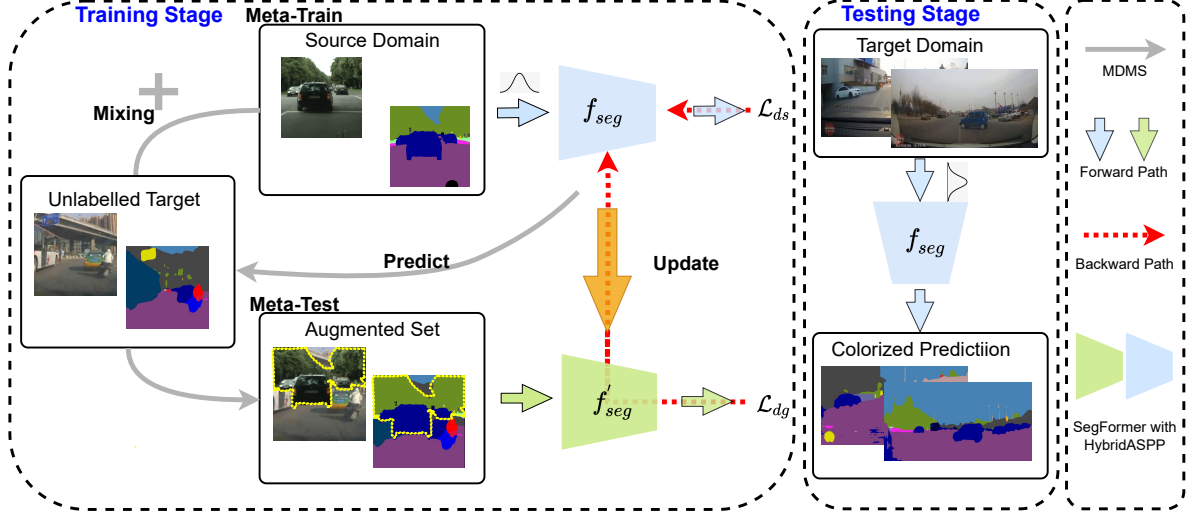


Figure 2. **Overview of the proposed Multi-source Meta-learning UDA (MMUDA) framework.** It includes *Multi-domain Mixed Sampling (MDMS)* and meta-learning with segmentation transformers. Given multiple source (normal) domains, the model fine-tuned by meta-training and meta-testing across various source domains with rich ontologies, can generalize well in the target (abnormal) domain.

source domains, thus the augmented set is finally created as $\{(X_{M_k}^i, \hat{Y}_{M_k}^i)\}^{N_U} \in \mathcal{D}_M$.

These new samples are then used to train the segmentation transformer model f'_{seg} . The process of MDMS is presented as gray arrows in Fig. 2, where an example of mixing image-label pairs from source- and target domain is shown. Pixels cut from the source are marked in yellow dotted boxes. In this way, we obtain a set of augmented samples from multiple source domains, which are then passed to the subsequent meta-learning-based network together with the original source domain images.

3.2. Meta-Learning in UDA

During the training phase, we adapt MLDG [22] (*i.e.*, meta-learning for domain generalization) to train our model. Different from MLDG, we perform meta-learning across multiple source domains and together with the target domain, intertwined via mixing sampling to obtain augmented images which instills the knowledge from the target scenes. Besides, instead of addressing image classification, we apply it to the normal-to-abnormal UDA setting for robustifying dense accident scene segmentation.

As shown in the training stage of our framework in Fig. 2, we use the images from sources for meta-training while the augmented images produced via MDMS are used for meta-testing. The cross-entropy is selected as the loss function \mathcal{L} for our semantic segmentation task. First, the domain-specific loss \mathcal{L}_{ds} is computed from the meta-training data through the network f_{seg} . The gradient $\nabla \mathcal{L}_{ds}$ is then used to update a new network f'_{seg} , *i.e.*, the green block in Fig. 2, which has the same backbone as the blue one. As we want the model to adapt well in the unseen target domain, the domain adaptation loss \mathcal{L}_{da} is calculated

from f'_{seg} with the updated parameters using the meta-test data. Finally, we employ the total loss $\mathcal{L}_{total} = \mathcal{L}_{da} + \alpha \mathcal{L}_{ds}$ to update the original f_{seg} , so that the network is optimized to perform well in the source and target domains. During the update for f'_{seg} , we use the inner learning rate η , while updating the original network f_{seg} after a meta-train and subsequent meta-test process, the outer learning rate γ is used. The parameters can then be updated with SGD (*i.e.*, Stochastic Gradient Descent).

In our segmentation task, the statistics in the source domain (*i.e.*, the *normal driving scenes*) are different from the target domain of traffic accident scenes. Therefore, normalizing the test data with the accumulated mean and variance during the training phase can be problematic. Considering this fact, we adopt target-specific normalization introduced by [57] to normalize features directly using statistics from the target domain in the testing stage. In the experiments, we further investigate different combinations of datasets for meta-learning, aiming for an optimal path to follow towards robust accident scene understanding.

