Unsupervised Monocular Depth Estimation for Foggy Images with Domain Separation and Self-depth Domain Conversion

Fuyang Liu
Central South University of Forestry and Technology
Changsha,China
20221100373@csuft.edu.cn

Jianjun Li
Central South University of Forestry and Technology
Changsha,China
T20010539@csuft.edu.cn

Abstract

The depth estimation of foggy images has always been a major challenge in the research field. The common depth estimation methods use supervised training to estimate the depth of foggy images, but their effectiveness is often limited by the domain adaptation characteristics of supervised training. Here, we propose an unsupervised domain separation depth estimation algorithm for foggy images. This algorithm adopts an unsupervised approach and designs a domain separation framework for foggy and clear images to perform depth estimation on foggy images. It utilizes the characteristic that depth information can be used for both dehazing and hazing, incorporating a self-depth domain conversion module that constructs a symmetric training framework. Domain separation separates the information of the image itself in the feature space dimension, breaking it down into two parts: exclusive domain information (color, lighting, fog degree, etc.) and common domain information (depth information). The experimental results show that our designed network can achieve state-of-the-art results on the NYUv2 dataset and SUN RGB-D dataset, which is superior to existing advanced depth estimation algorithms. Furthermore, the algorithm has strong robustness and can accurately estimate the corresponding depth maps for both non-foggy and foggy images.

***Keywords:** monocular depth estimation foggy image domain separation unsupervised learning domain conversion.*

1. Introduction

Monocular depth estimation is an important research subject in the field of computer vision, which is of great significance in stereo matching [39], scene understanding [5], 3D reconstruction [29], etc. It can mine and provide depth information and spatial position relationship information of images. Monocular depth estimation opens the possibility of mapping images from two-dimensional space to three-dimensional space, being able to explore the relative relationships between objects and perform numerical measurements while understanding the hierarchy, structure, and layout of the scene. In the field of image dehazing, monocular depth estimation can also serve as auxiliary information to guide image restoration [21].

However, monocular depth estimation itself is an ill-posed problem, as for an RGB-formatted scene image, it corresponds to countless depth space mapping relationships that conform to visual perception. There are infinite ways in which three-dimensional space can generate the same projection in two-dimensional space. The inherent nature of this problem adds great difficulty and cost to monocular depth estimation. A large number of existing work centers on paired scene image-depth map datasets. The depth estimation of such methods largely depends on the quality of the dataset [6]. When the dataset itself can provide relatively accurate depth, an elaborately designed network structure and algorithm can usually estimate a satisfactory depth map. When the quality of the ground truth depth map itself is poor, with a large amount of noise and inaccurate edge information, relevant algorithms usually cannot solve this problem well. Meanwhile, collecting paired RGB scene images and their corresponding depth maps is a time-consuming and laborious task. The publicly available datasets are limited, and the cost of creating such a dataset

is huge. In this situation, unsupervised monocular depth estimation methods have become a major pillar in the field of depth estimation.

A large part of existing unsupervised methods are models constructed based on consistency constraints on stereo image pairs [11, 12, 1, 33, 32, 44]. These methods rely on continuous video frame images for modeling and training of networks, as depth estimation requires geometric constraints using consistency information. This method heavily relies on the dataset, and in some cases where only one single scene image is provided, this method will naturally fail. Meanwhile, this method estimates depth maps with inaccuracies. In addition to self-supervised methods based on disparity maps, many existing networks implement unsupervised training by redesigning network architectures, including utilizing multi-scale or multi-level network structures, cascaded network models, multiple decoders, and other structures [34, 18, 45, 40, 28]. Essentially, these methods increase the complexity and number of parameters of the model to achieve better depth estimation fitting results. Such networks often have high training costs and weak generalization ability. In addition, in recent years, there have been works using Transformers and diffusion models for monocular depth estimation [47, 35, 26], but these network models often have high requirements for training resources such as GPUs, making it difficult to conduct research under limited costs.

In the light of the issues in the above methods, we propose an innovative unsupervised monocular depth estimation algorithm for foggy images with domain separation and self-depth domain conversion. This algorithm adopts an unsupervised approach to design domain separation for paired foggy and non-foggy images, extracting their common domain information, namely depth, to estimate the depth map. Exclusive domain features, such as lighting, texture, color, and degree of fog are extracted from the image itself and are combined with common domain information to obtain complete image information for reconstruction. The domain separation of images uses two encoding and decoding networks to extract common and exclusive information for each foggy/non-foggy image, relying on orthogonality loss to ensure complementary features. At the same time, for clear images without fog, a blurring operation is performed before sending them into the exclusive domain information extraction network, which increases the difficulty of network learning while losing some information, allowing the model to discover more robust features and pay more attention to the extraction of detailed information when learning depth. Networks for extracting common information share weights and impose consistency loss constraints on the results of two images. After obtaining the depth information, the original foggy and clear images use the estimated depth maps respectively, combined with the atmospheric scattering model, to carry out the self-depth domain conversion, and the corresponding fogless and foggy maps are acquired to implement the corresponding loss calculations, which indirectly constrains the depth estimation effect and completes the whole training process of the network.

- An unsupervised monocular depth estimation method for foggy images has been proposed. It can perform good monocular depth estimation on both foggy and non-foggy images.
- A targeted domain separation and self-depth domain conversion framework was designed to decompose the features of images into common and exclusive domains to extract depth information, and to use the self-estimated depth for domain conversion of input paired foggy and non-foggy images. In addition, blurring is introduced to enable the network to mine more details when extracting deep information, and to learn more robust features for reconstruction.
- The effectiveness was validated on the NYUv2 and SUN RGB-D datasets, achieving the state-of-the-art experimental results.

