AGDF-Net: Learning Domain Generalizable Depth Features With Adaptive Guidance Fusion

Lina Liu, Xibin Song, Mengmeng Wang, Yuchao Dai, *Member, IEEE*, Yong Liu, *Member, IEEE*, and Liangjun Zhang

***Abstract*—Cross-domain generalizable depth estimation aims to estimate the depth of target domains (i.e., real-world) using models trained on the source domains (i.e., synthetic). Previous methods mainly use additional real-world domain datasets to extract depth specific information for cross-domain generalizable depth estimation. Unfortunately, due to the large domain gap, adequate depth specific information is hard to obtain and interference is difficult to remove, which limits the performance. To relieve these problems, we propose a domain generalizable feature extraction network with adaptive guidance fusion (AGDF-Net) to fully acquire essential features for depth estimation at multi-scale feature levels. Specifically, our AGDF-Net first separates the image into initial depth and weak-related depth components with reconstruction and contrary losses. Subsequently, an adaptive guidance fusion module is designed to sufficiently intensify the initial depth features for domain generalizable intensified depth features acquisition. Finally, taking intensified depth features as input, an arbitrary depth estimation network can be used for real-world depth estimation. Using only synthetic datasets, our AGDF-Net can be applied to various real-world datasets (i.e., KITTI, NYUDv2, NuScenes, DrivingStereo and CityScapes) with state-of-the-art performances. Furthermore, experiments with a small amount of real-world data in a semi-supervised setting also demonstrate the superiority of AGDF-Net over state-of-the-art approaches.**

***Index Terms*—Depth estimation, domain generalization, domain generalizable depth features, adaptive guidance fusion.**

I. Introduction

MONOCULAR depth estimation is an important perception task that has been widely applied in many applications, such as autonomous driving [1], [2], 3D scene reconstruction [3], [4] and augmented reality [5], [6], etc. Promising results have been achieved by deep convolution neural networks (DCNNs) based supervised depth estimation methods [7], [8], [9], [10], which mainly use the annotated depth ground-truth to supervise the DCNNs to estimate well depth maps. However, supervised learning requires large amounts of depth acquired by the sensors and aligning with the images as ground truth, which are costly and time-consuming [11], [12]. Therefore, approaches [13], [14], [15] have been proposed to use video sequences or stereo images to estimate depth in a self-supervised manner. However, consecutive video sequences and stereo images are not always available in current datasets, and the results are commonly limited to a single training dataset, which is difficult to generalize to different unseen scenes.

To relieve the problems, several solutions are proposed, using synthetic data that depth annotations can be directly obtained for training, and testing in real data (synthetic to real) to generalize to multiple real scenarios, which we summarized as: (1) Direct approaches: Some methods [7], [13] try to estimate real-world depth using models trained by synthetic datasets that are easy to obtain ground truth annotations. However, there is a vast gap between synthetic and real-world domains, which essentially limits the performance of real-world depth estimation. (2) Domain adaptation based methods: In order to bridge the gap from synthetic to real, an intuitive consideration is to directly convert the synthetic domain images into the real, including reducing the gap of the domain in the image level [12], [16] or feature level [17], [18], [19]. The above methods are domain adaptation based methods that require both synthetic domain and real domain images for training to reduce the domain gap. Nevertheless, a large number of real-world images of various scenes are hard to obtain. (3) Domain generalization based methods [20]: To relieve the above limitations, synthetic to real domain generalization methods only use labeled synthetic data for training, and directly test in several real data scenes, which is a more difficult task because the style of real data cannot be obtained during training. These approaches aim to learn a depth specific feature map for depth estimation of both synthetic and real-world images using only synthetic datasets, and enhancement operations are commonly utilized on the depth specific feature map, thus obtaining good generalization results on the real datasets. However, the pre-trained encoder obtained with both synthetic and additional real-world images is commonly needed to obtain the depth specific features in these

Manuscript received 22 March 2023; revised 17 October 2023; accepted 28 November 2023. Date of publication 13 December 2023; date of current version 3 April 2024. This work was supported in part by Baidu Research, the National Natural Science Foundation of China under Grants U21A20484 and 62271410. Recommended for acceptance by G. Carneiro. *(Corresponding authors: Xibin Song; Yong Liu.)*

Lina Liu is with the Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou 310027, China, also with China Mobile Research Institute, Beijing 100053, China, and also with Robotics and Autonomous Driving Lab, Baidu Research, Beijing 100085, China (e-mail: linaliu@zju.edu.cn).

Xibin Song and Liangjun Zhang are with Robotics and Autonomous Driving Lab, Baidu Research, Beijing 100085, China (e-mail: song.sducg@gmail.com; liangjunzhang@baidu.com).

Mengmeng Wang and Yong Liu are with the Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou 310027, China (e-mail: mengmengwang@zju.edu.cn; yongliu@iipc.zju.edu.cn).

Yuchao Dai is with the School of Electronics and Information, Northwestern Polytechnical University, Xi'an 710071, China (e-mail: daiyuchao@gmail.com).

Digital Object Identifier 10.1109/TPAMI.2023.3342634

approaches, which brings in new domain problems and limits the performance of the depth specific feature map. Besides, simple attention operation is commonly utilized to enhance the depth specific feature map, and limited improvement can be obtained.

In order to relieve the above limitations, we expect to find the essential expression of information for depth estimation. We propose a novel domain generalization based framework, which learns multi-scale domain generalizable depth features from the feature level for depth estimation, and only synthetic data is used for training without any real data. Previous work [20], [21] has proved that the structures and textures of images play key important roles in observing depth, while style and illumination, etc., are disturbance terms for depth perception. Therefore, we aim to relieve the influence of disturbance terms and extract domain invariant components from the images for depth estimation. To obtain domain invariant components with better generalization, we first extract initial depth features from the image with reconstruction and contrary losses, and enhance the initial depth features using the adaptive guidance fusion module to strengthen the domain invariant information inspired by [17], [22], and finally obtaining intensified depth features that are domain generalizable for input into the subsequent depth estimation network. Note that the domain generalizable depth features are invariant for different domains. And we expect to learn domain generalizable depth features at the multi-scale feature levels, thus obtaining better generalized depth maps and further narrowing the synthetic-to-real domain gap.

To obtain domain generalizable features for depth estimation, we propose a framework, i.e., AGDF-Net, to separate images into multi-scale initial depth and weak-related depth parts using two network branches constrained by reconstruction loss and contrary loss. The initial depth part should be depth related information, which is similar to [20], [21], while the weak-related depth part should contain the disturbance terms of depth estimation. Then, the extracted initial depth features are enhanced by the adaptive guidance fusion module to obtain intensified depth features. Specifically, at each scale, the intensified depth features of the previous scale are used to enhance the extracted initial depth features with the adaptive guidance fusion module. The largest scale is guided by features extracted from the color image. The purpose of this adaptive guidance fusion module is to make the network pay more attention to the domain invariant parts. This module can enhance the initial depth features to further eliminate the disturbing information and recover the domain invariant information as well. Finally, the intensified depth features that are domain generalizable are obtained for subsequent arbitrary depth estimation networks, and constrained by depth loss. In general, the practical design of our framework (initial depth branch, weak-related depth branch and adaptive guidance fusion module) and different kinds of losses (contrary loss, reconstruction loss and depth loss) used to constrain feature extraction can help obtain domain generalizable features for depth estimation.

Fig. 1(b) and (c) show the visualization of the initial depth and weak-related depth features extracted by our AGDF-Net. The initial depth features contain more obvious structural information, especially on the object areas and the corresponding edges. Weak-related depth features are the remaining information separated from the image. Fig. 1(d) shows the intensified depth features, which are enhanced in domain invariant areas for depth estimation, and further weakened and almost eliminated in object interiors texture, etc. Meanwhile, the depth estimation results in Fig. 1(e) have distinct object boundaries, which also demonstrate generalization abilities in both synthetic and real domains. Note that the first and second columns in Fig. 1 show the results of different style images with similar initial depth features and almost the same intensified depth features, while weak-related depth features have different strengths for different styles.

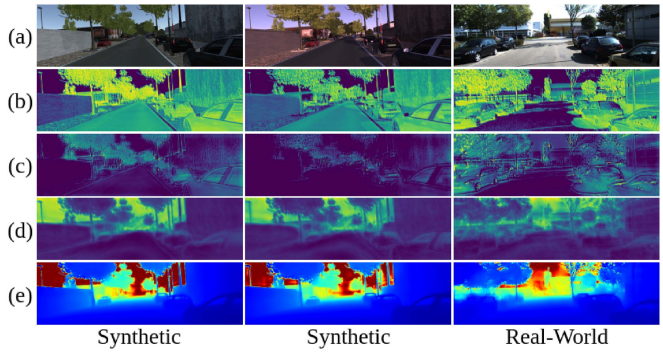


Fig. 1. Visualization of the learned intermediate features and depth maps of our approach. From top to bottom: (a) Input Image, (b) Initial Depth Feature, (c) Weak-related Depth Feature, (d) Intensified Depth Feature, and (e) Depth Map. The first and second columns represent the features learned on the synthetic dataset (vKITTI) of different image styles, and column 3 represents the features learned on the real-world dataset (KITTI). Additionally, the initial depth feature and weak-related depth feature in (b) and (c) are the single-channel features of f^{Din} (1) and f^{Dwr} (2) at maximum resolution, which are summed at the channel level for display. The intensified depth feature is the single-channel feature, which is the output of the intensified branch.

The main contributions of our paper can be summarized as:

- We propose an effective domain generalizable depth feature extraction framework (AGDF-Net), which separates the image into initial depth and weak-related depth components to efficiently extract depth related information for cross-domain generalizable depth estimation;
- An adaptive guidance fusion module is designed to sufficiently reuse and intensify the extracted initial depth features at multi-scale levels to get intensified depth features that are domain generalizable. This module can further enhance the domain invariant components for depth estimation. Finally, generalize well to unseen real domains after training only on synthetic domains;
- Without using any real-world dataset, our AGDF-Net can be well applied to the various depth estimation datasets (i.e., KITTI [23], NYUDv2 [24], NuScenes [25], DrivingStereo [26] and CityScapes [27]) and achieve state-of-the-art performance, more applicable to practical scenarios. Furthermore, the experiments using a small amount of labeled real-world data in a semi-supervised setting also demonstrate the superiority of our AGDF-Net.

The rest of our paper is organized as follows. We introduce and discuss the related work in Section II. Then the proposed AGDF-Net is introduced in Section III, including

the framework pipeline, image separation design, adaptive guidance module and loss functions. We give details of the experimental settings and show more detailed analysis in Section IV, including results of generalization experiments on various datasets (i.e., trained on vKITTI [28] and SUNCG [29], evaluated on KITTI [23], NYUDv2 [24], NuScenes [25], DrivingStereo [26] and CityScapes [27]); comparative experiments in semi-supervised settings; detailed analysis and validation of each module in the framework. Finally, we conclude the paper in Section V.

II. Related Work

With the great success of deep convolutional neural networks, DCNN-based monocular depth estimation methods have achieved exciting depth perception capabilities by studying the neural network structure or borrowing auxiliary information. In this section, we first introduce the progress made by previous monocular depth estimation methods, then introduce the related developments of domain adaptation methods, and finally, we discuss cross-domain depth estimation methods that have begun to receive attention in recent years.

A. Monocular Depth Estimation

Monocular depth estimation takes a color image as input and estimates the depth map through the encoding-decoding network [7], [30]. This task can be divided into image-only depth estimation tasks [13], [31] and depth completion tasks [32], [33], [34] with the help of other sensors, where image-only depth estimation tasks can be further divided into supervised and self-supervised methods. For supervised depth estimation methods, some approaches modify the structure of the network module to improve the depth estimation performance [8], [31], and some methods combine other tasks to help improve depth estimation performance, such as surface normal [35], [36], segmentation [37], [38] and optical flow [39], etc. For depth completion tasks, images are combined using related sensors such as LiDAR to obtain more accurate depth results. Most methods improve the completion accuracy by designing a feature fusion module [22], [40], [41]. All of the above approaches require obtaining annotated depth ground truths for training. Obtaining annotated depth ground truths requires additional sensors and algorithms to align with the color images, which is costly and time-consuming. Subsequently, self-supervised depth estimation methods appeared to solve the above problems. Self-supervised monocular depth estimation methods do not require labeled data for training, avoiding costly and time-consuming problems. Some methods are trained by estimating inter-frame poses and warping inter-frame images [13], [14], and other methods are trained by stereo matching [42], [43]. Most of the above methods are trained on a specific dataset and validated in a single domain, and all achieve superior results on specific datasets. These methods aim to achieve better depth estimation results on a single dataset, ignoring the multi-scene generalization of a single model. When generalizing to other scenes, the network estimation results will fail. In this work, the generalization of the network is paid attention to, and we aim to obtain an essential representation of depth estimation, and use this representation to learn depth, which generalizes well to real datasets from networks trained only on synthetic datasets.