3.3. HybridASPP

Next, we propose multiple improvements to the original SegFormer architecture in order to efficiently extract large regions of the global context and long-range dependencies. To this intent, our enhanced HybridASPP decoder replaces the vanilla MLP-based decoder of SegFormer. As shown in Fig. 3, in the first yellow branch, HybridASPP inherits the large window attention blocks of Large Window Attention Spatial Pyramid Pooling (LawinASPP) [46] to exploit global context information. By adjusting the ratio of the context regions to the query regions with $\{2, 4, 8\}$ as illustrated in Fig. 3, large-window attention is able to cap-

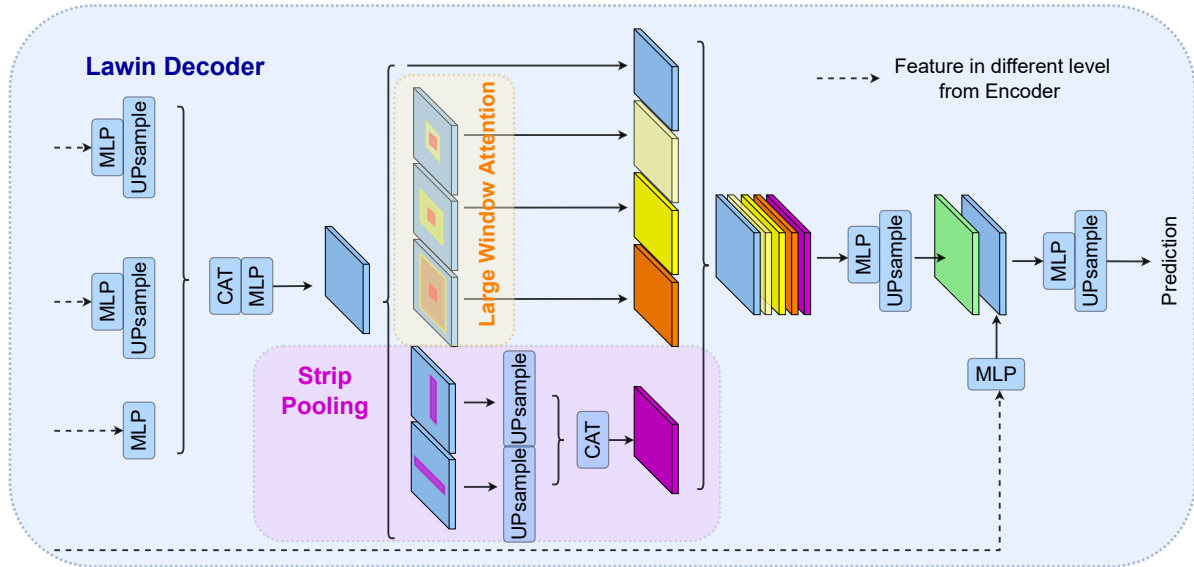


Figure 3. The structure of HybridASPP consists of large window attention and strip pooling to capture long-range context.

ture contextual information at multiple scales. The offered receptive fields are of sizes $\{16, 32, 64\}$, when setting the patch size of the local window to 8. Further, the study in [18] found that the standard spatial pooling operation with a shape of $H \times W$ makes the long-range context likely to contain many unrelated regions. In the bottom branch, we intertwine strip pooling modules, which encode long-range context along the horizontal and vertical spatial dimension based on long, narrow kernels ($1 \times W$ and $H \times 1$).

As shown in Fig. 3, we concatenate an aggregated feature from the last three stages of the encoder, the large window attended features, the strip pooling augmented feature, after which a learned linear transformation performs dimensionality reduction for producing the final segmentation map. The output of HybridASPP feature is upsampled to the size of a quarter of input image, then it is fused with the low-level feature coming from the first stage of the segmentation transformer (*i.e.*, SegFormer) via a linear layer. Last, dense semantic predictions, *i.e.*, the segmentation logits, are obtained from the final representation.

4. Experiments

4.1. Datasets

The statistics of the five source datasets (normal scenes) and the target dataset (abnormal accident scenes) that are used in our experiments are listed in Table 1.

Source datasets. For training, we leverage five semantic segmentation datasets as our multi-origin source domains: WildDash2 (**W**) [54], ACDC (**A**) [35], BDD10K (**B**) [50], IDD (**I**) [42], and Cityscapes (**C**) [8]. The datasets **A** and **B** and **C** are annotated with the same label set of 19 categories. Compared to these datasets, **W** has 5 five additional

Datasets	WildDash2	ACDC	BDD	IDD	Cityscapes	DADA-seg
#training	2,979	1,600	7,000	6,993	2,975	12,207
#evaluation	1,277	406	1,000	981	500	313

Table 1. Statistics of datasets for experiments.

classes, including *van*, *pickup*, *street light*, *billboard*, and *ego-vehicle*. Following the setting of [57], we merge the additional categories of **W** by mapping the added new Ids to the original Ids of **A**, **B**, and **C**. Although **I** also contains more classes, we use the provided public code² to directly generate the masks with the same label Ids as **C**.