2. Related Work

This part mainly summarizes the related research work in the monocular depth estimation area, centering on the unsupervised methods.

The most common unsupervised monocular depth estimation methods are based on the stereo pair disparity images, which imposes geometric constraint on the model. Godard et al. [11] utilized epipolar geometry constraints and disparity consistency loss for generating improved depth estimations unsupervisedly. In another work by Godard [12], several adjustments are proposed for large improvements when training with stereo pairs, including minimum reprojection loss, auto-masking loss and a multi-scale sampling method. Filippo et al. [1] utilized GAN [13] to tackle unsupervised depth estimation through warping images with depth maps generated to fool the discriminator. Poggi et al. [33] trained one CNN to predict the left and right stereo pair disparity images accompanying a central image to do unsupervised monocular depth estimation in a trinocular way. In another work by Poggi [32], a pyramid architecture with multi-level features extracted is designed to refine the ultimate depth map with up-sampling. Zhan et al. [44] proposed the usage of stereo sequences from both spatial and temporal aspects for learning depth and visual odometry simultaneously.

Some researchers cope with this problem using the semantic segmentation information as an aid when estimating depth. Chen et al. [5] delt with this problem by transforming the input image into a scene representation and extracting depth estimation and semantic segmentation at the same

time with their alignment ensured. Li et al. [5] extracted the semantic priors of objects using a semantic segmentation network for being fused with the original image to learn depth, which enhances structure perception.

Apart from this, some approaches are proposed based on a new design of the network architecture. Pilzer et al. [31] utilized a design named cycle-inconsistency within the refinement of the depth map. To be specific, it includes estimating a disparity map of a frame for recovery of the opposite view one, backward as well. The inconsistency of the two frames is exploited for refining a final depth map. Ren et al. [34] incorporated a depth basis decoder with multiple coefficient modules as a co-teaching ensemble for learning depth from diverse sources. Zhao et al. [47] first applied the ViT (Vision Transformer) to the self-supervised monocular depth estimation and combined it with convolutions to reason locally and globally. In Hui's work [18], monocular depth estimation and complete 3D motion prediction are jointly trained together to recurrently refine the estimated result with encoder and decoder features fused iteratively. Zhang et al.[45] proposed a multi-scale structure that downsampled the estimated depth map and implemented image synthesis at various resolutions and used a structural similarity pyramid loss to improve the locality of photometric error. Wang et al. [40] proposed a cascaded depth estimation network towards the ill-posed regions in estimation and designed corresponding feature extraction networks and a pose estimation network using attention mechanism. Armin et al. [28] extracted image features from an auto-encoder and utilized multi-scale graph convolutional networks to do depth estimation self-supervisedly. Saxena et al. [35] applied the denoising diffusion model to the monocular depth estimation with infilling and step-unrolled denoising diffusion training and achieved well effectiveness. Lin et al. [26] utilized the Transformer and convolutional neural network to model long-range dependencies and local correlations simultaneously when extracting hybrid image features for depth estimation. Besides, a bowknot-type fuser aimed at aligning features and bridging local and global semantic representations is devised. Guo et al. [14] finetuned an optical flow estimation network to supervise monocular depth estimation with optical flow and multi-scale feature maps generated for loss calculating.

Finally, there exist novel methods dealing with this problem from other aspects. Zhao et al. [46] used one image transfer-based framework with domain adaptation for highly complex scenes. They adapted the day-time training to night-time training and raised one image adaptation approach to improve the performance of model after adapted. Zhu et al.[49] substituted a flow distillation loss for the typically used photometric loss and used a prior flow-based mask to eliminate noise in training loss. Shao et al.[36] took monocular depth estimation as a classification-regression problem where bin centers are used to estimate depth and put forward an elastic target bin adjusting flexibly according to the depth uncertainty. Gasperini et al. [9] focused on the depth estimation under challenging conditions and designed a framework trained with a mixture of good weather images and translated adverse weather images. Sun et al.[24] introduced pseudo-depth from external pretrained monocular depth estimation network and designed corresponding modules to enhance self-supervised training. Zhen et al.[25] proposed a sparse depth densification method by unsupervised image segmentation combined with sparse depth and further corrected it by estimating the potential error. Han et al. [15] used the image activity measure to segment image features, which boosts the perception of network and designed depth consistency loss to give more accurate estimation in weak-texture regions. In addition, recently Piccinelli et al. [30] explored estimating metric 3D points solely from an input image using a pseudo-spherical output representation, achieving accurate monocular metric depth estimation.

3. Methodology

3.1. Network architecture

The overall network workflow is mainly divided into two parts, namely the domain separation part and the self-depth domain conversion part. Its structure is shown in Fig 1.

3.1.1 Domain Separation

The domain separation part refers to [27], which mainly involves orthogonal decomposition of the input image in the feature space to obtain exclusive features – exclusive domain, and common features – common domain.

Exclusive domain contains the exclusive information of an image, including texture, color, lighting, degree of fog, and other information. The information extracted from this domain for foggy and non-foggy images is different, representing the unique style and intuitive representation of two images in the same scene under different natural lighting environments. The common domain refers to the information shared by both foggy and non-foggy images for the same scene. In our research topic, it is evident that this part of information is the depth information of the scene.