B. Domain Adaptation and Generalization

Domain adaptation is a task that trains on one or more related source domains and the unlabeled target domain, and tests in the target domain [44]. This task is to bridge the gap between the source domain and the target domain by fine-tuning the network with target data to diminish the domain shift [45], [46], [47], using domain discriminators to encourage domain confusion through an adversarial objective [48], [49] and using data reconstruction as an auxiliary task to ensure feature invariance [17], [18]. [50] proposes a novel parameter-free adaptive feature norm approach for unsupervised domain adaptation by progressively adapting the feature norms of the two domains to a large range of scalars. [51], [52], [53] obtain stable action recognition results by narrowing the gaps between different domains(modalities, data forms and views), where [51] enhances action recognition in vision-sensor modality (videos) by adaptively transferring and distilling the knowledge from multiple wearable sensors, [52] enhances action recognition in videos by transferring knowledge from images using video keyframes as a bridge, and [53] addresses recognizing human actions from varied views by learning view-invariant representations hierarchically. Domain generalization is a task which only trained on the source dataset and tested on the target dataset [54], containing data augmentation and generation [55], [56], domain-invariant representation learning and feature disentanglement [57], [58] and learning strategies such as self-supervised learning [59] and gradient operation [60], etc. Among them, some methods transfer the image by decomposing the image into domain invariant and domain specific components [61], [62], [63]. In this work, we propose an efficient domain generalizable depth feature extraction framework that utilizes image separation, image reconstruction, and domain invariant representation learning to obtain domain generalizable depth features for depth estimation results with strong generalization across domains. In other words, our approach does not require any real-world data for training, but extracts domain generalizable features for depth estimation from synthetic images through feature disentanglement and reconstruction, and finally, superior depth estimation results with strong generalization are obtained in real images.

C. Cross-Domain Depth Estimation

Recently, many methods have focused on depth estimation across domains, including domain adaptation based and domain generalization based methods. For domain adaptation based approaches, in order to obtain superior results for cross-domain depth estimation, these methods are trained on labeled synthetic data and unlabeled real-world data, and tested on real-world data. [11], [16], [64] perform feature alignment between the synthetic domain and the real domain to transfer the depth estimation from the synthetic domain to the real domain. Specifically, [64] mitigates the inherent shift across domains through

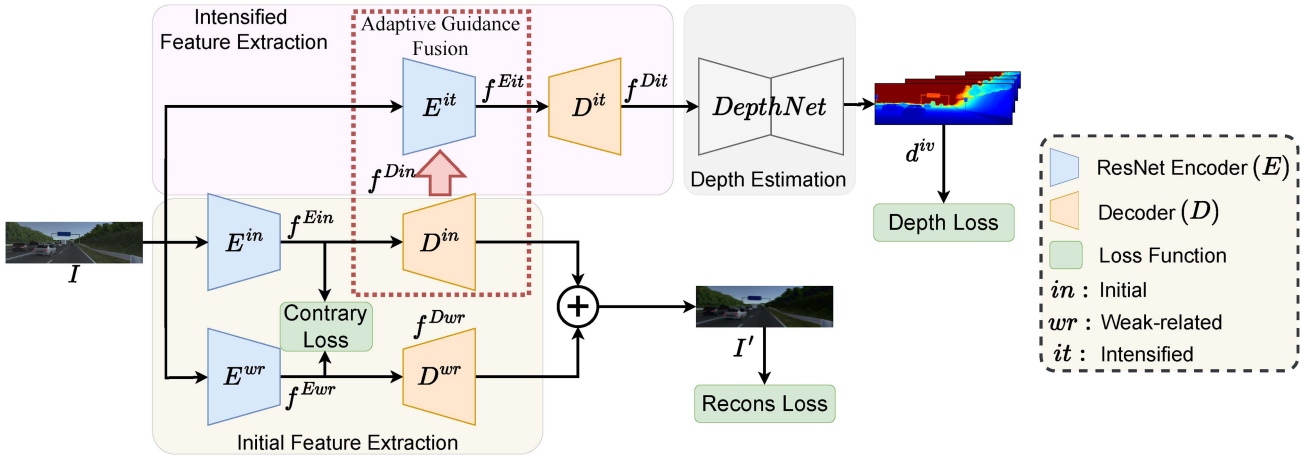


Fig. 2. Overview of the network architecture. Taking image I as input, the (E^{in}, D^{in}) and (E^{wr}, D^{wr}) are used to extract initial depth features and weak-related depth features, respectively. Then the initial depth features from D^{in} are fused in E^{it} and D^{it} to get intensified depth features (f^{Eit} and f^{Dit}) that are domain generalizable. Finally, the intensified depth features are input to *DepthNet* to get the domain generalizable depth result. The adaptive guidance fusion process is framed by a red-dotted rectangle. The loss functions in green boxes are introduced in Section III-F.

adversarial learning and explicitly imposes content consistency on the adapted target representation. [11] take advantage of style transfer and adversarial training to predict pixel perfect depth on real-world data by training on synthetic data. [16] uses an image translation network to enhance the realism of the input image, and then obtains a cross-domain invariant depth estimation result through a depth prediction network. [12] proposes a geometry-aware symmetric domain adaptation framework to explore the labels in the synthetic data and epipolar geometry in the real data jointly. The above methods mainly perform cross-domain depth estimation by aligning the synthetic domain and a single real domain at the image level or feature level. [65] makes the model have strong cross-domain generalization ability by designing consistent loss applicable to multiple datasets. This method needs to use a large number of real-world datasets and their corresponding labeled data for training, which is very costly and time-consuming. For domain generalization based approaches, these methods focus on the study of model generalization, i.e., training only on synthetic domains and achieving superior results on real domains. The goal of [20] is to learn a depth-specific feature to improve generalization, which extracts structural information using a pre-trained encoder obtained by synthesizing images and appending real-world images. And use an additional network to learn a single weight map as an attention module to attenuate useless information. Our work achieves a more general representation using an adaptive guided fusion strategy for domain generalizable feature learning for depth estimation at multi-scale levels using only synthetic data. It avoids new cross-domain problems caused by the introduction of additional datasets, and the enhancement and fusion of features at different scales can avoid the problem of incomplete information purification by a single-scale simple attention mechanism. Our approach demonstrates that transforming the representation of domain generalizable features from the single level to the multi-scale feature levels can lead to more vital generalization ability.

III. Approach

Remarkable progress has been achieved in monocular depth estimation [8], [13], where the training and testing processes of most methods are in the same domain. However, the performance is heavily limited when training and testing are in different domains. To relieve the problem, we propose a generalizable depth estimation framework by learning domain generalizable depth features with adaptive guidance fusion, i.e., AGDF-Net, and we provide more details of the proposed network in this section.

A. Overview

[20], [21] have proven that structural information is more related to depth while style and illumination, etc., are disturbance terms for depth perception. Therefore, our approach aims to extract domain invariant representations for cross-domain generalizable depth estimation. As shown in the light yellow area in Fig. 2, to extract features that are related to depth estimation preliminarily, taking a color image I as input, our AGDF-Net first utilizes the initial branch (initial depth encoder and decoder: E^{in} and D^{in}) and weak-related branch (weak-related depth encoder and decoder: E^{wr} and D^{wr}) to extract initial depth features from the color image. The purpose is to extract initial depth features that are related to depth for depth estimation while the interferential information can be preliminarily removed. In this process, contrary loss and reconstruction loss are employed to separate and extract the two kinds of initial depth features and weak-related depth features. Precisely, the initial branch and the weak-related branch extract two kinds of features from the same image, and these two kinds of features are constrained by contrary loss to obtain features that are mutually exclusive. In order to avoid losing information, the reconstruction loss is used to reconstruct the two features back to the original image, so as to ensure that the information extracted by the two branches is different but complementary.

To further intensify the initial depth features to get domain generalizable depth features, in the lilac region in Fig. 2, the intensified branch (intensified depth encoder and decoder: E^{it} and D^{it}) reuses color images to further guide the initial depth features for feature enhancement, resulting in more effective domain generalizable features for depth estimation. Specifically, after obtaining two completely different features, one of the features (the feature extracted by the initial branch) is sent to the adaptive guidance fusion module to intensify the multi-scale initial depth features to obtain multi-scale intensified depth features (domain generalizable features for depth estimation) for subsequent cross-domain depth estimation.

Finally, in the light gray area of Fig. 2, arbitrary depth estimation networks can be applied to obtain domain generalizable depth estimation results. In this process, depth loss is imposed during training to constrain the predicted depth results.

Better depth results are obtained, domain generalizable features for cross-domain depth estimation can be obtained by the initial branch and adaptive guidance fusion module, and the remaining information can be separated into the weak-related branch (with contrary loss and reconstruction loss). Note that the initial and weak-related branches are for preliminary feature disentangling. The resulting initial depth features are not entirely domain invariant for depth estimation. The initial depth features can be further enhanced using the proposed adaptive guidance module, where the small texture information is further eliminated and adequate information is further strengthened. Finally, the fully decoupled domain generalizable intensified depth features are obtained for subsequent depth estimation.

B. Initial Feature Extraction

This process aims to extract common information from images of different domains. Inspired by [17], two encoding-decoding branches are used to separate images of different domains into initial depth features and weak-related depth features. These two branches are defined as initial branch (E^{in} and D^{in}) and weak-related branch (E^{wr} and D^{wr}) as shown in the light yellow area in Fig. 2.

Initial Branch: The initial branch consists of an initial depth encoder E^{in} and an initial depth decoder D^{in}. E^{in} aims to extract initial depth encoding features f^{Ein} from images of different domains. Then f^{Ein} are fed into D^{in} to extract initial depth decoding features f^{Din}, and reconstruct the initial depth map $I^{'in}$ for subsequent image reconstruction. Note that commonly used ResNet18 [66] encoder is utilized here, and the process can be formulated as:

$$\begin{aligned} f^{Ein} &= E^{in}(I) \\ f^{Din}, I^{'in} &= D^{in}(f^{Ein}) \end{aligned} \tag{1}$$

where $f^{Ein} = \{f_1^{Ein}, \ldots, f_n^{Ein}\}$, $f^{Din} = \{f_1^{Din}, \ldots, f_n^{Din}\}$, n is the number of features extracted at each of the different encoding/decoding scales. And the shapes of these features are $B \times C \times H \times W$, where B is the batch size, C is the number of channels of features, H and W are the height and width of the features, C, H, W are variable. Similar features in the following sections have the same shapes and will not be described again.

Weak-Related Branch: The weak-related branch consists of weak-related depth encoder E^{wr} and weak-related depth decoder D^{wr}. E^{wr} aims to extract weak-related depth encoding features f^{Ewr} that are disturbances for depth estimation from images of different domains. Then f^{Ewr} are fed into D^{wr} to extract weak-related depth decoding features f^{Dwr}, and reconstruct the weak-related depth map $I^{'wr}$ for subsequent image reconstruction. The process can be formulated as:

$$\begin{aligned} f^{Ewr} &= E^{wr}(I) \\ f^{Dwr}, I^{'wr} &= D^{wr}(f^{Ewr}) \end{aligned} \tag{2}$$

where $f^{Ewr} = \{f_1^{Ewr}, \ldots, f_n^{Ewr}\}$, $f^{Dwr} = \{f_1^{Dwr}, \ldots, f_n^{Dwr}\}$, n is the number of features extracted at each of the different encoding/decoding scales.

Initial Depth Feature Extraction: The purpose of the initial branch and weak-related branch is to separate the image into initial depth and weak-related depth parts. According to the previous work [20], [21], when the neural network observes depth, structure and texture play a significant role, which is domain invariant for depth estimation. This is also in line with the conclusion that humans observe 3D geometry (depth) through sketches and structures [67], [68]. Therefore, inspired by image decomposition [69], the initial depth information (related to depth perception) and weak-related depth information (the remaining interference information) separated from the same image should be independent and complementary. In this process, the initial depth encoding features f^{Ein} and weak-related depth encoding features f^{Ewr} are constrained by contrary loss, so that the two branches can extract different information independently of each other, where the initial branch is used for domain generalizable depth estimation. The reconstructed images from the initial depth and weak-related depth features ($I^{'in}$ and $I^{'wr}$) extracted by the two branches can be jointly reconstructed to the input image, which is formulated as:

$$I' = I^{'in} + I^{'wr} \tag{3}$$

where I' is the reconstructed image, which should be the same as the input image I. The reconstructed image I' and the input image I are constrained by the reconstruction loss. See Section III-F for the details of the contrary loss (8) and reconstruction loss (9). Note that only the initial branch is involved in inference. Visualization and quantitative results are provided in Section IV to clarify the extracted features and prove the necessity of our architecture. Please see Figs. 7, and 8, Table VIII and Section IV-B for details.