W has a large collection of road scenes from different countries, weather and lighting conditions, including 2,979 training- and 1,277 validation images of size $1,920 \times 1,080$. **A** includes four common adverse conditions, *i.e.*, fog, nighttime, rain, and snow. Each of these conditions has 1,000 images: 400 for training, 100 for validation and 500 for testing (resolution $1,920 \times 1,080$). **B** has geographic, environmental, and weather diversity and provides 7,000 training images and 1,000 validation images, with a resolution of $1,280 \times 720$. **I** features a higher diversity of within-class appearance compared to **C**. A total of 10,004 images with 6,993 training- and 981 validation examples are captured from Indian roads with a resolution of $1,280 \times 964$. The **C** dataset contains street scenes of 50 different cities and 19 categories with high resolution of $2,048 \times 1,080$. The 5,000 images are divided into 2,975 training-, 500 validation-, and 1,525 test samples.

Target dataset. We utilize DADA-seg (**D**) [60] as our target dataset used for testing our approach, which covers 313 evaluation images. Following [60, 61], 12,207 *unlabelled* images are used for unsupervised adaptation. All images

²<https://github.com/AutoNUE/public-code>

	Method	mIoU	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle
Source-only	MobileNetV2 [37]	16.05	31.87	8.50	26.55	3.60	5.38	13.96	19.51	10.87	44.99	11.09	67.05	8.11	5.23	28.58	11.77	2.17	-	1.90	3.86
	PSPNet [64]	17.07	31.62	11.42	32.48	4.16	8.52	12.38	17.93	13.39	50.82	13.85	67.19	9.86	3.13	31.54	6.97	3.15	-	2.97	2.89
	ResNet50 [16]	18.96	34.19	8.24	31.05	4.56	7.39	19.04	27.05	15.35	33.30	12.40	61.52	10.04	3.95	42.59	14.15	27.02	-	3.72	4.72
	SemFPN [20]	19.59	37.90	10.12	23.80	3.74	9.64	22.06	28.64	15.55	40.95	12.13	51.93	9.24	5.93	52.08	13.89	26.54	-	3.66	4.36
	DNLNet [49]	19.72	41.68	13.26	30.45	6.17	11.04	21.91	28.03	17.99	40.05	14.13	56.06	10.75	5.41	34.78	8.01	28.01	-	3.55	3.39
	ResNeSt [56]	19.99	39.63	11.38	33.68	2.81	9.73	22.76	27.35	18.09	45.24	14.22	71.23	13.34	5.03	36.45	6.91	13.08	-	3.94	4.87
	DANet [12]	22.24	46.49	10.17	42.20	3.81	10.65	13.46	18.69	22.59	55.76	22.22	83.84	6.68	11.75	39.59	7.96	12.64	-	7.98	6.12
	ResNet101 [16]	23.60	57.96	11.16	39.94	6.43	9.46	23.67	27.37	17.32	45.65	16.47	69.21	13.19	4.51	47.29	13.75	30.44	-	6.64	8.01
	OCRNet [52]	24.85	42.13	11.54	34.49	6.63	12.70	22.76	29.03	22.28	42.41	15.15	85.43	14.31	6.65	53.94	20.65	34.86	-	9.30	7.87
	FastSCNN [32]	26.32	69.91	16.30	52.53	6.09	9.63	19.98	19.30	22.58	57.04	22.95	90.81	11.19	13.95	46.16	22.65	9.74	-	4.49	4.75
Cross-source	CLAN [26]	28.76	79.80	18.61	51.56	8.32	13.60	15.51	17.15	21.51	63.20	21.99	80.53	8.37	6.32	63.47	33.43	33.12	-	3.69	6.21
	BDL [23]	29.66	81.44	19.18	57.18	8.61	16.26	14.65	8.78	16.77	66.60	26.83	85.87	10.51	7.16	65.45	35.18	34.78	-	2.71	5.57
	ISSAFE [60]	29.97	80.23	19.51	52.02	6.43	14.68	16.19	17.03	19.50	65.39	21.69	79.84	9.95	8.82	65.60	39.51	39.73	-	6.09	7.03
	EDCNet [61]	32.04	73.03	19.47	57.31	11.60	14.30	20.70	12.27	27.22	70.54	18.98	88.64	10.69	8.70	68.14	49.80	50.86	-	9.02	4.12
	Trans4Trans-M [58]	39.20	71.10	15.57	70.39	10.34	16.53	31.63	37.16	37.38	71.88	19.61	93.04	21.27	14.97	64.04	53.76	81.53	-	24.63	10.07
	Our Baseline	40.73	84.93	23.66	68.34	16.27	20.58	25.96	31.25	28.20	71.89	22.39	93.16	17.92	26.84	73.89	55.09	69.26	-	34.77	9.49
	+Meta	45.03	86.20	25.44	70.63	14.21	19.75	26.56	28.01	29.23	74.45	25.29	93.18	20.40	31.53	75.02	64.73	76.84	-	38.05	10.95
	+Meta+MDMS	46.11	87.10	27.71	71.11	22.94	20.64	32.25	29.49	34.34	75.48	24.02	92.18	20.65	33.33	74.64	63.35	71.14	-	39.08	12.04
	Our MMUDA	46.97	87.51	27.97	74.76	16.16	21.93	29.94	29.43	31.62	75.67	26.69	93.57	24.40	29.57	77.35	68.24	84.02	-	36.96	10.44