For the depth information of the common domain, we can use it to estimate the depth map of the image. Combining the information of the common domain and the exclusive domain, after decoding with the decoder, we can use it to recover the image and obtain the reconstructed images of the original foggy and non-foggy input images, respectively. This design of reconstruction is the main cornerstone of our unsupervised framework, as we separate the exclusive and common features of the input and use their combination to restore the input, which avoids the requirement

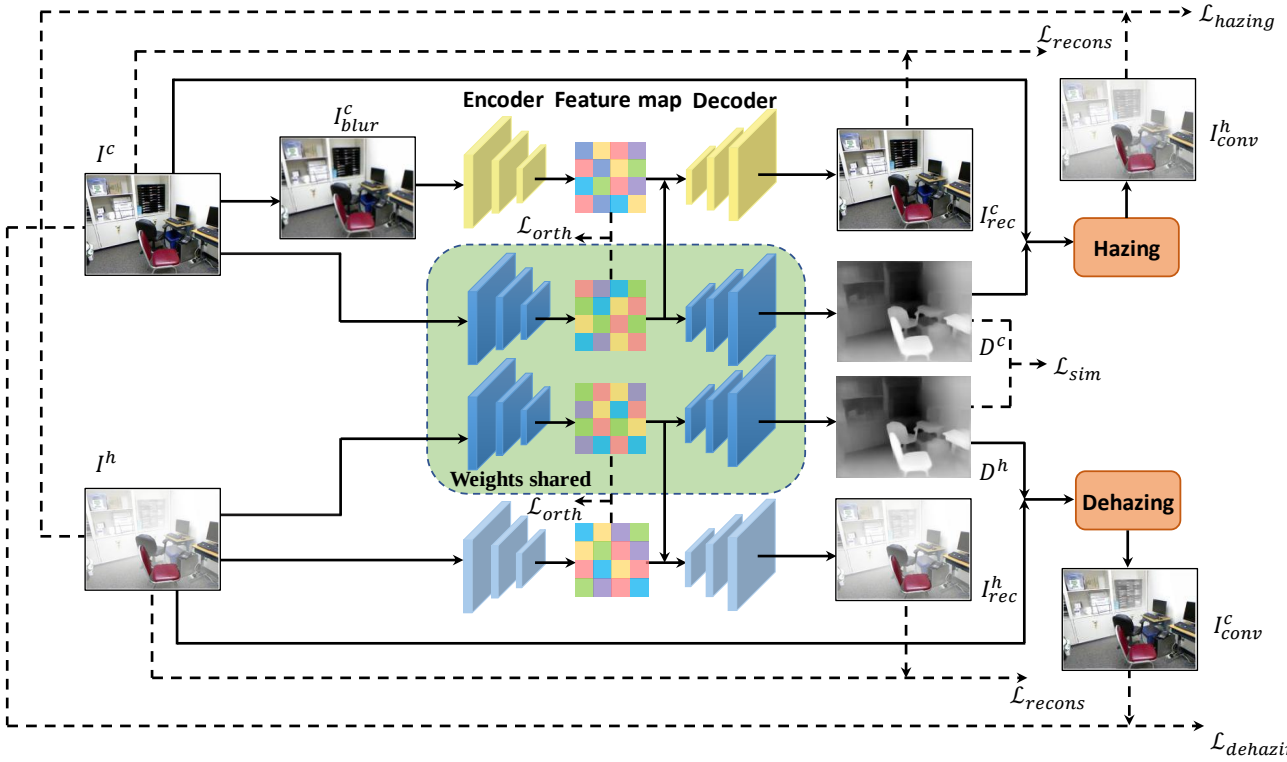

Figure 1. Architecture of our network. The solid line represents the actual data flow of the network, and the dotted line represents the connection between calculated losses. The green section indicates weight sharing, where the networks use the same weight and perform gradient backpropagation. 'Hazing' and 'Dehazing' represent the fogging and defogging modules that perform domain conversion, respectively.

for ground-truth depth. Due to the fact that the common domain encoder-decoder always processes shared depth information from both foggy and non-foggy images, these two encoder-decoders (green part in Figure 1) share the same weight.

3.1.2 Self-Depth Domain Conversion

The self-depth domain conversion part is aimed at the inherent characteristics of non-foggy/foggy images, using the depth map obtained by itself to add/remove fog on the image, thereby achieving cross-domain conversion of the image and further optimizing the depth. The main design route is to use the atmospheric scattering model [20] and its estimated depth map to add fog to the clear image, thereby achieving cross-domain conversion from fog-free to foggy, and calculating losses with foggy images. Similarly, using a foggy image and its estimated depth map, the image can be dehazed to achieve cross-domain conversion from foggy to non-foggy. Due to this process, the image uses its own information to estimate its depth map and participates in the conversion without introducing external information. At the same time, it utilizes the inverse process of hazing and dehazing, so it is called self-depth domain conversion, which improves the effectiveness of unsupervised learning.

3.1.3 Gaussian Blurring

Before sending paired foggy and non-foggy images into the network, we apply a slight Gaussian blurring (with a standard deviation = 2) blurring to clear images when extracting exclusive domain features, distorting some of the image's detailed information for two reasons. Firstly, according to [19], it has been verified that adding deblurring operations to the Masked Autoencoder (MAE) [16] can help the network better recover the detailed information in images. Taking inspiration from this, we also used blurring operation on the clear image I^c before our domain separation. The blurred clear image I^c_{blur} was fed into the exclusive domain feature extraction network (the common domain feature extraction network is still fed in the clear image I^c that was not blurred), and the final reconstruction result was calculated based on the reconstruction loss of the original clear image that was not blurred. This intuitively increased the difficulty of learning the exclusive domain feature extraction network, allowing it to learn more robust features and expose the common domain feature extraction network to more original information, enabling the network to learn more detailed

information. The second is to consider the characteristics of our paired data. The original task of the network was to discover common depth information from non-degraded clear images without fog and degraded images with fog. This encoder-decoder network shares weights, while the network itself is relatively laborious in obtaining accurate depth information from foggy images. The network tends to learn not that fine depths. Under the convergence effect of weight sharing, this will result in the depth estimation of the details in the face of non-foggy clear images being smoothed out, and the depth mapping relationship corresponding to the original details is difficult to learn. When extracting exclusive domain features from clear images suffers from information loss, the network will rely more on the common domain to excavate the original information lost due to blurring during reconstruction. In this way, the learning of these details will be naturally incorporated into the learning process of shared depth information. This compensates for the problem of missing depth details and inaccurate estimation caused by the damage of scene information in foggy images themselves.