C. Intensified Feature Extraction

Our goal is to learn domain generalizable depth features for cross-domain generalizable depth estimation. To further remove the influence of interference information for depth estimation, such as some detailed textures, in this process, we design the intensified depth branch to further optimize the extracted initial depth features, removing the residual disturbing information and

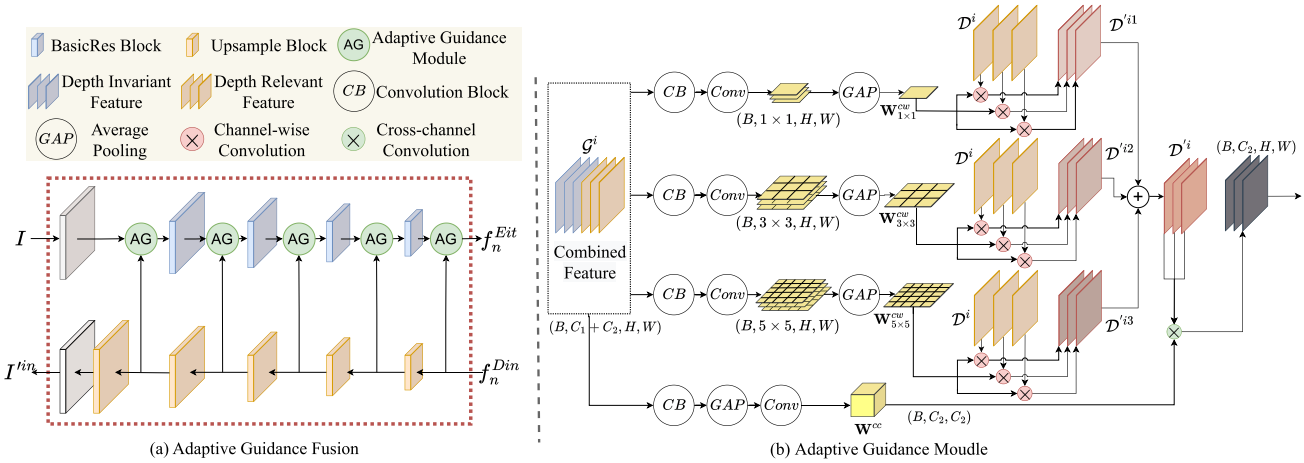


Fig. 3. Details of adaptive guidance fusion. Each subsection in this figure represents: (a) Adaptive guidance fusion, (b) Adaptive guidance module (AG in (a)).

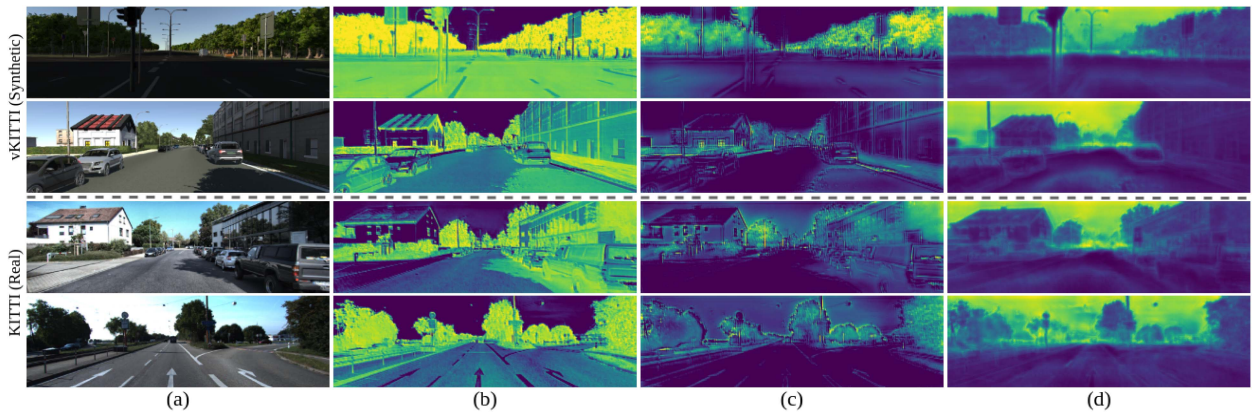


Fig. 4. Visualization of the learned initial depth, weak-related depth and intensified depth features in outdoor scenes. From left to right: (a) Input Image, (b) Initial Depth Feature, (c) Weak-related Depth Feature, (d) Intensified Depth Feature. The top two rows are features of synthetic images on vKITTI, the bottom two rows are features of real-world images on KITTI. Note that the brighter the color, the higher the value. Additionally, the initial depth feature and weak-related depth feature in (b) and (c) are the single-channel features of f^{Din} (1) and f^{Dwr} (2) at maximum resolution, which are summed at the channel level for display. The intensified depth feature is the single-channel feature, which is the output of the intensified branch.

TABLE I
QUANTITATIVE COMPARISON WITH STATE-OF-THE-ART METHODS ON KITTI

Method	dataset	SUP	EXT	MAX	Abs Rel	Sq Rel	RMSE	RMSE_{log}	δ <1.25	δ <1.25^2	δ <1.25^3
Eigen *et al.* [7]	K	Yes	-	80m	0.215	1.515	7.156	0.270	0.692	0.899	0.967
MiDaS [65]	MIX 5	Yes	-	80m	0.156	1.281	5.990	0.226	0.784	0.936	0.976
DPT-Hybrid [74]	MIX 6	Yes	K	80m	0.062	0.222	2.575	0.092	0.959	0.995	0.999
Kundu *et al.* [64]	V	UDA	K	80m	0.214	1.932	7.157	0.295	0.665	0.882	0.950
T^2Net [16]	V	UDA	K	80m	0.171	1.351	5.944	0.247	0.757	0.918	0.969
S2R-DepthNet [20]	V	DG	PBN	80m	0.165	1.351	5.695	0.236	0.781	0.931	0.972
DepthNet Baseline	V	DG	-	80m	0.230	2.169	6.865	0.300	0.665	0.879	0.952
Ours	V	DG	-	80m	**0.155**	**1.049**	**5.388**	**0.226**	**0.782**	**0.936**	**0.977**
Kundu *et al.* [64]	V	UDA	K	50m	0.203	1.734	6.251	0.284	0.687	0.899	0.958
T^2Net [16]	V	UDA	K	50m	0.164	1.019	4.469	0.231	0.773	0.928	0.974
S2R-DepthNet [20]	V	DG	PBN	50m	0.158	1.000	4.321	0.223	0.793	0.939	0.976
DepthNet Baseline	V	DG	-	50m	0.224	1.808	5.483	0.288	0.679	0.889	0.957
Ours	V	DG	-	50m	**0.149**	**0.822**	**4.113**	**0.214**	**0.796**	**0.944**	**0.980**

K, V and PBN denote KITTI, vKITTI and Painter By Numbers datasets, respectively. Mix 5 means training on mix five datasets in [65], MIX 6 means training on mix six datasets containing about 1.4 million images in [74]. SUP means supervision, and for this column, Yes is supervised learning, UDA is unsupervised domain adaptation, DG is domain generalization. EXT means extra images for training, MAX means max depth. DepthNet Baseline is the baseline result of ours with only depth estimation network. DPT-Hybrid [74] is trained on MIX 6 and fine-tuned on the KITTI dataset. The best results are marked in bold.

TABLE II
COMPARISON ON KITTI FOR SEMI-SUPERVISED SETTING

Method	dataset	MAX	Abs Rel	Sq Rel	RMSE	$RMSE_{log}$	δ <1.25	δ <1.25^2	δ <1.25^3
Kundu *et al.* [64]	V+K(Small)	80m	0.167	1.257	5.578	0.237	0.771	0.922	0.971
Zhao *et al.* [75]	V+K(Small)	80m	0.143	0.927	4.694	0.252	0.796	0.922	0.968
S2R-DepthNet-S [20]	V+K(Small)	80m	0.116	0.766	4.409	0.185	0.858	0.955	0.984
Ours-DG	V	80m	0.155	1.049	5.388	0.226	0.782	0.936	0.977
Ours-S	V+K(Small)	80m	**0.114**	**0.704**	**4.305**	**0.180**	**0.870**	**0.961**	**0.985**
kuznietsov *et al.* [76]	K+Stereo	80m	0.113	0.741	4.621	0.189	0.862	0.960	0.986
Kundu *et al.* [64]	V+K(Small)	50m	0.162	1.041	4.344	0.225	0.784	0.930	0.974
S2R-DepthNet [20]	V+K(Small)	50m	0.111	0.642	3.463	0.176	0.870	0.959	0.986
Ours-DG	V	50m	0.149	0.822	4.113	0.214	0.796	0.944	0.980
Ours-S	V+K(Small)	50m	**0.110**	**0.573**	**3.357**	**0.170**	**0.881**	**0.966**	**0.987**
kuznietsov *et al.* [76]	K+Stereo	50m	0.108	0.595	3.518	0.179	0.875	0.964	0.988

V means vKITTI, K(Small) means the first 1000 KITTI frames. The best results are marked in bold.

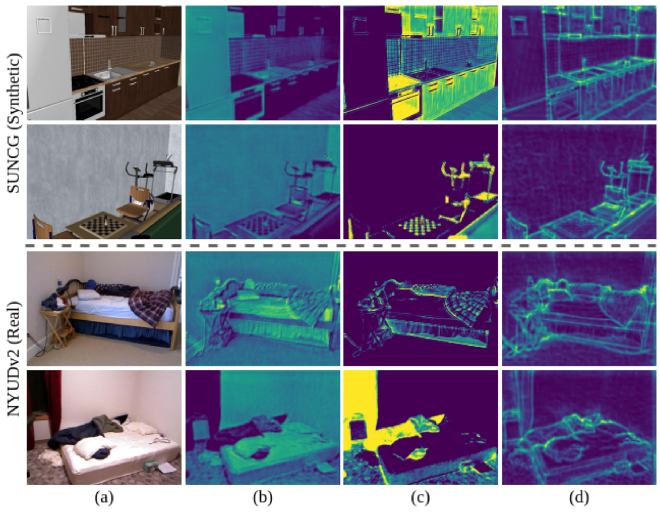


Fig. 5. Visualization of the learned initial depth, weak-related depth and intensified depth features in indoor scenes. From left to right: (a) Input Image, (b) Initial Depth Feature, (c) Weak-related Depth Feature, (d) Intensified Depth Feature. The top two rows are features of synthetic images on SUNCG, the bottom two rows are features of the real-world images on NYUDv2. Note that the brighter the color, the higher the value. Additionally, the initial depth feature and weak-related depth feature in (b) and (c) are the single-channel features of f^{Din} (1) and f^{Dwr} (2) at maximum resolution, which are summed at the channel level for display. The intensified depth feature is the single-channel feature, which is the output of the intensified branch.

TABLE III
QUANTITATIVE COMPARISON ON NYUDV2

Method	Abs	Rel	RMSE	log10	δ <1.25	δ <1.25^2	δ <1.25^3
Li *et al.* [58]	0.232	0.821	0.094	0.621	0.886	0.968	
Eigen *et al.* [7]	0.215	0.907	-	0.611	0.887	0.971	
MiDaS [65]	0.125	0.526	-	0.861	0.970	0.992	
DPT-Hybrid [74]	0.109	0.357	0.045	0.904	0.988	0.998	
S2R-DepthNet [20]	0.205	0.690	0.085	0.687	0.899	0.967	
Ours	**0.199**	**0.685**	**0.084**	**0.693**	**0.904**	**0.970**	

MiDaS [65] is trained on MIX 5 dataset, where MIX 5 means mix five datasets in [65]. DPT-Hybrid [74] is trained on MIX 6 and fine-tuned on the NYUDv2 dataset, where MIX 6 means mixing six datasets containing about 1.4 million images in [74].

enhancing the domain invariant information for depth estimation. Images contain a large amount of texture information and have certain corresponding relationships with the depth map in structure and texture. Therefore, in the intensified feature extraction process, image information is reused to learn adaptive convolution parameters to guide the optimization of domain invariant information.

Intensified Branch: As shown in the lilac area in Fig. 2, the intensified branch consists of intensified depth encoder E^{it} and intensified depth decoder D^{it}. E^{it} aims to intensify the initial depth decoding features f^{Din} with adaptive guidance fusion (Section III-D), getting intensified depth encoding features f^{Eit}. Then f^{Eit} are fed into D^{it} to extract the multi-scale intensified depth maps f^{Dit} that are domain generalizable for the subsequent domain generalizable depth estimation.

D. Adaptive Guidance Fusion

Adaptive Guidance Fusion: The details of the adaptive guidance fusion are shown in Fig. 3(a). At each scale of the neural network, the image information is combined to guide the optimization of initial depth information with the adaptive guidance module (AG) to obtain multi-scale intensified depth features. The whole process can be formulated as:

$$\begin{aligned} f_0^{Eit} &= E_1^{it}(I) \\ f_1^{Eit} &= E_2^{it}(AG_1(f_0^{Eit}, f_1^{Din})) \\ f_2^{Eit} &= E_3^{it}(AG_2(f_1^{Eit}, f_2^{Din})) \\ &\cdots \\ f_{n-1}^{Eit} &= E_n^{it}(AG_{n-1}(f_{n-2}^{Eit}, f_{n-1}^{Din})) \\ f_n^{Eit} &= (AG_n(f_{n-1}^{Eit}, f_n^{Din})) \\ f^{Dit} &= D^{it}(f^{Eit}) \end{aligned} \tag{4}$$

where $f^{Eit} = \{f_0^{Eit}, f_1^{Eit}, \ldots, f_n^{Eit}\}$, f_0^{Eit} is the extracted image feature, $f^{Dit} = \{f_1^{Dit}, \ldots, f_n^{Dit}\}$, $E^{it} = \{E_1^{it}, \ldots, E_n^{it}\}$, n is the number of features extracted at each of the different encoding/decoding scales. AG_i means the adaptive guidance operation on the i-th feature scale. For f_i^{Eit} and f_i^{Din}, feature scales are from large to small as i increases.