Table 2. **Comparison of state-of-art methods.** The source-only models are trained on the Cityscapes dataset, while the other models are domain-transferred using a single source [23, 26], multiple sources [58, 61] or a different modality [60]. While our baseline model has no mixed sampling and performs normal supervised learning based on ResNet101 only using the source domain, our MMUDA framework based on the SegFormer backbone uses MDMS and meta-learning (Meta for short) across the source and target domain.

are captured in abnormal driving scenes, *i.e.*, *traffic accident scenes*. The images of **D** are labelled with 19 classes, which are consistent with the classes of the **C** dataset (Cityscapes). The resolution of the images is 1, 584×660.

4.2. Implementation Details

Our approach leverages the ImageNet1K-pretrained MiT-B2 SegFormer [45] as the encoder backbone, and the public mmsegmentation framework implementation³. The meta-learning inner and outer learning rates (*i.e.* η and γ) are set to $1e^{-3}$ and $5e^{-3}$, respectively, with Polynomial learning rate decay with the power 0.9. The network is trained for 120 epochs, unless otherwise specified. The weight α of the \mathcal{L}_{ds} loss is set to 1. We train the model with a mini-batch size of 1 using stochastic gradient descent (SGD), momentum of 0.9 and weight decay of $5e^{-4}$. For training data augmentation, we use random resizing with ratio 0.5 to 2.0, random flipping, random Gaussian blur, and random cropping with a size of 600×600. Mean Intersection over Union (mIoU) is our main evaluation metric.

4.3. Results of Accident Scenes Segmentation

Table 2 compares the mean IoU and the per-class IoU scores achieved by our proposed approach to the state-of-the-art methods on DADA-seg. The source-only models trained on Cityscapes experience a considerable performance degradation and achieve a rather low accuracy when deployed in abnormal accident scenes. For example, the ResNet101 [16], whose results are illustrated in Fig. 1, only reaches 23.60% in mIoU. The previous state-of-the-art Trans4Trans model [58] attains 39.20% in mIoU

³<https://github.com/open-mmlab/msegmentation>

with a vision transformer and multi-source training. Our proposed MMUDA model surpasses all previous results, yielding a significantly higher recognition rate of 46.97% in mIoU, which is >7.50% higher than the past state-of-the-art. For per-class IoU, our approach achieves the best scores in 16 out of all 19 categories. The improvements over Trans4Trans are compelling (>10.00% performance gain) on categories which are safety-critical for accidental scene understanding, in particular, for *road*, *sidewalk*, *rider*, *car*, *truck*, and *motorcycle*.

The ablation results of our proposed modules are also depicted in Table 2, where all experiments are based on all five multi-origin source datasets described in Sec. 4.1. Our baseline model with ResNet101 uses only the normal source-supervised learning by aggregating multiple source domains and obtains a mIoU of 40.76%, whereas the model with meta-learning (+Meta) leads to a further 4.26% improvement. In addition, our proposed MDMS and transformer model with HybridASPP further elevate mIoU to 46.11% and 46.97%, respectively. Overall, these results demonstrate the effectiveness of the proposed framework and clear benefits of the MDMS module and the transformer model with HybridASPP, as well as the importance of our multi-source meta-learning training strategy.

4.4. Ablation Study of Meta-learning

Effect of meta-learning. The previously described Table 2 indicates clear advantages of meta-learning over the baseline. Next, we study the impact of our meta-learning strategy in more detail though an ablation study featuring a variety of source datasets (Table 3). All of these experiments employ the target-specific normalization and ResNet101 pretrained on ImageNet1K [9] as the backbone.