The network used in the entire architecture is based on an encoder-decoder structure, with the encoder using ResNet-18 [17] and the decoder using a custom multi-layer CNN network, as referenced in Monodeth2 [12]. We named the network **FODS-Net** (**FO**g **D**omain **S**eparation **Net**), indicating that our network is a depth estimation network designed with domain separation for foggy images.

3.2. Atmospheric scattering and depth-transmittance model

The hazing and dehazing modules in the self-depth domain conversion section use a mutually inverse solution process based on the atmospheric scattering model [20]. Clear images and depth maps can be used to add fog to images, while foggy images and depth maps can be used to remove fog from images. The formula for the atmospheric scattering model is as follows:

$$I(x) = J(x) \cdot t(x) + A(1 - t(x)), \tag{1}$$

where $I(x)$ represents degraded foggy images, $J(x)$ represents clean non-foggy images, A represents global atmospheric light, and $t(x)$ represents the transmittance of unscattered light reaching the camera.

By using this formula, we can perform hazing on clean non-foggy images to obtain a foggy image. According to the atmospheric scattering model, in order to add fog to an image, we need to obtain two key parameters in the model – global atmospheric light A and transmittance $t(x)$.

Based on the characteristics of the NYUv2 dataset we used, the paired foggy and non-foggy images we used were synthesized using the [2] paper. Among them, the global atmospheric light A remains consistent with the definition

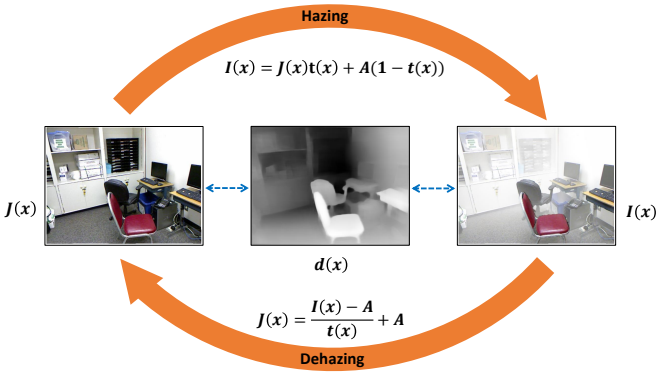

Figure 2. Self-depth domain conversion module. The clear image is represented as J(x), the foggy image is represented as I(x), and d(x) is the depth map, which has a fixed non-linear relationship with the transmittance t(x). A represents global atmospheric light.

of [2], which is [1, 1, 1]. The transmittance $t(x)$ has a non-linear relationship with the depth $d(x)$ of the image as follows:

$$t(x) = e^{-\beta \cdot d(x)}, \tag{2}$$

where β is the atmospheric scattering coefficient. Based on comprehensive experiments, we found that the fog map is most consistent with the synthetic fog map provided in paper [2] when β is 1.75.

By solving the inverse process of this formula, we can obtain the corresponding non-foggy image from the foggy image. The formula is as follows:

$$J(x) = \frac{I(x) - A}{t(x)} + A. \tag{3}$$

Similarly, under the condition of global atmospheric light $A = [1, 1, 1]$, we use image depth $d(x)$ to solve the transmittance $t(x)$. Here, based on multiple experiments, we select β as 1.73, which can achieve the best defogging effect.

After we complete the hazing/dehazing work on the image, we can calculate losses for the obtained foggy/non-foggy images corresponding to the original foggy/clear ones, in order to further refine and evaluate the estimation of image depth.

3.3. Loss function

The overall loss consists of five parts, namely reconstruction loss $\mathcal{L}_{recons}$, similarity loss $\mathcal{L}_{sim}$, orthogonality loss $\mathcal{L}_{orth}$, hazing loss $\mathcal{L}_{hazing}$, and dehazing loss $\mathcal{L}_{dehazing}$. Reconstruction loss mainly measures the loss between the reconstructed image obtained through encoding and decoding structure and the original input image. This part of the

loss includes two parts: MSE and SIMSE, and the formula is as follows:

$$\mathcal{L}_{recons} = \frac{1}{N}\sum_{x}(I^{c}_{rec_x} - I^{c}_{x})^2 + \frac{1}{N^2}\Big(\sum_{x}(I_{rec_X c} - I^{c}_{x})\Big)^2 + \frac{1}{N}\sum_{x}(I^{h}_{rec_x} - I^{h}_{x})^2 + \frac{1}{N^2}\Big(\sum_{x}(I^{h}_{rec_X} - I^{h}_{x})\Big)^2, \tag{4}$$

where I^c represents the original clear input image, I^c_{rec} represents the reconstructed clear image, I^h represents the original foggy input image, I^h_{rec} represents the reconstructed foggy image, x represents the pixel position in the image, and N represents the number of pixels in the image.

Similarity loss refers to the loss calculated based on the similarity between the depth maps estimated for both foggy and non-foggy images, ensuring consistency in using shared depth features to estimate depth maps for both. The formula is as follows:

$$\mathcal{L}_{sim} = \frac{1}{N}\sum_{x}(D^{c}_{x} - D^{h}_{x})^2, \tag{5}$$

where D^c represents the depth map estimated from the clear image, and D^h represents the depth map estimated from the foggy image.