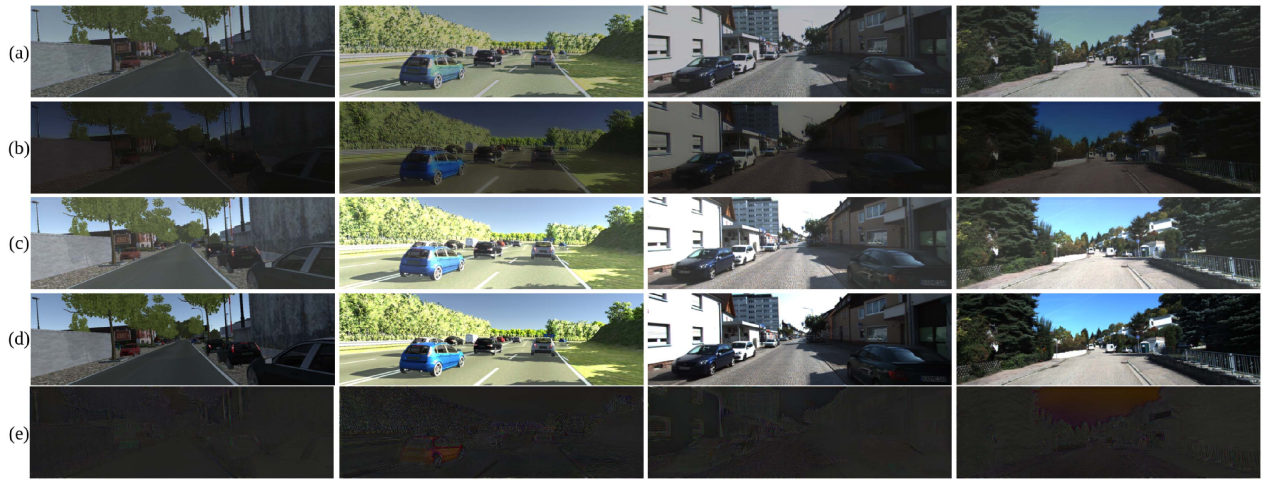


Fig. 6. Visualization of reconstruction images and errors. From top to bottom: (a) Reconstructed Image $I^{'in}$ from Initial Depth Features, (b) Reconstructed Image $I^{'wr}$ from Weak-related Depth Features, (c) Final Reconstructed Image I', (d) Input Image I, (e) Reconstruction Error Image. The first two columns are images on vKITTI, and the last two columns are images on KITTI. The reconstructed image is almost the same as the input image. The darker the reconstruction error color, the smaller the error, and the overall reconstruction error is kept at a low level.

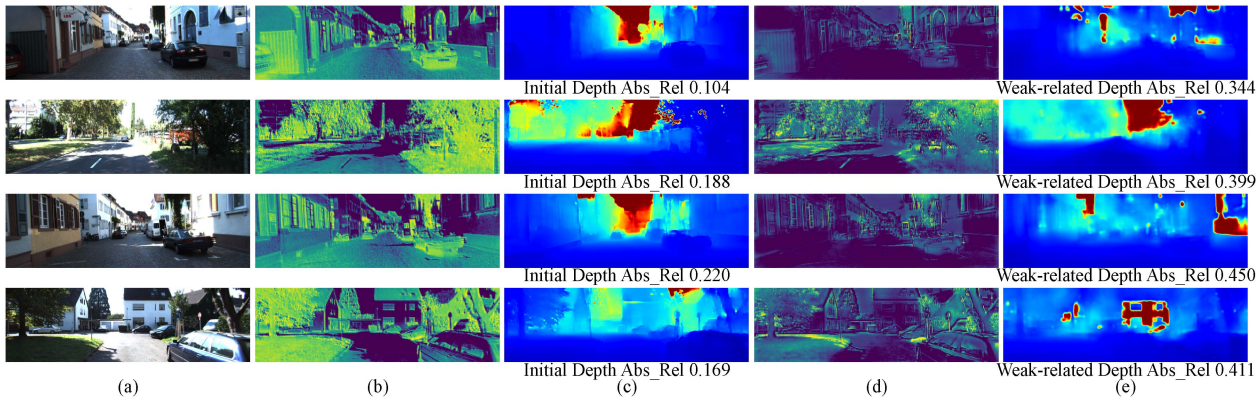


Fig. 7. Depth prediction using initial depth and weak-related depth features on the KITTI dataset. From left to right: (a) Input Image, (b) Initial Depth Feature, (c) Initial Depth Prediction Map (d^{in}), (d) Weak-related Depth Feature, and (e) Weak-related Depth Prediction Map (d^{wr}). d^{in} is clearer in object boundaries, while the result of d^{wr} has blurred object boundaries and even predicts wrongly. For a quantitative comparison, we compare d^{in} and d^{wr} on the Abs_Rel, where the weak-related depth maps are much worse than the initial depth maps.

TABLE IV
GENERALIZATION ON MORE DATASETS

Method	Dataset	Abs Rel	Sq Rel	RMSE	$RMSE_{log}$	δ <1.25	δ <1.25^2	δ <1.25^3
T^2Net [16]	NuScenes	0.231	2.744	10.099	0.371	0.626	0.845	0.914
S2R-DepthNet [20]	NuScenes	0.201	2.304	8.960	0.308	0.703	0.880	0.936
Ours	NuScenes	**0.180**	**2.010**	**8.910**	**0.297**	**0.739**	**0.883**	**0.941**
T^2Net [16]	DrivingStereo	0.389	14.157	19.565	0.426	0.512	0.741	0.860
S2R-DepthNet [20]	DrivingStereo	0.206	2.988	9.954	0.256	0.690	0.909	0.972
Ours	DrivingStereo	**0.183**	**2.433**	**9.138**	**0.226**	**0.745**	**0.935**	**0.983**
T^2Net [16]	CityScapes	0.261	3.887	12.282	0.369	0.579	0.804	0.910
S2R-DepthNet [20]	CityScapes	0.173	2.225	9.254	0.246	0.747	0.919	0.972
Ours	CityScapes	**0.155**	**1.777**	**9.025**	**0.240**	**0.759**	**0.920**	**0.973**

All datasets are evaluated within 80m, which is consistent with KITTI. The best results are marked in bold.

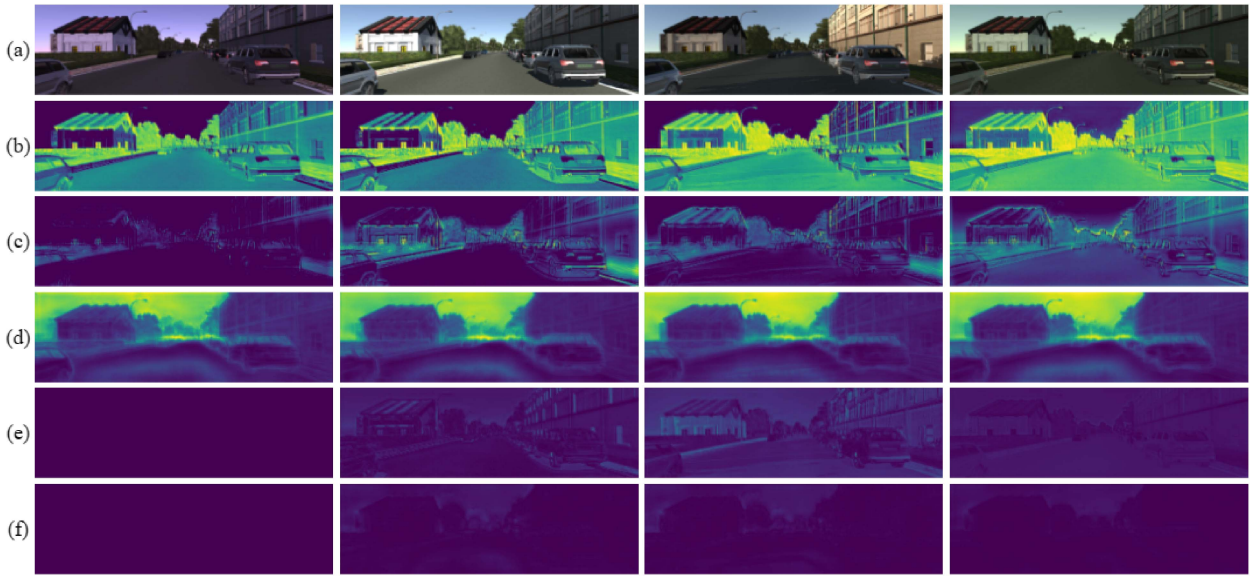


Fig. 8. Different styles of images from the same scene with similar initial depth and nearly the same intensified depth features on vKITTI. From top to bottom: (a) Image, (b) Initial Depth Feature, (c) Weak-related Depth Feature, (d) Intensified Depth Feature, (e) Differential Map of Initial Depth Feature relative to the first column, and (f) Differential Map of Intensified Depth Feature relative to the first column. For images of different styles, our approach can generate similar initial depth features and almost the same intensified depth features, while weak-related depth features have different strengths for different styles, and depth detailed information is lost.

TABLE V

COMPARISON WITH STATE-OF-THE-ART METHODS WITH DIFFERENT DEPTH ESTIMATION NETWORKS. DEPTHNET MEANS DEPTH ESTIMATION NETWORK USED IN THE PREVIOUS WORK [16], [20] AND RB-NET MEANS COMMONLY USED RESNET BASED DEPTH ESTIMATION NETWORK [13]. RB-NET (PRE-TRAINED) MEANS RB-NET PRE-TRAINED ON IMAGENET

Method	Depth Estimation Network	Abs Rel	Sq Rel	RMSE	RMSE_{log}	δ <1.25	δ <1.25^2	δ <1.25^3
S2R-DepthNet [20]	DepthNet [20]	0.165	1.351	5.695	0.236	0.781	0.931	0.972
Ours	DepthNet [20]	**0.155**	**1.049**	**5.388**	**0.226**	**0.782**	**0.936**	**0.977**
S2R-DepthNet [20]	RB-Net [13]	0.185	1.449	5.842	0.250	0.738	0.919	0.971
Ours	RB-Net [13]	0.168	1.404	5.760	**0.235**	0.776	**0.932**	**0.974**
Ours	RB-Net (pre-trained) [13]	**0.164**	**1.340**	**5.749**	0.236	**0.777**	0.931	0.973

DepthNet means depth estimation network used in the previous work [16], [20] and RB-Net means commonly used ResNet based depth estimation network [13]. RB-Net (pre-trained) means RB-Net pre-trained on ImageNet.

TABLE VI

ABLATION STUDY RESULTS TO ILLUSTRATE THE EFFECTIVENESS OF EACH MODULE. INE MEANS INITIAL FEATURE EXTRACTION, WHICH IS THE ENTIRE PROCESS OF INITIAL DEPTH FEATURE AND WEAK-RELATED DEPTH FEATURE EXTRACTION, ITE MEANS INTENSIFIED FEATURE EXTRACTION. INB MEANS INITIAL BRANCH, WHICH IS A SINGLE BRANCH

Method	Abs Rel	Sq Rel	RMSE	RMSE_{log}	δ <1.25	δ <1.25^2	δ <1.25^3
Baseline	0.230	2.169	6.865	0.300	0.665	0.879	0.952
+INE	0.173	1.372	6.229	0.254	0.750	0.916	0.967
+INE+ITE(Add)	0.164	1.457	6.282	0.238	0.776	0.925	0.971
+INE+ITE(Concatenate)	0.163	1.334	6.262	0.245	0.762	0.923	0.972
+INB+ITE(Adaptive Guidance)	0.171	1.224	6.248	0.255	0.744	0.912	0.967
+INE+ITE(Adaptive Guidance)	0.155	1.049	5.388	0.226	0.782	0.936	0.977

INE means Initial Feature Extraction, which is the entire process of initial depth feature and weak-related depth feature extraction, ITE means Intensified Feature Extraction. INB means Initial Branch, which is a single branch.

TABLE VII

ANALYSIS OF INITIAL FEATURE EXTRACTION AND LOSS FUNCTIONS. INB MEANS INITIAL BRANCH, WRB MEANS WEAK-RELATED BRANCH. THE INITIAL EXTRACTION PROCESS (INE) CONSISTS OF INB AND WRB. $\mathcal{L}_{rec}$ MEANS RECONSTRUCTION LOSS, $\mathcal{L}_{con}$ MEANS CONTRARY LOSS

Method	$\mathcal{L}_{rec}$	$\mathcal{L}_{con}$	Abs Rel	Sq Rel	RMSE	RMSE_{log}	δ <1.25	δ <1.25^2	δ <1.25^3
INB			0.184	1.531	6.125	0.249	0.738	0.918	0.971
INB+WRB	✓		0.179	1.608	6.322	0.249	0.756	0.920	0.970
INB+WRB	✓	✓	0.173	1.372	6.229	0.254	0.750	0.916	0.967

INB means Initial Branch, WRB means Weak-related Branch. The initial extraction process (INE) consists of INB and WRB. $\mathcal{L}_{rec}$ means reconstruction loss, $\mathcal{L}_{con}$ means contrary loss.

TABLE VIII
ANALYSIS OF CASCADE OF HOURGLASS ARCHITECTURES BEFORE BASELINE DEPTH ESTIMATION NETWORK. THE HOURGLASS ARCHITECTURES ARE THE SAME AS THE INITIAL BRANCH

Method	Abs Rel	Sq Rel	RMSE	RMSE_{log}	δ <1.25	δ <1.25^2	δ <1.25^3
Baseline	0.230	2.169	6.865	0.300	0.665	0.879	0.952
+ one hourglass architecture	0.184	1.531	6.125	0.249	0.738	0.918	0.971
+ two hourglass architectures	0.181	1.354	6.275	0.260	0.729	0.912	0.966
+ three hourglass architectures	0.184	1.427	6.043	0.247	0.735	0.920	0.974
+ four hourglass architectures	0.181	1.443	6.038	0.251	0.740	0.919	0.970

The hourglass architectures are the same as the initial branch.