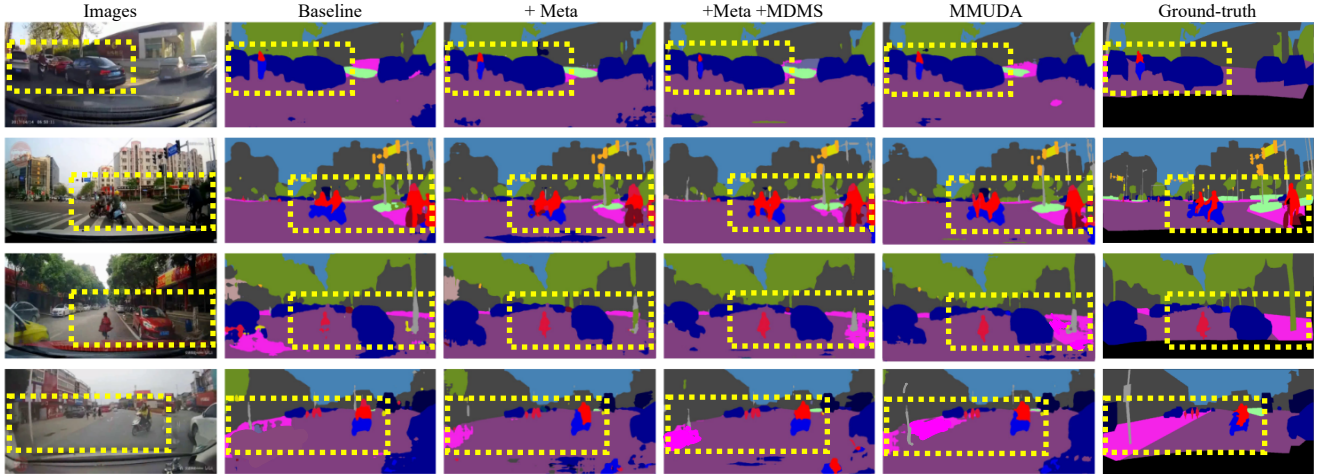


Figure 4. **Qualitative ablation study** of different modules used in our framework.

Source domain	mIoU(%)	
	Meta	without Meta
WildDash2 [54]	-	35.35
ACDC [35]	-	23.56
BDD [50]	-	31.18
IDD [42]	-	28.15
Cityscapes [8]	-	25.48
W + I + C	43.74	40.01
W + A + B	42.70	38.75
W + A + I	43.36	39.98
W + B + I	43.90	40.20
B + I + C	41.03	39.03
W + B + I + C	44.67	40.71
W + A + B + I + C	45.03	40.73

Table 3. **Ablation of meta-learning** with different source domains and their combinations.

The baseline models are trained with traditional learning on different single sources, respectively. When using multiple sources, meta-learning-based approaches (*e.g.*, 43.74% with **W+I+C**) are evidently more effective than the common aggregated-source-supervised learning (*e.g.*, 40.01%).

Influence of multi-source combination. In Table 3, we investigate the impact of combining different datasets on the segmentation performance. In single-source studies, training with **WildDash2**, **BDD**, and **IDD**, respectively, yields the top three mIoU scores thanks to their especially diverse examples. For example, **WildDash2** leads to a decent performance, as it offers many composite scenes and various visual hazards. As a result, when these three sources are used together, the combination of **W+B+I** yields the highest accuracy. In general, leaning with more sources improves the performance for accident scene segmentation, and our five-source meta-learning model attains 45.03% in mIoU, clearly standing out in front of all other models.

Method	GFLOPs↓	mIoU(%)↑	
		Cityscapes	DADA-seg
SegFormer + Vanilla MLP [45]	717.1	74.00	18.50
SegFormer + LawinASPP [46]	569.6	75.86	25.16
SegFormer + HybridASPP (Ours)	553.8	76.41	25.21

Table 4. **Ablation of decoders.** GFLOPs are calculated with a size of $2,048 \times 1,024$. We train all models on **C** on a single GPU with a batch size of 1 for $80k$ iterations and an input size of 768×768 .

4.5. Ablation Study on HybridASPP

To analyze the efficiency and effectiveness of our HybridASPP module, we replace the decoder of SegFormer-B2 [45], and compare it to LawinASPP [46] and the proposed decoder in Table 4. In terms of computation costs, HybridASPP reduces the GFLOPs of LawinASPP by 15.8 after introducing strip pooling to capture long-range context instead of using traditional image pooling. Looking at the performance, our decoder on **D** is evidently more reliable than Vanilla MLP. Meanwhile, our HybridASPP achieves higher mIoU scores than LawinASPP with less GFLOPs.

4.6. Model Efficiency Analysis

To investigate the efficiency of MMUDA, we compare it with state-of-the-art approaches evaluated on DADA-seg following [58]. Our model employs MiT-B2 [45] as the encoder and the proposed HybridASPP as the decoder. The comparison of mIoU, GFLOPs, and #Params is shown in Table 5. Compared to DeepLabV3+ [3] and HRNet [43], our model largely improves the performance, while saving a great amount of computation demands. In comparison with the previous state-of-the-art Trans4Trans-M, we obtain a significant gain of 7.8% in mIoU with 19.1M less parameters and an increase in GFLOPs. Capturing rich contextual cues requires more computation but also ensures the robustness of the model. While our model achieves high efficiency in general, in future work, we will consider a more lightweight encoder to further reduce the computational cost.