Orthogonality loss is a loss designed to ensure that the exclusive domain features and common domain features obtained in the feature space dimension are orthogonal and unrelated to each other during the domain separation stage. It consists of two parts, namely, directly calculating the orthogonality of the inner product of the feature vectors and calculating the Gram matrix of the features, and then stretching them into one-dimensional feature vectors to calculate the inner product of the vectors as a judgment of orthogonality. The formula is as follows:

$$\mathcal{L}_{orth} = V(v^c_E)\cdot V(v^c_C) + V(v^h_E)\cdot V(v^h_C) + V(g^c_E)\cdot V(g^c_C) + V(g^h_E)\cdot V(g^h_C), \tag{6}$$

where $v^c_E = C(f^c_E)$, $v^c_C = C(f^c_C)$, $v^h_E = C(f^h_E)$, $v^h_C = C(f^h_C)$, $g^c_E = G(f^c_E)$, $g^C_c = G(f^c_C)$, $g^h_E = G(f^h_E)$, and $g^h_C = G(f^h_C.)$. Within these equations, f^c_E presents the exclusive domain feature of clear images, f^c_C is the common domain feature of clear images, f^h_E denotes the exclusive domain feature of foggy images, and f^h_C is the common domain feature of foggy images. $C(\cdot)$ represents the 1×1 convolution operation used for feature dimensionality reduction, $G(\cdot)$ represents the calculation of Gram matrix, $V(\cdot)$ represents flattening feature vectors to one dimension, and $\cdot$ represents dot product of vectors.

Hazing loss and dehazing loss are a set of co-existing losses. The hazing loss is mainly calculated by adding fog to the fog-free image using the atmospheric scattering model after obtaining the depth map, and calculating the loss on the foggy image paired with the original clear image. The dehazing loss, on the other hand, is the opposite. It is mainly calculated by using the atmospheric scattering model to dehaze the foggy image after obtaining the depth map, and calculating the loss with paired non-foggy images of the foggy image. The formulas are as follows:

$$\mathcal{L}_{hazing} = \frac{1}{N}\sum_{x}(I^{h}_{conv_x} - I^{h}_{x})^2, \tag{7}$$

$$\mathcal{L}_{dehazing} = \frac{1}{N}\sum_{x}(I^{c}_{conv_x} - I^{c}_{x})^2, \tag{8}$$

where I^h_{conv} represents the hazed image obtained from domain conversion, and I^c_{conv} represents the dehazed image obtained from domain conversion. Finally, the total loss is expressed as:

$$\mathcal{L}_{total} = \lambda_1\mathcal{L}_{recons} + \lambda_2\mathcal{L}_{sim} + \lambda_3\mathcal{L}_{orth} + \lambda_4\mathcal{L}_{hazing} + \lambda_5\mathcal{L}_{dehazing}, \tag{9}$$

4. Experiments

4.1. Datasets

We mainly used the NYUv2 dataset [37] to train our network. The NYUv2 dataset is a depth dataset provided by New York University, collected using Kinect RGB-D cameras for 464 indoor scenes. The dataset contains 1449 pairs of raw RGB images and their corresponding depth maps. The original image has a resolution of 480×640. We used the official 654 images with ground truth maps in the test set.

Meanwhile, our unsupervised model training process requires paired foggy and non-foggy images. Here, we adopted the work of [2], which synthesized corresponding foggy versions of images based on the NYUv2 dataset and named them the D-HAZY dataset. In addition to the NYUv2 dataset, we also used the SUN RGB-D dataset [38] for additional quantitative analysis and evaluation of the model's generalization ability. The SUN RGB-D dataset is also a dataset collected by depth cameras for indoor scenes, with 10335 RGB-D images. We used the official 5050 images for its testing set.

4.2. Implementation details

Our experiment was run on a computer equipped with NVIDIA RTX 4090 GPU, with a single card of 24GB graphics memory. The optimizer uses the Adam optimizer, with an initial learning rate set to 1e-4 and multiplied by 0.1 every 20 epochs. The program ran a total of 100 epochs.

Table 1. Quantitative comparison to state-of-the-art methods on the NYUv2 dataset (clear images).

	Lower is better ↓				Higher is better ↑		
Method	Abs Rel.	Sq Rel.	RMSE (lin)	RMSE (log)	δ_1	δ_2	δ_3
Zhao *et al.* [48]	0.189	-	0.686	0.079	0.701	0.912	0.987
Monodepth2 [12]	0.165	-	0.686	0.070	0.765	0.937	0.983
DCL-depth [15]	0.137	-	0.534	0.059	0.820	0.958	0.990
DORN [8]	0.115	-	0.509	0.051	0.828	0.965	0.992
VNL [42]	0.108	-	0.416	0.048	0.875	0.976	0.994
BTS [22]	0.110	0.066	0.392	0.047	0.885	0.978	0.994
PWA [23]	0.105	-	0.374	0.045	0.892	0.985	0.997
TransDepth [41]	0.106	-	0.365	0.045	0.900	0.983	0.996
AdaBins [3]	0.103	-	0.364	0.044	0.903	0.984	0.997
NeWCRFs [43]	0.095	0.045	0.334	0.041	0.922	**0.992**	**0.998**
IEBins [36]	0.087	**0.040**	0.314	0.038	**0.936**	**0.992**	**0.998**
Ours	**0.085**	**0.040**	**0.311**	**0.037**	0.928	0.991	**0.998**

Table 2. Quantitative comparison to state-of-the-art methods on the NYUv2 dataset (foggy images).