Adaptive Guidance Module (AG): Convolution is applied as an adaptive form of feature enhancement. AG aims to learn convolution kernel parameters combined with image information, and optimize initial depth decoding features (f^{Din}) with convolution to obtain intensified depth features (f^{it}). Fig. 3(b) demonstrates the details of the AG. Inspired by the separable convolution [22], [70], convolution operations can be separated into channel-wise convolution and cross-channel convolution. AG optimizes initial depth features by adaptively learning channel-wise convolution kernels and cross-channel convolution kernels. Specifically, the combined feature groups input into AG are defined as $\mathcal{G}$, where $\mathcal{G}^i = \{f_{i-1}^{Eit}, f_i^{Din}\}, i \in [1, n]$. The initial depth feature groups that need to be optimized are defined as $\mathcal{D}$, where $\mathcal{D}^i$ represents $f_i^{Din}, i \in [1, n]$.

More specifically, in the channel-wise convolution process, given an initial depth feature group $\mathcal{D}^i$ with m channels, AG uses the combined feature group $\mathcal{G}^i$ to learn single-channel convolution kernels of different sizes, and performs convolution optimization of $\mathcal{D}^i$ for each channel. Here, 1×1, 3×3 and 5×5 kernels adaptively learned from $\mathcal{G}^i$ (as shown in Fig. 3(b)) are used respectively, denoted as $\mathbf{W}_{1\times1}^{cw}$, $\mathbf{W}_{3\times3}^{cw}$ and $\mathbf{W}_{5\times5}^{cw}$. The optimized feature $\mathcal{D}_j^{'i}$ of the j-th channel obtained after channel-wise convolution can be expressed as:

$$\begin{aligned}\mathcal{D}_j^{'i1} &= \mathbf{W}_{1\times1}^{cw} \otimes \mathcal{D}_j^i \\ \mathcal{D}_j^{'i2} &= \mathbf{W}_{3\times3}^{cw} \otimes \mathcal{D}_j^i \\ \mathcal{D}_j^{'i3} &= \mathbf{W}_{5\times5}^{cw} \otimes \mathcal{D}_j^i \\ \mathcal{D}_j^{'i} &= \mathcal{D}_j^{'i1} + \mathcal{D}_j^{'i2} + \mathcal{D}_j^{'i3}\end{aligned} \tag{5}$$

where $j \in [1, m]$, $\mathcal{D}^i = \{D_1^i, \ldots, D_j^i, \ldots, D_m^i\}$, and $\mathcal{D}^{'i} = \{D_1^{'i}, \ldots, D_j^{'i}, \ldots, D_m^{'i}\}$. $\otimes$ indicates the convolution operation.

In the cross-channel convolution process, since the kernel of this process is only related to the number of channels, the previously optimized initial depth features are added first to obtain $\mathcal{D}^{'i}$ as shown in (5), and then the cross-channel convolution operation is used for further optimization, which can be formulated as:

$$\hat{\mathcal{D}}^i = \mathbf{W}^{cc} \otimes \mathcal{D}^{'i} \tag{6}$$

where $\hat{\mathcal{D}}^i$ is the final optimized feature at the i-th scale. $\mathbf{W}^{cc}$ is the cross-channel kernel adaptively learned from $\mathcal{G}^i$ as shown in Fig. 3(b), and the size is $m \times m$.

E. Domain Generalizable Depth Estimation

After obtaining intensified depth maps f^{Dit} that are domain generalizable (introduced in Sections III-C and III-D), an arbitrary depth estimation network can be used for subsequent domain generalizable depth estimation, getting cross-domain generalizable depth results d^{it} as follows:

$$d^{it} = DepthNet(f^{Dit}) \tag{7}$$

We will give more analysis in the experiments.

F. Loss Functions

1) Contrary Loss: In order to separate initial depth and weak-related depth features from the same image, an opposite constraint is imposed on initial depth encoding feature f_n^{Ein} and weak-related depth encoding feature f_n^{Ewr} to extract independent and complementary information from two branches (initial and weak-related), which is defined as contrary loss. Specifically, in order to ensure that initial depth and weak-related depth features are as independent as possible, the contrary loss constrains these two features to be orthogonal in the vector space, which can be expressed as:

$$\begin{aligned}\upsilon^{in} &= \frac{\Theta(f_n^{Ein})}{||\Theta(f_n^{Ein})||_2 + \gamma} \\ \upsilon^{wr} &= \frac{\Theta^{\mathrm{T}}(f_n^{Ewr})}{||\Theta^{\mathrm{T}}(f_n^{Ewr})||_2 + \gamma} \\ \mathcal{L}_{con} &= \upsilon^{in} \cdot \upsilon^{wr} + \lambda_1 + \lambda_2\end{aligned} \tag{8}$$

where Θ is a layer of convolution operation and straightens the feature into a one-dimensional vector. $||\cdot||_2$ is 2-norm operation, γ is set to $1e-6$ to avoid a denominator of zero. λ_1 and λ_2 are regularization terms. The purpose is to avoid the feature terms being zero, where $\lambda_1 = abs(||\upsilon^{in}||_2 - 1)$, $\lambda_2 = abs(||\upsilon^{wr}||_2 - 1)$, $abs(\cdot)$ is absolute operation.

2) Reconstruction Loss: As shown in Fig. 2, in order for the initial depth branch and the weak-related depth branch to learn complementary information, reconstruction constraints are imposed on the reconstructed image I' (3) and the original input image I, named reconstruction loss, which is defined as:

$$\mathcal{L}_{rec} = ||I' - I||_2 + ||I' - I||^2 + \sigma_1 + \sigma_2 \tag{9}$$

where $||\cdot||$ is 1-norm operation. σ_1 and σ_2 are regularization terms, where $\sigma_1 = -||I^{'in}||$, $\sigma_2 = -||I^{'wr}||$. Since (8) avoids the feature items of the encoder from converging to zero, $I^{'in}$ and $I^{'wr}$ integrate the features of the encoder, so under the constraints

of all losses, it can also prevent the final reconstructed output of the decoder from being zero. Various results demonstrated in Fig. 6(a) and (b) further prove that the output $I^{'in}$ and $I^{'wr}$ do not converge to zero.

3) Depth Loss: The predicted multi-scale cross-domain generalizable depth maps d^{it} and the corresponding depth ground truth maps d^{gt} are constrained by depth loss, which is formulated as:

$$\mathcal{L}_d = w_1 \sum_{i}^{n} ||d_i^{gt} - d_i^{it}|| + w_2 \sum_{i}^{n} \text{SSIM}(d_i^{gt}, d_i^{it}) \quad (10)$$

where w_1 and w_2 are the weighted parameters, which are empirically set to 1 and 5. SSIM means structural similarity loss [71].

The total loss is defined as:

$$\mathcal{L}_{total} = a_1 \cdot \mathcal{L}_{con} + a_2 \cdot \mathcal{L}_{rec} + a_3 \cdot \mathcal{L}_d \quad (11)$$

where a_1 to a_3 are the weighted parameters, which are empirically set as 0.1, 1.0 and 1.0.

All the above three losses are used to separate the image into initial depth and weak-related depth features. The contrary loss is used to separate the two features of the image as far as possible, and the reconstruction loss is used to constrain the separated features to be different but complementary, and can be combined back to the original image without losing information. Depth loss guarantees that final depth estimation results can be obtained, better depth estimation results are with domain generalizable features, and the remaining information can be separated into the weak-related branch (with contrary loss and reconstruction loss). Qualitative results are provided in Figs. 7 and 8 to demonstrate the separated features. Please see Section IV for details.

G. Implementation Details

For the initial branch, weak-related branch and intensified branch, all encoders used in our framework are based on ResNet18 [66]. The encoder-decoder backbones are the same as [13], and there are skip connections between the encoder and decoder. For DepthNet, two network structures are applied to demonstrate the generalizability of our AGDF-Net, including the depth estimation network architecture of previous works [12], [16], [20] and the widely used ResNet18 [66] depth estimation network [13].

The whole network is trained in an end-to-end manner for 20 epochs. The training procedure starts with an initial learning rate of 1e-4 and reduces by 50% every 5 epochs. We use a step learning rate decay with Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$), and the batch size is set as 24.

IV. Experiments

In this section, we first introduce the details of our experimental setup. Then, we verify the effectiveness of our approach in the synthetic to real generalization of depth estimation both in outdoor and indoor scenes. A comparison of a semi-supervised setting is also provided to further prove the capability of the network. Then, generalization results on more datasets further prove the generalizability of our approach. Finally, the ablation study is provided to analyze the effectiveness of each part of our architecture.

A. Experimental Setup

1) Outdoor Datasets: Virtual KITTI (vKITTI) [28]: vKITTI is a photo-realistic synthetic video dataset designed to learn and evaluate computer vision models for several video understanding tasks, which contains 21260 frames generated from five different virtual worlds in urban settings under different imaging and weather conditions. In this paper, this dataset is used as the outdoor scene source domain dataset for training. Following previous works [12], [16], [20], 20760 image-depth pairs are randomly selected as our training dataset. The image resolution is downsampled from 375×1242 to 192×640. Following prior works [16], [20], the range of the depth ground truth is clipped to 80 m.

KITTI [23]: KITTI is a large real outdoor autonomous driving dataset, which includes color images and depth collected from Velodyne HDL64. In this paper, this dataset is used as the real-world outdoor dataset for evaluation. Following [16], [20], 697 test frames are used for evaluation, and the frames are downsampled to 192×640.

2) Indoor Datasets: SUNCG [29]: SUNCG is a large-scale indoor dataset of synthetic 3D scenes, which includes 45622 3D houses with various room types. In this paper, this dataset is used as the indoor source domain dataset for training. Following previous works [16], [20], 130 k image-depth pairs are chosen for training. Since the input resolution of our network needs to be a multiple of 16, similar to previous works [7], the original image of resolution 480×640 pixels is downsampled and cropped to 224×304 pixels as input.

NYUDv2 [24]: NYUDv2 is a real-world indoor dataset, which contains a large set of video frames captured from Microsoft Kinect, with 1449 test frames. In this paper, this dataset is used for evaluation. The evaluation split is selected from 1449 test frames following [7], which contains 654 frames. The original image of resolution 480×640 is downsampled and cropped to 224×304 as input.

3) Evaluation Metrics: All results of our approach are evaluated with standard evaluation metrics described in [7], [64], [72], including RMSE, Abs Rel, RMSE$_{log}$, Sq Rel, log10 and Threshold δ. Let d_i and $\hat{d}_i$ denote the ground truth depth and estimated depth at the pixel location $i, i \in [1, N]$, N is the number of valid pixels in the ground truth depth. The evaluation metrics are specified as follows:

- RMSE: $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{d}_i - d_i)^2}$
- Abs Rel: $\frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{d}_i - d_i|}{d_i}$
- RMSE$_{log}$: $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log(\hat{d}_i) - \log(d_i))^2}$
- Sq Rel: $\frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{d}_i - d_i|^2}{d_i}$
- log10: $\frac{1}{N}\sum_{i=1}^{N}|\log_{10}(\hat{d}_i) - \log_{10}(d_i)|$
- Threshold δ: percentage of $\hat{d}_i$, s.t. $\max(\frac{\hat{d}_i}{d_i}, \frac{d_i}{\hat{d}_i}) < \delta$, $\delta \in \{1.25, 1.25^2, 1.25^3\}$.

B. Analysis of ADGF-Net

Our AGDF-Net only uses synthetic datasets for training, and directly tests on real-world data both in outdoor and indoor scenes. Note that real-world data is not utilized in the training process. In order to obtain domain generalizable depth estimation results, our AGDF-Net is divided into three steps, including: (1) initial feature extraction; (2) intensified feature extraction; and (3) domain generalizable depth estimation. Note that our method is trained in an end-to-end manner.

1) Initial Feature Extraction: For initial feature extraction, our AGDF-Net decomposes the image into initial depth and weak-related depth parts with two encoder-decoder branches, and only the former is used for the subsequent cross-domain generalizable depth estimation. The final map of each part obtained by the network is added to obtain the full image. For outdoor scenes, the visualization results of initial depth and weak-related depth features are shown in Fig. 4(b) and (c) represent f^{Din} and f^{Dwr} at maximum resolution, which are summed at the channel level for display, respectively. As shown in Fig. 4, the initial depth features contain more structure-related information. Weak-related depth features are the remaining information separated from the image. It is consistent with [20], [21], which proves that the depth map is mainly recovered from the structural information of the image. Furthermore, we also provide the visualization of the reconstructed images ($I^{'in}$) from initial depth features of D^{in}, reconstructed images ($I^{'wr}$) from weak-related depth features of D^{wr}, final reconstructed images I', input images I and reconstruction errors in Fig. 6 for reference. The results show that $I^{'in}$ and $I^{'wr}$ are the components of the final reconstructed image, and the two are summed to form the final reconstructed image, which is almost the same as the input image. The darker the reconstruction error color, the smaller the error, and the overall reconstruction error is kept at a low level. For indoor scenes, as shown in Fig. 5(b) and (c) represent the initial depth and weak-related depth features, which are summed at the channel level using the f^{Din} and f^{Dwr} at maximum resolution. Consistent with outdoor scenes, initial depth features contain more overall structural information, while weak-related depth features contain the remaining information separated from the image, such as tile grids, checkerboards, textures of quilt and carpet, etc.