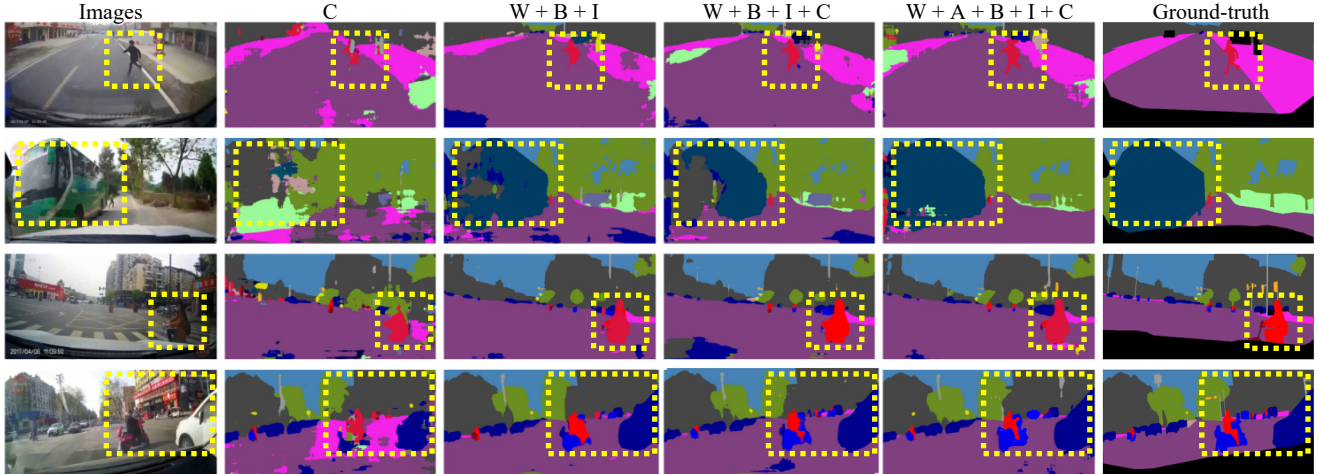


Figure 5. **Qualitative analysis** of the effect of using different numbers of source domains in our multi-source meta-learning framework.

Method	mIoU (%) \uparrow	GFLOPs \downarrow	#Params (M) \downarrow
DeepLabV3+ [3]	26.8	178.1	18.70
HRNet [43]	27.5	210.5	65.86
Trans4Trans-M [58]	39.2	41.9	49.55
Ours	47.0	105.5	30.49

Table 5. **Model efficiency analysis.** GFLOPs are calculated with the input size of 768×768 .

4.7. Qualitative Analysis

Effect of MMUDA. In our final study, we showcase multiple examples of representative qualitative results in Fig. 4, which corresponds to the numerical results in Table 2. Our baseline model produces relatively noisy segmentation results, such that the rushing person in the third row can not be completely segmented, whereas the model using meta-learning method makes more accurate predictions in this regard. When a collision happens, the model with the additive MDMS module is better at distinguishing between the *motorcyclist* and the *motorcycle*. Furthermore, the segmentation of *sidewalk* is also improved. In comparison, the noise of the predictions made by our MMUDA model with HybridASPP is much lower and the model segments the sidewalks well even under low light and with occlusion. Overall, our model adapts well to the accident scenario and is capable of providing robustness-improved semantic segmentation for the safety of autonomous driving.

Effect of source data. Fig. 5 visualizes the performance of our model when using different numbers of source domains. The images contain distortion and blur of foreground objects. It can be seen that the Cityscapes-trained model yields fragmented segmentation that disqualifies its application in self-driving scenarios, as it poses great threats in potential accident scenes. Comparing the predictions in the dashed box, our MMUDA provides more robust segmentation results when intertwining more sources. Since DADA-seg is a complex driving scenario, the model will be more ef-

fective when the sources provide more diverse information. The five-source meta-learned model clearly robustifies and improves semantic segmentation of accident scenes, and we believe that the high-quality predictions can be propagated to the upper-level driving applications.

5. Conclusion

Semantic segmentation of road scenes is a key ingredient for safe autonomous driving, but requires models that reliably operate under unusual circumstances. In this work, we specifically focus on segmentation of *abnormal* accident scenes since unexpected objects or traffic scenarios forms one of the common cause of dangerous situations. To tackle this challenge, we introduce a new framework which transfers knowledge from the well-studied domain of *standard* image segmentation to our target domain of *abnormal* scenes. Our *Multi-source Meta-learning UDA (MMUDA)* framework leverages multi-domain mixed sampling targeting at a better adaptation to the unknown accident scenes and is optimized using meta-learning. We further introduce a HybridASPP decoder design to improve the SegFormer segmentation backbone, which proved to be effective for our task. We verify the robustness of our model through extensive quantitative and qualitative experiments on the public DADA-seg benchmark, demonstrating superior generalization ability to abnormal accident scenes and surpassing previous state-of-the-art by a large margin.