	Lower is better ↓				Higher is better ↑		
Method	Abs Rel.	Sq Rel.	RMSE (lin)	RMSE (log)	δ_1	δ_2	δ_3
BTS [22]	5.689	16.734	2.592	0.943	0.244	0.441	0.596
Zhao *et al.* [48]	5.546	17.184	2.580	0.908	0.264	0.476	0.634
VNL [42]	2.260	4.532	2.056	0.721	0.306	0.567	0.740
AdaBins [3]	2.964	5.418	1.861	0.775	0.343	0.588	0.747
TransDepth [41]	2.600	4.410	1.639	0.683	0.394	0.641	0.783
NeWCRFs [43]	1.129	1.344	1.830	0.644	0.317	0.588	0.764
IEBins [36]	1.680	1.945	1.455	0.575	0.431	0.689	0.831
Ours	**0.666**	**0.356**	**0.751**	**0.438**	**0.728**	**0.885**	**0.925**

The encoders are all ResNet-18 (He K et al., 2022) with the same structure. The image was resized to 512 $\times$ 512 before being sent to the network.

4.3. Quantitative analysis

We conducted quantitative analysis experiments on the NYUv2 dataset and the SUN RGB-D dataset. The quantitative indicators used in the experiment are commonly used in the industry, mainly including: the absolute relative error (Abs Rel.) index, squared relative error (Sq Rel.) index, linear root mean squared error (RMSE (lin)) index, log root mean squared error (RMSE (log)) index, and the threshold accuracy δ_n (% of pixels $s.t.$ $max(d_i/\hat{d}_l, \hat{d}_l/d_i) < 1.25^n, n = \{1, 2, 3\}$, where d_i denotes the ground truth depth at pixel i, $\hat{d}_l$ denotes the predicted depth at pixel i. The maximum scene depth is limited to 10 m.

From Table 1, Table 2 and Table 3, it can be seen that our method achieved the best overall performance among multiple advanced methods on the NYUv2 dataset. Specifically, Table 3 demonstrates the quantitative results of mainstream methods after fine-tuning on the NYUv2 dataset (foggy images).For clear haze-free images, our method achieved thse best results on 5 indicators, with the other 2 indicators ranking second only to the best one, demonstrating the best overall performance. For foggy images, our method achieved the best results in all 7 metrics and far exceeded other methods, surpassing the NeWCRFs [43] algorithm by 41% in Abs Rel metrics and 73.51% in Sq Rel metrics. We also exceeded the IEBins [36] algorithm by 70.4%, 23.83%, 68.91%, 28.45%, and 11.31% in RMSE (lin), RMSE (log), δ_1, δ_2, and δ_3 metrics, respectively. This once again demonstrates the additional effectiveness of our monocular depth estimation algorithm for foggy images.

According to the results shown in table 4, our method is significantly superior to other algorithms on the SUN RGB-D dataset. This further demonstrates the effectiveness and robustness of our algorithm.

4.4. Qualitative analysis

In the qualitative analysis section, we first demonstrate the input and output images involved in the training process of the network. The network inputs both clear images and corresponding foggy images, and the clear images undergo a Gaussian blurring operation when extracting exclusive domain features. The results in Figure 3 show that consistency is ensured when extracting common domain features,

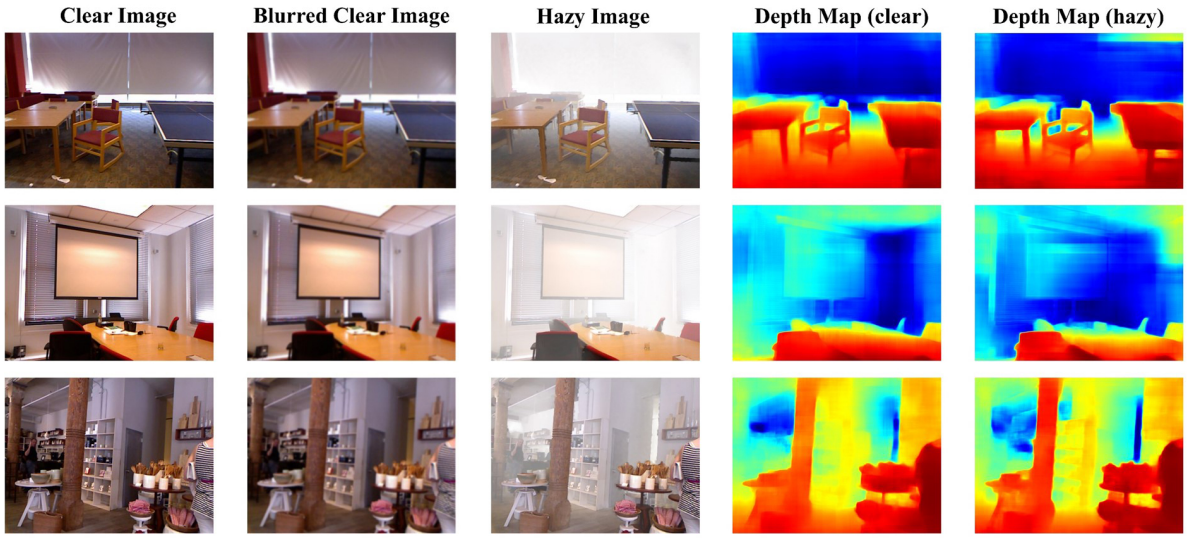

Figure 3. Input and output involved in training on the NYUv2 dataset.