To further demonstrate the extracted initial depth and weak-related depth features, these pre-trained two features are directly used as the input to predict the depth maps with the same depth estimation networks on KITTI, respectively. Fig. 7 shows the results of the pre-trained initial depth and weak-related depth features, and the retrained results of the depth map using the corresponding features as input, respectively. Note that the pre-trained initial depth features and weak-related depth features are directly passed into the corresponding new DepthNet models, respectively. During training, the initial depth features (E^{in} and D^{in}) and weak-related depth features (E^{wr} and D^{wr}) are frozen, and the corresponding DepthNet models are trained from scratch respectively. As shown in Fig. 7, the initial depth prediction map (d^{in}) is clearer in object boundaries, while the results of the weak-related depth prediction map (d^{wr}) have blurred object boundaries and even predicts wrongly. For a quantitative comparison, we give comparisons of d^{in} and d^{wr} on the Abs_Rel, where the d^{wr} is much worse than d^{in} (0.344 versus 0.104, etc.). The clear boundaries of d^{in} indicate that initial depth features are related to depth estimation. On the contrary, using weak-related depth features that are missing depth details information separated from the image cannot correctly estimate the depth of the object, resulting in poor estimation results.

2) Intensified Feature Extraction: For intensified feature extraction, the intensified branch is applied to further optimize the extracted initial depth features, enhancing the domain invariant information. Finally, intensified depth features that are domain generalizable are obtained with adaptive guidance fusion and input to the final depth estimation network. For outdoor scenes, as shown in Fig. 4(d), where (d) is the intensified depth features f^{Dit} at maximum resolution, lane lines and building surface textures, etc., have been almost removed, resulting in domain generalizable depth feature maps that do not change with image domains. It is worth noting that our intensified depth features have a high contribution in the sky, which is the same as the depth-specific maps of S2R-DepthNet [20] have the stronger response in the sky region, and is also similar to [73]. The reason for this phenomenon is that the sky with the farthest depth value represents the vanishing point, which is an essential clue for depth estimation, and has the same infinite depth representation for the sky in different domain images, which can be seen as a strong domain invariant depth information. For indoor scenes, as shown in Fig. 5(d), interference information such as tile grids, checkerboards, and textures of quilts and carpets are almost eliminated. An interesting observation is that the indoor intensified depth features have more obvious lines and edges. Indoor scenes have a smaller depth of field than outdoor scenes, and have richer structure information than outdoor scenes. Furthermore, the sky area of outdoor scenes has a large value, and the distribution of depth range in the image is nonlinear. [20], [21] prove that image structure has a stronger contribution to depth estimation, so our intensified depth feature has more obvious lines and edges in the indoor scene with a small depth of field than in the outdoor scene with a large depth of field. And because of the nonlinear depth range, the structural lines of outdoor scenes in the visualization are not obvious enough, but the overall structure edges, such as buildings, are clearly visualized, which is also consistent with the results of [20].

Besides, these domain generalizable intensified depth features are almost consistent between the synthetic domain and the real domain. Then the intensified depth features are input into subsequent depth estimation networks to estimate cross-domain depth maps, and obtain excellent results.

3) Feature Stability of Different Style Images: To further demonstrate the extracted initial depth, weak-related depth and intensified depth features, we provide the three extracted features of images with different styles from the same scene in Fig. 8, which have similar initial depth features and nearly the same intensified depth features. As shown in Fig. 8, for images captured in the same scene with different styles, our approach can generate similar initial depth features in row (b) and almost the same intensified depth features in row (d), while weak-related depth

features have different strengths for different styles and depth detailed information is lost in row (c). We also provide the differential map of initial depth features and intensified depth features in row (e) and (f), where row (e) represents the differential map between the initial depth feature of the corresponding column and the initial depth feature of the first column, row (f) represents the differential map between the intensified depth feature of the corresponding column and the intensified depth feature of the first column. The results show that the differential map of the first column is zero relative to itself, and the differential maps of other columns are relatively small on the initial depth features and almost zero on the intensified depth features. These similar initial depth features further prove that the proposed framework can extract features that are related to depth estimation. And the almost same intensified depth features prove that our proposed intensified feature extraction process (with adaptive guidance module) can obtain fully decoupled domain generalizable depth features for subsequent depth estimation.

C. Experimental Results

Following [20], we only train on synthetic datasets and test on real-world datasets. We conduct experimental comparisons on the outdoor dataset vKITTI to KITTI and indoor dataset SUNCG to NYUDv2, including synthetic to real experiments and semi-supervised learning experiments. To show the generalization ability of our AGDF-Net, we provide comparison results from vKITTI to NuScenes, DrivingStereo and CityScapes. Besides, to further prove the generalization ability of the extracted domain generalizable depth features, we also provide comparison results of our AGDF-Net with state-of-the-art approaches with different depth estimation networks for real-world depth estimation.

1) vKITTI to KITTI: Synthetic to Real: Table I demonstrates the experimental results of vKITTI to KITTI. Following [20], we compare with the state-of-the-art unsupervised domain adaptation methods [16], [64] and domain generalization method [20]. Meanwhile, we also compare with real-world supervised depth estimation method [7], [65], [74], where [7] is trained on the KITTI dataset, MiDaS [65] is trained on MIX 5 dataset, and DPT-Hybrid [74] is trained on MIX 6 dataset which contains about 1.4 million images and finetuning on KITTI dataset. Note that the results of MiDaS are evaluated on the official pre-trained model[1] using the same input as ours, the results of DPT-Hybrid are pulled from the official paper which are finetuned on KITTI. As shown in Table I, our approach outperforms the existing state-of-the-art methods trained on vKITTI under the depth range of 50 m and 80 m for all evaluation metrics, while our approach is trained without any real-world data, and the result is still significantly better than the existing unsupervised domain adaptation methods. At the same time, our approach does not use any external images for training except vKITTI, and is still superior to the domain generalization method [20], which trained on the external real-world PBN dataset.[2] Specifically, compared with [20], Abs_Rel is reduced by 6.06%, Sq_Rel is reduced by 22.35% in the 80 m range. Besides, the qualitative comparisons are illustrated in Fig. 9, and compared with T^{2}Net [16] and S2R-DepthNet [20], our approach can obtain more obvious structures and details in depth.

Semi-Supervised Learning: Following previous methods [20], [64], [75], we selected the first 1000 frames of real-world captured KITTI for further training, which only contains 4.42% of the total dataset. This is more in line with actual usage scenarios and can be considered as a semi-supervised setting. The semi-supervised methods [64], [75] and the domain generalization based semi-supervised learning method [20] are compared here with the same settings, as shown in Table II. Our approach outperforms the compared methods on all metrics. Specifically, our approach does not use any additional dataset other than vKITTI in the domain generalization training process, and is still superior to [20] using an external dataset (PBN) on all metrics. Besides, though [76] uses more KITTI frames (7346, 32.5% of total dataset) and 12600 stereo pairs for training, our approach still surpass [76] on most evaluation metrics.

2) SUNCG to NYUDv2: Next, we report a SUNCG [29] to NYUDv2 [24] transfer experiment in the indoor scenes in Table III, which means the network is trained on SUNCG [29] and directly tested on NYUDv2 [24]. Following [7], [16], [20], we use the Eigen 654 split as the testing dataset. To compare equally, we retrain S2R-DepthNet [20] on the SUNCG under the same setting with the input size 224 $\times$ 304. The results show that our approach yields better results than S2R-DepthNet on all metrics in the indoor scene, and Abs_Rel is reduced by 2.93%. In addition, we also provide the comparison results of the deep learning based real-world supervised methods [7], [58], [65], [74] for reference, where [7], [58] are trained on the NYUDv2 dataset, MiDaS [65] is a strong supervised method trained on MIX 5 dataset, and DPT-Hybrid [74] is another strong supervised method trained on MIX 6 dataset which contains about 1.4 million images and finetuning on NYUDv2 dataset. Note that the results of MiDaS are evaluated on the official pre-trained model[3] using the same input as ours, the results of DPT-Hybrid are pulled from the official paper which are finetuned on NYUDv2. In order to evaluate the effectiveness of our approach in indoor scenes, we report qualitative results on NYUDv2 in Fig. 10, which shows that our approach can capture more complete object boundaries than S2R-DepthNet, such as chairs, desk lamps, etc.

D. Generalization on More Datasets

We then verify the cross-dataset generalized performance on different datasets, including vKITTI [28] to NuScenes [25], DrivingStereo [26] and CityScapes [27]. In other words, the network is trained on the synthetic vKITTI dataset and directly evaluated on NuScenes, DrivingStereo and CityScapes.

NuScenes [25] dataset is a large-scale autonomous driving dataset with 3D object annotations, including 1000 driving scenes, where 850 scenes are for training and validation, and 150 scenes for testing. In this paper, we randomly select 174 frames

[1][Online]. Available: https://github.com/isl-org/MiDaS
[2][Online]. Available: https://www.kaggle.com/c/painter-by-numbers
[3][Online]. Available: https://github.com/isl-org/MiDaS

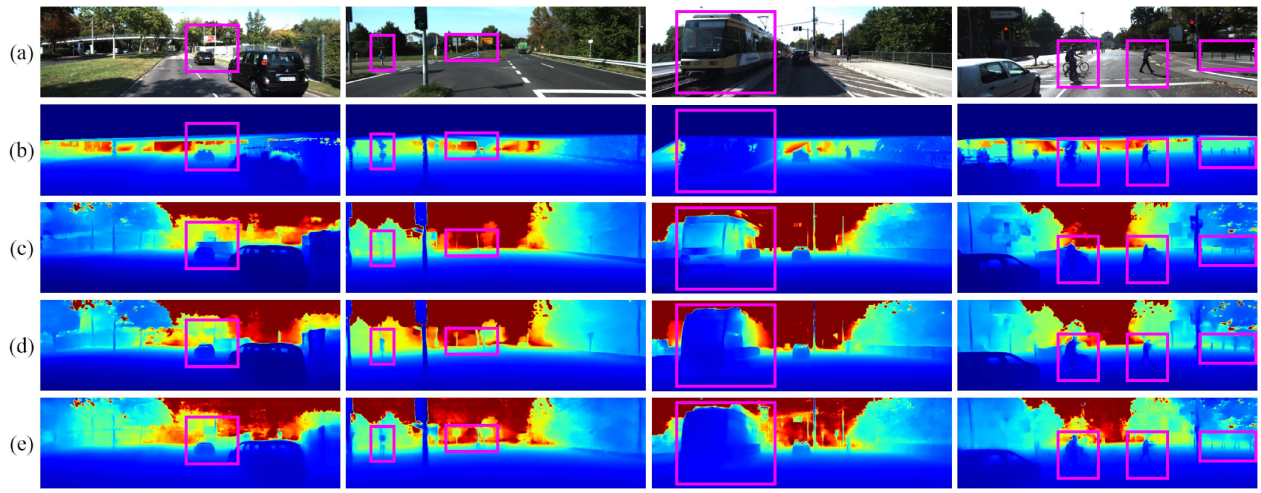


Fig. 9. Qualitative comparison with state-of-the-art methods on KITTI. From top to bottom: (a) Image, (b) Ground Truth, (c) T^{2}Net [16], (d) S2R-DepthNet [20], and (e) Ours. Compared with [16], [20], our approach can obtain clearer object boundaries, such as persons and signs.

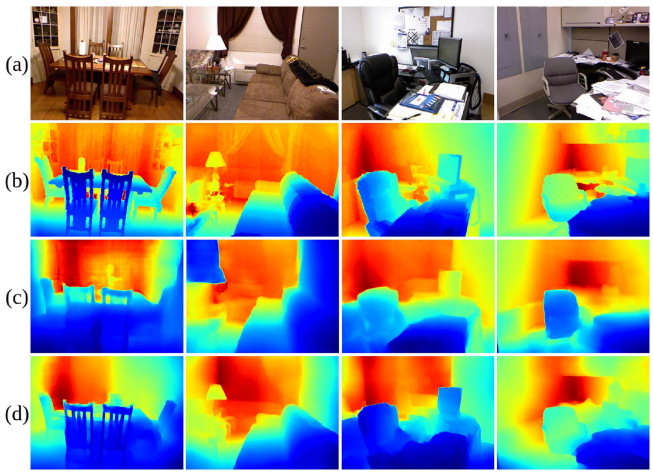


Fig. 10. Qualitative comparison with state-of-the-art methods on NYUDv2. From top to bottom: (a) Image, (b) Ground Truth, (c) S2R-DepthNet [20], and (d) Ours.

from the test set for evaluation. In evaluation, the original image resolution is first cropped from 900 $\times$ 1600 to 480 $\times$ 1600, then downsampled to 192 $\times$ 640.

DrivingStereo [26] dataset is a large-scale stereo dataset, containing over 180 k images covering a diverse set of driving scenarios. We randomly select 463 frames from the testing data for evaluation. The original image resolution is first cropped from 800 $\times$ 1762 to 528 $\times$ 1762, then downsampled to 192 $\times$ 640 for evaluation.

CityScapes [27] dataset is a large-scale dataset that contains a diverse set of stereo video sequences recorded in street scenes from 50 different cities. We randomly select 495 frames from the testing data for evaluation. The original image resolution is first cropped from 1024 $\times$ 2048 to 614 $\times$ 2048, then downsampled to 192 $\times$ 640 for evaluation.