Limitation and broader impact. The model training is limited by not using the augmented set as the meta-train set. To address this issue, a potential approach is to combine multiple source domains and target domain into a fusion set, and then to conduct a cross-combination meta-learning process. We leave it to our further research. Besides, the current experiments are conducted based on the referred datasets, thus there are still data biases when the model is applied in real-world test fields.

References

- [1] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. MetaReg: Towards domain generalization using meta-regularization. In *NeurIPS*, 2018. 3
- [2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *TPAMI*, 2018. 2
- [3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 7, 8
- [4] Shoufa Chen, Enze Xie, Chongjian Ge, Ding Liang, and Ping Luo. CycleMLP: A MLP-like architecture for dense prediction. In *ICLR*, 2022. 2
- [5] Ziliang Chen, Jingyu Zhuang, Xiaodan Liang, and Liang Lin. Blending-target domain adaptation by adversarial meta-adaptation networks. In *CVPR*, 2019. 3
- [6] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021. 2
- [7] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T. Kim, Seungryong Kim, and Jaegul Choo. RobustNet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *CVPR*, 2021. 3
- [8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 2, 5, 7
- [9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 6
- [10] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *NeurIPS*, 2019. 3
- [11] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017. 3
- [12] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, 2019. 2, 6
- [13] Rui Gong, Yuhua Chen, Danda Pani Paudel, Yawei Li, Ajad Chhatkuli, Wen Li, Dengxin Dai, and Luc Van Gool. Cluster, split, fuse, and update: Meta-learning for open compound domain adaptive semantic segmentation. In *CVPR*, 2021. 3
- [14] David Ha, Andrew M. Dai, and Quoc V. Le. Hypernetworks. In *ICLR*, 2017. 3
- [15] Jianzhong He, Xu Jia, Shuaijun Chen, and Jianzhuang Liu. Multi-source domain adaptation with collaborative learning for semantic segmentation. In *CVPR*, 2021. 3
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6
- [17] Qibin Hou, Zihang Jiang, Li Yuan, Ming-Ming Cheng, Shuicheng Yan, and Jiashi Feng. Vision permutator: A permutable MLP-like architecture for visual recognition. *TPAMI*, 2022. 2
- [18] Qibin Hou, Li Zhang, Ming-Ming Cheng, and Jiashi Feng. Strip pooling: Rethinking spatial pooling for scene parsing. In *CVPR*, 2020. 2, 5
- [19] Jiaying Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. FSDR: Frequency space domain randomization for domain generalization. In *CVPR*, 2021. 3
- [20] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *CVPR*, 2019. 6
- [21] Da Li and Timothy Hospedales. Online meta-learning for multi-source and semi-supervised domain adaptation. In *ECCV*, 2020. 3
- [22] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018. 2, 3, 4
- [23] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *CVPR*, 2019. 6
- [24] Ziwei Liu, Zhongqi Miao, Xingang Pan, Xiaohang Zhan, Dahua Lin, Stella X. Yu, and Boqing Gong. Open compound domain adaptation. In *CVPR*, 2020. 3
- [25] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 2
- [26] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *CVPR*, 2019. 6
- [27] Viktor Olsson, Wilhelm Traneheden, Juliano Pinto, and Lennart Svensson. ClassMix: Segmentation-based data augmentation for semi-supervised learning. In *WACV*, 2021. 3
- [28] Marin Orsic, Ivan Kreso, Petra Bevanđić, and Sinisa Segvić. In defense of pre-trained ImageNet architectures for real-time semantic segmentation of road-driving images. In *CVPR*, 2019. 1, 2
- [29] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via IBN-net. In *ECCV*, 2018. 3
- [30] Kwanyong Park, Sanghyun Woo, Inkyu Shin, and In So Kweon. Discover, hallucinate, and adapt: Open compound domain adaptation for semantic segmentation. In *NeurIPS*, 2020. 3
- [31] Kunyu Peng, Juncong Fei, Kailun Yang, Alina Roitberg, Jiaming Zhang, Frank Bieder, Philipp Heidenreich, Christoph Stiller, and Rainer Stiefelhagen. MASS: Multi-attentional semantic segmentation of LiDAR data for dense top-view understanding. *T-ITS*, 2022. 1
- [32] Rudra P. K. Poudel, Stephan Liwicki, and Roberto Cipolla. Fast-SCNN: Fast semantic segmentation network. In *BMVC*, 2019. 2, 6
- [33] Eduardo Romera, Jose M. Alvarez, Luis Miguel Bergasa, and Roberto Arroyo. ERFNet: Efficient residual factorized ConvNet for real-time semantic segmentation. *T-ITS*, 2018. 1, 2