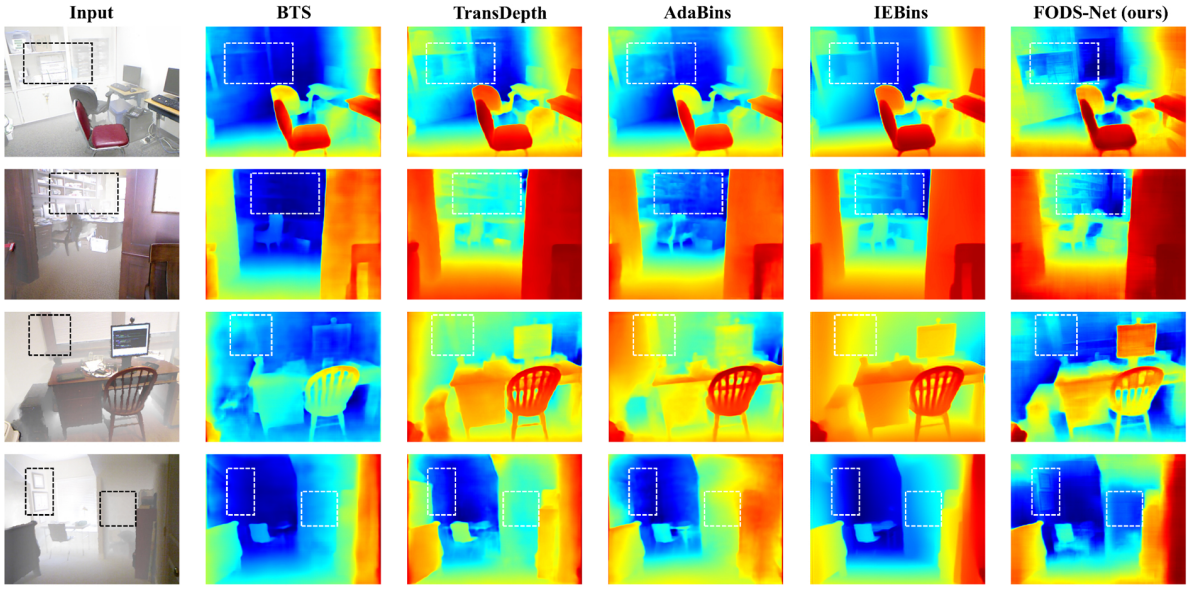

Figure 4. Qualitative comparison on the NYUv2 dataset.

namely depth maps, for both clear and foggy images. This demonstrates the reliability of our network's designed structure.

We conducted a qualitative comparative analysis with advanced methods on the NYUv2 dataset, and the results are shown in Figure 4. From the graph, it can be seen that our proposed method can estimate the most accurate and detailed depth map among these methods. The dashed box in the figure (such as the box in the first row, the bookshelf in the second row, the wooden board in the third row, and the decorative painting in the fourth row) shows the depth region with distinct detail features estimated by our method for foggy images, which can reflect more information about the background area of the original image without losing accuracy. However, the BTS [22] model, TransDepth model, and AdaBins [41] model all suffer from inaccurate depth estimation. Although the IEBins [36] algorithm can accurately estimate the overall depth, its accuracy in depicting details is not as good as our model. Meanwhile, for some areas covered by fog, such as the dashed box area on the right side of the fourth row images in Figure 4, other methods will estimate it as a smooth solid depth area, while our method can estimate the true depth of the void area.

Table 3. Quantitative comparison of state-of-the-art methods on NYUv2 dataset (foggy images) after fine-tuning.

	Lower is better ↓				Higher is better ↑		
Method	Abs Rel.	Sq Rel.	RMSE (lin)	RMSE (log)	δ_1	δ_2	δ_3
AdaBins [3]	1.457	2.423	1.609	0.733	0.427	0.587	0.754
TransDepth [41]	1.413	2.014	1.577	0.675	0.478	0.694	0.801
NeWCRFs [43]	0.686	0.676	1.123	0.543	0.581	0.741	0.863
IEBins [36]	0.679	0.573	0.824	0.512	0.712	0.819	0.907
Ours	**0.666**	**0.356**	**0.751**	**0.438**	**0.728**	**0.885**	**0.925**

Table 4. Quantitative comparison to state-of-the-art methods on the SUN RGB-D dataset

	Lower is better ↓			Higher is better ↑		
Method	Abs Rel.	RMSE (lin)	RMSE (log)	δ_1	δ_2	δ_3
VNL [42]	0.183	0.541	0.082	0.696	0.912	0.973
BTS [22]	0.172	0.515	0.075	0.740	0.933	0.980
AdaBins [3]	0.159	0.476	0.068	0.771	0.944	0.983
LocalBins [4]	0.156	0.470	0.067	0.777	0.949	0.985
Ours	**0.146**	**0.451**	**0.061**	**0.792**	**0.963**	**0.990**

Table 5. Ablation study of the proposed FODS-Net on the NYUv2 dataset (foggy images).

	Lower is better ↓				Higher is better ↑		
Method	Abs Rel.	Sq Rel.	RMSE (lin)	RMSE (log)	δ_1	δ_2	δ_3
Basic	0.804	0.487	0.861	0.523	0.575	0.782	0.889
Basic + blur	0.753	0.421	0.825	0.505	0.619	0.793	0.894
Basic + $\mathcal{L}_{orth}$	0.721	0.395	0.801	0.482	0.641	0.811	0.905
Basic + blur + $\mathcal{L}_{orth}$	0.695	0.373	0.779	0.466	0.673	0.840	0.917
Basic + blur + $\mathcal{L}_{orth}$ + H./DH.	**0.666**	**0.356**	**0.751**	**0.438**	**0.728**	**0.885**	**0.925**

Table 6. Ablation study of the proposed FODS-Net on the SUN RGB-D dataset.