For a fair comparison, all approaches are evaluated under the same setting, which is consistent with vKITTI to KITTI. Note that T^{2}Net [16] and S2R-DepthNet [20] are evaluated on different datasets with officially provided pre-trained models.[4] As shown in Table IV, our approach outperforms the unsupervised domain adaptation method [16] and the domain generalization method [20] on all metrics. The qualitative results are shown in Fig. 11, from which we can see that our approach can better estimate the object boundaries, such as people, cars, trunks, signs, etc.

E. Generalization on Different Depth Estimation Networks

To further demonstrate the generalization of our approach, we present the experimental results of our AGDF-Net and state-of-the-art approaches with different depth estimation networks, including the DepthNet network used in the previous work [16], [20] (DepthNet in Table V), commonly used ResNet based depth estimation network [13] without pre-trained on ImageNet (RB-Net in Table V) and commonly used ResNet based depth estimation network [13] with pre-trained on ImageNet (RB-Net (pre-trained) in Table V). Note that when using RB-Net [13] as the depth estimation network, the results of S2R-DepthNet are retrained with the officially provided code. Furthermore, as shown in Table V, Ours with pre-trained models on ImageNet can achieve better results than without pre-trained ones, where Abs_rel of ours with and without pre-trained models are 0.164 and 0.168, respectively. As shown in Table V, using DepthNet and RB-Net, the Abs_rel results of S2R-DepthNet [20] are 0.165 and 0.186, respectively, and the results of our approach are 0.155 and 0.168, respectively. Our approach has more stable results on all evaluation metrics under different depth estimation networks. Furthermore, the learned domain generalizable depth features are completely decoupled from the subsequent depth estimation network. Therefore, these features can be cascaded

[4] T^{2}Net: https://github.com/lyndonzheng/Synthetic2Realistic, S2R-DepthNet: https://github.com/microsoft/S2R-DepthNet

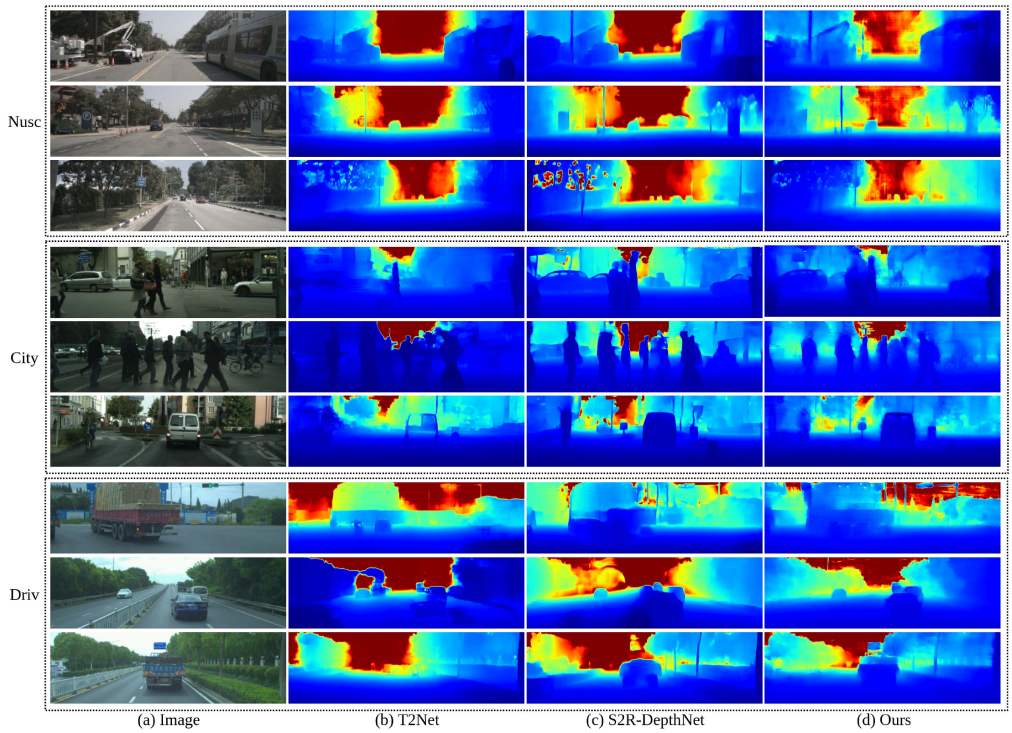


Fig. 11. Qualitative comparison with state-of-the-art methods on more datasets. From left to right: (a) Image, (b) T^{2}Net [16], (c) S2R-DepthNet [20], (d) Ours. "Nusc", "City" and "Driv" denote NuScenes, CityScapes and DrivingStereo datasets, respectively.

TABLE IX
ADAPTIVE GUIDANCE ANALYSIS ON DIFFERENT KERNEL SIZES

Kernel Size	Abs Rel	Sq Rel	RMSE	RMSE$_{log}$	δ <1.25	δ <1.25^2	δ <1.25^3
1x1	0.162	1.159	5.356	0.229	0.784	0.934	0.976
3x3	0.163	1.184	5.628	0.235	0.773	0.928	0.974
5x5	0.171	1.422	5.851	0.242	0.769	0.926	0.971
ALL	0.155	1.049	5.388	0.226	0.782	0.936	0.977

with any other depth estimation networks to improve the generalization of depth estimation, and further narrow the gap between the synthetic domain and the real domain.

F. Inference Time

Our approach achieves real-time performance, and the average inference time of our network on the KITTI dataset is 0.037 s using a GTX 1080 GPU. The size of the input image is 192×640.

G. Ablation Study

We provide the ablation study in the outdoor scene vKITTI to KITTI in Tables VI, VII, VIII and IX. All results are obtained within the depth range of 80 m.

Initial Feature Extraction (INE): The baseline of our network is the same as [16], [20], and the results are shown in the first row of Table VI. The initial feature extraction (INE) process consists of an initial branch (INB) and a weak-related branch (WRB), where the initial branch is used to extract initial depth features, and the weak-related branch is used to extract weak-related depth features. The initial depth features extracted by INE are directly input into the baseline network for depth estimation, and the results are shown in the second row of Table VI. The results show that depth estimation using the initial depth features extracted by our framework can significantly improve the Abs_rel from 0.230 to 0.173, compared to directly inputting images into DepthNet [20]. The comparison shows that separating the image into initial depth and weak-related depth components can extract depth related information preliminarily for cross-domain depth estimation, and generalize it from the synthetic scene to the real.

Specifically, in order to analyze the effectiveness of each module and its loss function in the INE process, the ablation study results are shown in Table VII, where INB and WRB denote the initial depth branch and weak-related depth branch respectively. When adding both INB and WRB to the baseline (second row), the Abs_rel reduced from 0.184 to 0.179 compared to adding only INB (first row). This shows that separating images into initial depth and weak-related depth components is effective for cross-domain generalizable depth estimation and can improve generalization. Note that after adding INB and WRB to the

baseline, the reconstruction loss ($\mathcal{L}_{rec}$) is used to constrain the images of the two branches to reconstruct back to the input image. The third row in Table VII shows the results of adding contrary loss ($\mathcal{L}_{con}$), Abs_rel decreased from 0.179 to 0.173. This shows that the contrary loss can effectively disentangle the features of the initial branch and the weak-related branch, and obtain more generalized initial depth features.

Another interesting result is the improvement of adding the initial branch (hourglass architecture) only to the baseline (Abs_Rel 0.230 versus 0.184). As shown in Table VIII, we provide the experiments of cascading multiple (1-4) hourglass architectures before the depth estimation network. With the stacking of hourglass structures from 2 to 4, there is little impact on the depth estimation results. It is proved that the results of stacking the hourglass network before the depth estimation network are limited and the features related to depth estimation can not be extracted only with cascaded hourglass architecture. Besides, our approach with initial and weak-related branches can improve the performance from 0.184 to 0.155, which also proves that the proposed framework can separate effective features related to depth estimation from the image.

Intensified Feature Extraction (ITE): The intensified feature extraction (ITE) process consists of the intensified branch. The obtained initial depth features are input to the intensified branch to get intensified depth features that are domain generalizable. Rows 3,4,6 of Table VI show the analysis of the ITE process, where the third row represents adding the intensified branch, and fuses the initial depth features and the features of the intensified branch in an additive manner. The purpose of the intensified branch is to enhance the previously obtained initial depth features. Traditional fusion methods such as addition and concatenation cannot adaptively fuse and enhance features, so the performance is still limited, as shown in rows 3 and 4 of Table VI. Therefore, we propose an adaptive feature fusion and enhancement module named adaptive guidance (AG). By learning different convolution kernel parameters, the interference information in the features is further eliminated, and the domain invariant information is enhanced. As shown in row 6 of Table VI, after adding AG, Abs_rel dropped from 0.164 to 0.155, demonstrating the effectiveness of AG. Additionally, rows 5 and 6 of Table VI further prove the effectiveness and essentiality of extracting initial depth features using two branches, where row 5 denotes the full method using a single initial branch to extract initial depth features(Abs_Rel: 0.171), and row 6 means the full method with two branches to separate initial and weak-related depth features(Abs_Rel: 0.155).

Adaptive Guidance Kernel Analysis: We design convolution kernels of different sizes in the AG for feature fusion. The performance of feature fusion of different sizes of convolution kernels is provided in Table IX. Lines 1–3 show the results of only learning 1×1, 3×3 and 5×5 kernel parameters and guiding feature fusion, respectively. Row 4 shows the result of adding convolution kernels of three sizes before guiding feature fusion. The results show that compared with learning a single kernel parameter, the fusion of multiple convolution kernels can enhance the features in different receptive fields, and then obtain better depth estimation results.

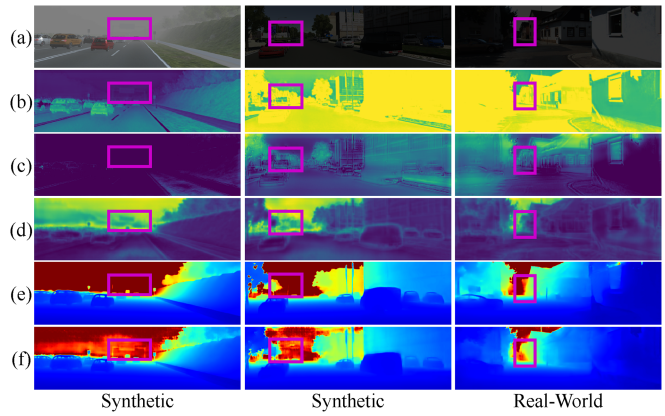


Fig. 12. Visualization of the learned intermediate features and depth maps of failure cases. From top to bottom: (a) Input Image, (b) Initial Depth Feature, (c) Weak-related Depth Feature, (d) Intensified Depth Feature, (e) S2R-DepthNet [20] Depth Map, and (f) Ours Depth Map. The first and second columns represent the features learned on the synthetic dataset (vKITTI), and column 3 represents the features learned on the real-world dataset (KITTI). The results show that our approach performs unsatisfactorily in extremely harsh conditions, such as fog (the first column) and darkness (columns 2–3). The areas circled in purple specifically indicate imperfect depth estimation results for low-visibility and distant objects in the image (signs or houses).

H. Failure Cases

Fig. 12 shows some failure examples of our approach (row (f)) and S2R-DepthNet [20] (row (e)). The results show that both our approach and S2R-DepthNet perform unsatisfactorily in extremely harsh conditions, such as fog (the first column) and darkness (columns 2–3). The areas circled in purple specifically indicate imperfect depth estimation results for low-visibility and distant objects in the image (signs or houses), where our approach cannot estimate the complete circled signs or houses while the S2R-DepthNet cannot recover them. Due to the limited image collection information under extremely harsh conditions (fog or darkness, etc.), some objects are obscured by fog or the outline of the object is unclear due to darkness. This limits the proposed framework to separate depth-related information and extract domain generalizable features, ultimately leading to depth estimation failure. Extremely harsh conditions have always been one of the unaddressed problems in the field of depth estimation. Most current depth estimation methods, such as S2R-DepthNet and our approach, cannot estimate the depth under such conditions well. Future research can focus on depth estimation for low-visibility images. More general depth estimation results can be obtained by designing separation modules under harsh conditions, or uniform image depth estimation can be achieved after improving visibility through pre-image processing.

V. Conclusion

In this paper, we propose a novel domain generalizable depth feature extraction network with adaptive guidance fusion, i.e., AGDF-Net, to more fully acquire essential depth features that are domain generalizable at multi-scale feature levels. Our approach can be well generalized to unseen real-world images with models only trained on synthetic data. We separate the image into initial depth and weak-related depth components with reconstruction loss and contrary loss to efficiently extract useful information for cross-domain generalizable depth estimation. The key is to extract intensified depth features from initial depth features. Therefore, an adaptive guidance fusion module is designed to sufficiently reuse and intensify the extracted initial depth features at multi-scale levels to get intensified depth features that are domain generalizable. This module can further enhance the depth related components and eliminate the disturbing components. Therefore, with the extracted intensified depth features that are domain generalizable, the gap bottleneck of the synthetic domain to the real domain is broken and further narrowed. Our AGDF-Net can be well applied to the various depth estimation datasets and achieve state-of-the-art performance without using any real-world dataset. The experiments using a small amount of labeled real-world data in a semi-supervised setting also demonstrate the superiority of our AGDF-Net over state-of-the-art approaches. In the future, we would like to apply our AGDF-Net to other pixel-level estimation tasks, such as stereo matching and semantic segmentation, etc. Furthermore, another challenging domain in monocular depth estimation is outdoor to indoor domain generalization. In the future, we would like to apply our AGDF-Net from outdoor to indoor domain generalization to obtain the depth of field generalization abilities by designing loss functions like [65] or other disentangle strategies.