- [34] Eduardo Romera, Luis Miguel Bergasa, Kailun Yang, Jose M. Alvarez, and Rafael Barea. Bridging the day and night domain gap for semantic segmentation. In *IV*, 2019. 3
- [35] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *ICCV*, 2021. 2, 5, 7
- [36] Fatemeh Sadat Saleh, Mohammad Sadegh Aliakbarian, Mathieu Salzmann, Lars Petersson, and Jose M. Alvarez. Effective use of synthetic data for urban scene semantic segmentation. In *ECCV*, 2018. 3
- [37] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobilenetV2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018. 6
- [38] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segformer: Transformer for semantic segmentation. In *ICCV*, 2021. 2
- [39] Lei Sun, Kaiwei Wang, Kailun Yang, and Kaite Xiang. See clearer at night: Towards robust nighttime semantic segmentation through day-night image conversion. In *SPIE*, 2019. 3
- [40] Ilya O. Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. MLP-mixer: An all-MLP architecture for vision. In *NeurIPS*, 2021. 2
- [41] Wilhelm Truhedden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. DACS: Domain adaptation via cross-domain mixed sampling. In *WACV*, 2021. 2, 3
- [42] Girish Varma, Anbumani Subramanian, Anoop M. Nambodiri, Manmohan Chandraker, and C. V. Jawahar. IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *WACV*, 2019. 5, 7
- [43] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2021. 7, 8
- [44] Ying Wang, Chiunan Ho, Wenju Xu, Ziwei Xuan, Xudong Liu, and Guo-Jun Qi. Dual-flattening transformers through decomposed row and column queries for semantic segmentation. *arXiv preprint arXiv:2201.09139*, 2022. 2
- [45] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021. 2, 6, 7
- [46] Haotian Yan, Chuang Zhang, and Ming Wu. Lawin transformer: Improving semantic segmentation transformer with multi-scale representations via large window attention. *arXiv preprint arXiv:2201.01615*, 2022. 2, 4, 7
- [47] Guanglei Yang, Zhun Zhong, Hao Tang, Mingli Ding, Nicu Sebe, and Elisa Ricci. Bi-mix: Bidirectional mixing for domain adaptive nighttime semantic segmentation. *arXiv preprint arXiv:2111.10339*, 2021. 3
- [48] Kailun Yang, Jiaming Zhang, Simon Reiß, Xinxin Hu, and Rainer Stiefelwagen. Capturing omni-range context for omnidirectional segmentation. In *CVPR*, 2021. 2
- [49] Minghao Yin, Zhuliang Yao, Yue Cao, Xiu Li, Zheng Zhang, Stephen Lin, and Han Hu. Disentangled non-local neural networks. In *ECCV*, 2020. 2, 6
- [50] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020. 5, 7
- [51] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. MetaFormer is actually what you need for vision. In *CVPR*, 2022. 2
- [52] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *ECCV*, 2020. 2, 6
- [53] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *ICCV*, 2019. 3
- [54] Oliver Zendel, Katrin Honauer, Markus Murschitz, Daniel Steininger, and Gustavo Fernandez Dominguez. WildDash - Creating hazard-aware benchmarks. In *ECCV*, 2018. 5, 7
- [55] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2017. 3
- [56] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Mueller, R. Manmatha, Mu Li, and Alexander J. Smola. ResNeSt: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020. 6
- [57] Jian Zhang, Lei Qi, Yinghuan Shi, and Yang Gao. Generalizable model-agnostic semantic segmentation via target-specific normalization. *PR*, 2022. 2, 3, 4, 5
- [58] Jiaming Zhang, Kailun Yang, Angela Constantinescu, Kunyu Peng, Karin Müller, and Rainer Stiefelwagen. Trans4Trans: Efficient transformer for transparent object and semantic scene segmentation in real-world navigation assistance. *T-ITS*, 2022. 2, 6, 7, 8
- [59] Jiaming Zhang, Kailun Yang, Chaoxiang Ma, Simon Reiß, Kunyu Peng, and Rainer Stiefelwagen. Bending reality: Distortion-aware transformers for adapting to panoramic semantic segmentation. In *CVPR*, 2022. 2
- [60] Jiaming Zhang, Kailun Yang, and Rainer Stiefelwagen. IS-SAFE: Improving semantic segmentation in accidents by fusing event-based data. In *IROS*, 2021. 1, 2, 5, 6
- [61] Jiaming Zhang, Kailun Yang, and Rainer Stiefelwagen. Exploring event-driven dynamic context for accident scene segmentation. *T-ITS*, 2022. 1, 5, 6
- [62] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *CVPR*, 2021. 3
- [63] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *ICCV*, 2017. 3
- [64] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 2, 6

- [65] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H. S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, 2021. 2
- [66] Qianyu Zhou, Zhengyang Feng, Qiqi Gu, Jiangmiao Pang, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. Context-aware mixup for domain adaptive semantic segmentation. *arXiv preprint arXiv:2108.03557*, 2021. 3
- [67] Zhen Zhu, Mengde Xu, Song Bai, Tengeng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In *ICCV*, 2019. 2