References

[1] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6602–6611.

[2] H. Zhan, R. Garg, C. S. Weerasekera, K. Li, H. Agarwal, and I. Reid, "Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 340–349.

[3] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "MVSNet: Depth inference for unstructured multi-view stereo," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 785–801.

[4] Z. Teed and J. Deng, "DeepV2D: Video to depth with differentiable structure from motion," 2018, *arXiv: 1812.04605*.

[5] R. T. Azuma, "A survey of augmented reality," *Presence: Teleoperators Virtual Environ.*, vol. 6, pp. 355–385, 1997.

[6] J. Carmigniani, B. Furht, M. Anisetti, P. Ceravolo, E. Damiani, and M. Ivkovic, "Augmented reality technologies, systems and applications," *Multimedia Tools Appl.*, vol. 51, pp. 341–377, 2011.

[7] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," 2014, *arXiv:1406.2283*.

[8] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2002–2011.

[9] J. Li, R. Klein, and A. Yao, "A two-streamed network for estimating fine-scaled depth maps from single RGB images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2017, pp. 3392–3400.

[10] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe, "Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 161–169.

[11] A. Atapour-Abarghouei and T. P. Breckon, "Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2800–2810.

[12] S. Zhao, H. Fu, M. Gong, and D. Tao, "Geometry-aware symmetric domain adaptation for monocular depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9780–9790.

[13] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3827–3837.

[14] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, "3D packing for self-supervised monocular depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2482–2491.

[15] A. Gordon, H. Li, R. Jonschkowski, and A. Angelova, "Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8976–8985.

[16] C. Zheng, T.-J. Cham, and J. Cai, "T2Net: Synthetic-to-realistic translation for solving single-image depth estimation tasks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 798–814.

[17] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 343–351.

[18] Z. Yi, H. Zhang, P. Tan, and M. Gong, "DualGAN: Unsupervised dual learning for image-to-image translation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2017, pp. 2868–2876.

[19] A. Tonioni, M. Poggi, S. Mattoccia, and L. D. Stefano, "Unsupervised domain adaptation for depth prediction from images," *IEEE Trans. Pattern Analalysis Mach. Intell.*, vol. 42, no. 10, pp. 2396–2409, Oct. 2020.

[20] X. Chen, Y. Wang, X. Chen, and W. Zeng, "S2R-DepthNet: Learning a generalizable depth-specific structural representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3033–3042.

[21] T. V. Dijk and G. D. Croon, "How do neural networks see depth in single images?," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 2183–2191.

[22] J. Tang, F.-P. Tian, W. Feng, J. Li, and P. Tan, "Learning guided convolutional network for depth completion," *IEEE Trans. Image Process.*, vol. 30, pp. 1116–1129, Dec. 2021.

[23] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, pp. 1231–1237, 2013.

[24] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2012, pp. 746–760.

[25] H. Caesar et al., "nuScenes: A multimodal dataset for autonomous driving," 2019, *arXiv: 1903.11027*.

[26] G. Yang, X. Song, C. Huang, Z. Deng, J. Shi, and B. Zhou, "DrivingStereo: A large-scale dataset for stereo matching in autonomous driving scenarios," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 899–908.

[27] M. Cordts et al., "The Cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.

[28] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4340–4349.

[29] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic scene completion from a single depth image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 190–198.

[30] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2015, pp. 2650–2658.

[31] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *Proc. Int. Conf. 3D Vis.*, 2016, pp. 239–248.

[32] F. Ma, G. V. Cavalheiro, and S. Karaman, "Self-supervised sparse-to-dense: Self-supervised depth completion from LiDAR and monocular camera," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2019, pp. 3288–3295.

[33] X. Cheng, P. Wang, and R. Yang, "Learning depth with convolutional spatial propagation network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2361–2379, Oct. 2020.

[34] L. Liu, Y. Liao, Y. Wang, A. Geiger, and Y. Liu, "Learning steering kernels for guided depth completion," *IEEE Trans. Image Process.*, vol. 30, pp. 2850–2861, Feb. 2021.

[35] H. Yan, S. Zhang, Y. Zhang, and L. Zhang, "Monocular depth estimation with guidance of surface normal map," *Neurocomputing*, vol. 280, pp. 86–100, 2018.

[36] X. Qi, Z. Liu, R. Liao, P. H. Torr, R. Urtasun, and J. Jia, "GeoNet++: Iterative geometric neural network with edge-aware refinement for joint depth and surface normal estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 2, pp. 969–984, Feb. 2022.
[37] A. Mousavian, H. Pirsiavash, and J. Košecká, "Joint semantic segmentation and depth estimation with deep convolutional networks," in *Proc. Int. Conf. 3D Vis.*, 2016, pp. 611–619.
[38] Z. Zhang, Z. Cui, C. Xu, Z. Jie, X. Li, and J. Yang, "Joint task-recursive learning for semantic segmentation and depth estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 238–255.
[39] J. Chen, X. Yang, Q. Jia, and C. Liao, "DENAO: Monocular depth estimation network with auxiliary optical flow," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2598–2610, Aug. 2021.
[40] L. Liu et al., "FCFR-Net: Feature fusion based coarse-to-fine residual learning for depth completion," in *Proc. 35th AAAI Conf. Artif. Intell.*, 2021, pp. 2136–2144.
[41] J. Qiu et al., "DeepLiDAR: Deep surface normal guided depth prediction for outdoor scene from sparse LiDAR data and single color image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3308–3317.
[42] J. Watson, M. Firman, G. J. Brostow, and D. Turmukhambetov, "Self-supervised monocular depth hints," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 2162–2171.
[43] A. Johnston and G. Carneiro, "Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4755–4764.
[44] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018.
[45] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2015, pp. 4068–4076.
[46] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2208–2217.
[47] Y. Li, N. Wang, J. Shi, J. Liu, and X. Hou, "Revisiting batch normalization for practical domain adaptation," 2016, *arXiv:1603.04779*.
[48] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by back-propagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.
[49] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 95–104.
[50] R. Xu, G. Li, J. Yang, and L. Lin, "Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1426–1435.
[51] Y. Liu, K. Wang, G. Li, and L. Lin, "Semantics-aware adaptive knowledge distillation for sensor-to-vision action recognition," *IEEE Trans. Image Process.*, vol. 30, pp. 5573–5588, Jun. 2021.
[52] Y. Liu, Z. Lu, J. Li, T. Yang, and C. Yao, "Deep image-to-video adaptation and fusion networks for action recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 3168–3182, Dec. 2020.
[53] Y. Liu, Z. Lu, J. Li, and T. Yang, "Hierarchically learned view-invariant representations for cross-view action recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 8, pp. 2416–2430, Aug. 2019.
[54] J. Wang, C. Lan, C. Liu, Y. Ouyang, W. Zeng, and T. Qin, "Generalizing to unseen domains: A survey on domain generalization," 2021, *arXiv:2103.03097*.
[55] S. Shankar, V. Piratla, S. Chakrabarti, S. Chaudhuri, P. Jyothi, and S. Sarawagi, "Generalizing across domains via cross-gradient training," 2018, *arXiv: 1804.10745*.
[56] F. Qiao, L. Zhao, and X. Peng, "Learning to learn single domain generalization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 12553–12562.
[57] X. Jin, C. Lan, W. Zeng, Z. Chen, and L. Zhang, "Style normalization and restitution for generalizable person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3140–3149.
[58] B. Li, C. Shen, Y. Dai, A. Van Den Hengel, and M. He, "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1119–1127.
[59] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi, "Domain generalization by solving jigsaw puzzles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2224–2233.
[60] Z. Huang, H. Wang, E. P. Xing, and D. Huang, "Self-challenging improves cross-domain generalization," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2020, pp. 124–140.
[61] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2017, pp. 1510–1519.
[62] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 179–196.
[63] H. Kazemi, S. M. Iranmanesh, and N. Nasrabadi, "Style and content disentanglement in generative adversarial networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2019, pp. 848–856.
[64] J. N. Kundu, P. K. Uppala, A. Pahuja, and R. V. Babu, "AdaDepth: Unsupervised content congruent adaptation for depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2656–2665.
[65] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1623–1637, Mar. 2022.
[66] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[67] J. T. Todd, "The visual perception of 3D shape," *Trends Cogn. Sci.*, vol. 8, pp. 115–121, 2004.
[68] D. Marr, *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*. Cambridge, MA, USA: MIT Press, 2010.
[69] F. Zhou, Q. Chen, B. Liu, and G. Qiu, "Structure and texture-aware image decomposition via training a neural network," *IEEE Trans. Image Process.*, vol. 29, pp. 3458–3473, Dec. 2020.
[70] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1800–1807.
[71] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[72] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6612–6619.
[73] J. Hu, Y. Zhang, and T. Okatani, "Visualization of convolutional neural networks for monocular depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3868–3877.
[74] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 12159–12168.
[75] Y. Zhao, S. Kong, D. Shin, and C. Fowlkes, "Domain decluttering: Simplifying images to mitigate synthetic-real domain shift and improve depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3327–3337.
[76] Y. Kuznietsov, J. Stuckler, and B. Leibe, "Semi-supervised deep learning for monocular depth map prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2215–2223.

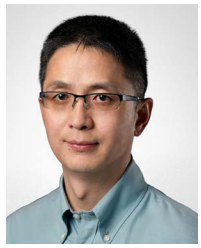
Lina Liu received the BS and PhD degrees in control science and engineering from Zhejiang University, Zhejiang, China, in 2018 and 2023, respectively. She is currently a researcher with China Mobile Research Institute, China Mobile Limited. She served as reviewer of *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Circuits and Systems for Video Technology*, ICCV, AAAI, ICRA, IROS, etc. Her research interests include 3D vision, information fusion, computer vision, and deep learning.

Xibin Song received the BE degree in digital media and technology from Shandong University, Jinan, China, in 2011, and the PhD degree from the School of Computer Science and Technology, Shandong University, in 2017. He is a senior researcher with Robotics and Autonomous Driving Laboratory (RAL) of Baidu. He worked as a joint PhD student with the Research School of Engineering, Australian National University, Canberra, Australia in 2015–2016. He served as reviewer of the *IEEE Transactions on Image Processing*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *IEEE Transactions on Circuits and Systems for Video Technology*, CVPR, ICCV, ECCV, AAAI, etc. His research interests include computer vision and augmented reality.

Mengmeng Wang received the BS and MS degrees in control science and engineering from Zhejiang University, Zhejiang, China, in 2015 and 2018, respectively. She is currently working toward the PhD degree with the Laboratory of Advanced Perception on Robotics and Intelligent Learning, College of Control Science and Engineering, Zhejiang University. Her research interests include visual tracking, action recognition, computer vision, and deep learning.

Yuchao Dai (Member, IEEE) received the BE, ME, and PhD degrees all in signal and information processing from Northwestern Polytechnical University (NPU), in 2005, 2008, and 2012, respectively. He is currently a professor with the School of Electronics and Information, Northwestern Polytechnical University (NPU), Xi'an, China. He was an ARC DECRA fellow with the Research School of Engineering, Australian National University, Canberra, Australia. His research interests include structure from motion, multi-view geometry, low-level computer vision, deep learning, compressive sensing, and optimization. He won the Best Paper Award at IEEE CVPR 2012, the Best Paper Award Nominee at IEEE CVPR 2020, the DSTO Best Fundamental Contribution to Image Processing Paper Prize at DICTA 2014, the Best Algorithm Prize in NRSFM Challenge at CVPR 2017, the Best Student Paper Prize at DICTA 2017 and the Best Deep/Machine Learning Paper Prize at APSIPA ASC 2017. He served as area chair for IEEE CVPR, ICCV, ACM MM, and etc.

Yong Liu (Member, IEEE) received the BS degree in computer science and engineering and the PhD degree in computer science from Zhejiang University, Zhejiang, China, in 2001 and 2007, respectively. He is currently a professor with the Institute of Cyber-Systems and Control, Zhejiang University. His main research interests include Robot perception and vision, deep learning, Big Data analysis, and multi-sensor fusion. His research interests include machine learning, computer vision, information fusion, and robotics.



Liangjun Zhang received the BS/MS degrees from Zhejiang University, and the PhD degree in computer science from the University of North Carolina at Chapel Hill, in 2009. He is currently the director with Robotics and Autonomous Driving Lab (RAL), Baidu Research USA and China. He was an NSF Computing Innovation fellow with the Computer Science Department, Stanford University from 2009 to 2011. His research interests span robotics, autonomous driving, computer vision, simulation, and geometric computing. He published research papers at *Science Robotics*, *The International Journal of Robotics Research*, *IEEE Transactions on Robotics*, *IEEE Transactions on Intelligent Transportation Systems*, *IEEE Transactions on Multimedia*, ICCV, CVPR, ECCV, RSS, ICRA, IROS, ACC, and AAAI. He has received a number of awards including the First Place of nuScenes Detection Challenge organized in conjunction with ICRA 2021, the Best Paper Award at the International CAD Conference 2008 and the UNC Linda Dykstra Distinguished PhD Dissertation Award.