	Lower is better ↓			Higher is better ↑		
Method	Abs Rel.	RMSE (lin)	RMSE (log)	δ_1	δ_2	δ_3
Basic	0.189	0.572	0.087	0.621	0.895	0.972
Basic + blur	0.181	0.524	0.084	0.667	0.907	0.975
Basic + $\mathcal{L}_{orth}$	0.176	0.503	0.080	0.682	0.919	0.977
Basic + blur + $\mathcal{L}_{orth}$	0.169	0.481	0.075	0.711	0.930	0.982
Basic + blur + $\mathcal{L}_{orth}$ + H./DH.	**0.146**	**0.451**	**0.061**	**0.792**	**0.963**	**0.990**

4.5. Ablation study

4.5.1 Common features and exclusive features

To demonstrate the effectiveness of our network's domain separation design, we plotted the intermediate feature maps of common and exclusive domains extracted during the domain separation process in Figure 5. From the figure, it can be seen that for the input clear and foggy images, the common domain features focus on the relative position and depth relationship of objects in the image. The extracted features have clear and layered large continuous pixels (pixels in the same depth area are mostly the same), which is in line with our intuitive understanding of depth. The features extracted from the exclusive domain focus more on the original details and scene information of the image itself. Although there are slight differences in the intermediate features between clear images and foggy images, they are more inclined towards the original scene representation of the image, such as lighting, texture, details, etc. The above analysis verifies the effectiveness of our domain separation design.

4.5.2 Self-depth domain conversion

Figure 6 shows the results of domain conversion using network's self-estimated depth maps. For the original clear images and foggy images, we use self-depth and atmospheric

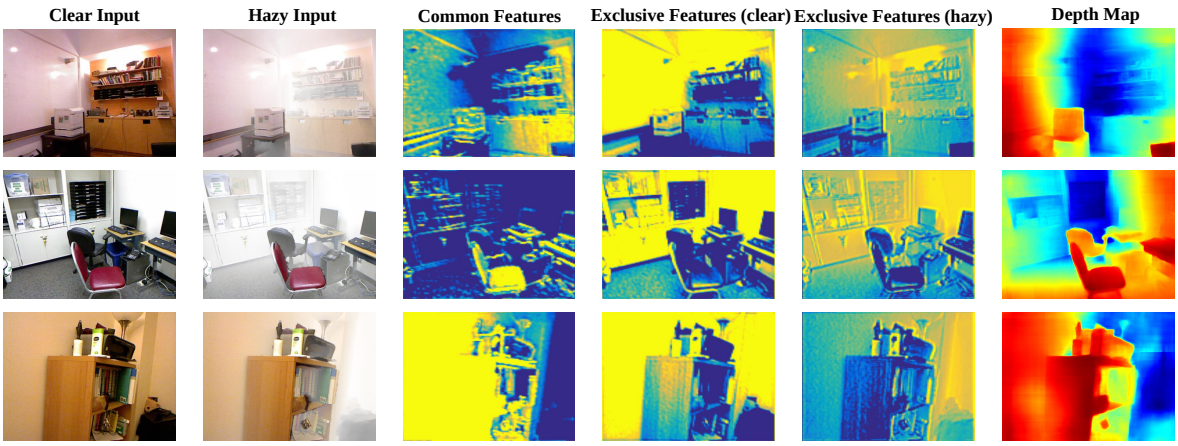

Figure 5. Domain separation demonstration on the NYUv2 dataset.

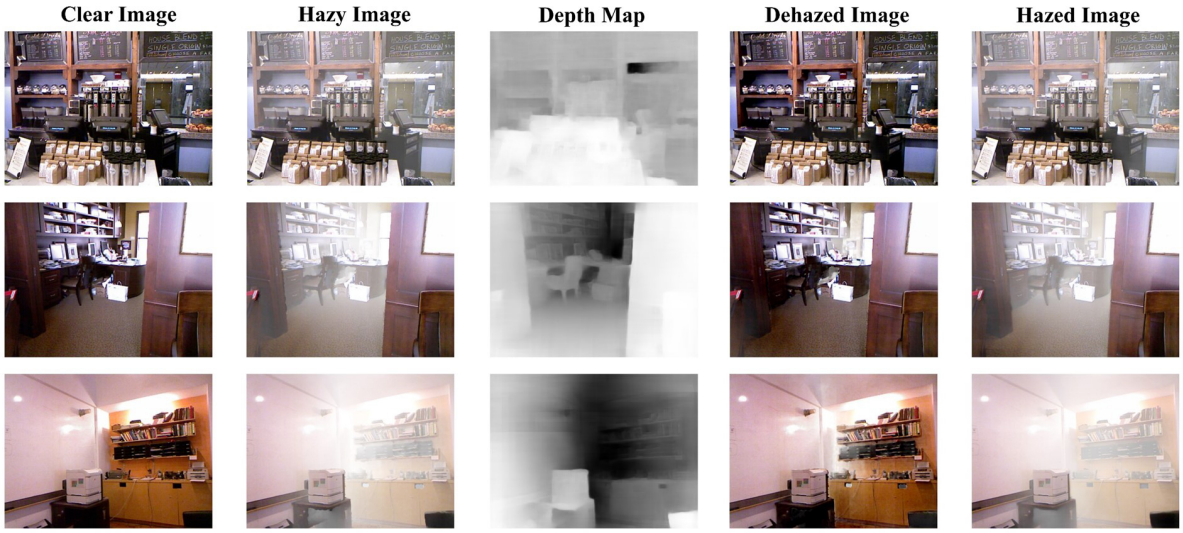

Figure 6. Self-depth domain conversion demonstration on the NYUv2 dataset.

scattering models to calculate their corresponding hazing and dehazing maps. From the graph, it can be seen that our self-estimated depth map first has high accuracy and details. The resulting dehazing and hazing maps have high consistency with the original clear and foggy images, with only slight differences. And this part of the difference mainly comes from the inherent information loss when the image is degraded by fog, which cannot restore the information of the original pixel space during the dehazing process, and does not significantly affect the performance of our network.

4.5.3 Analysis of the effectiveness of modules

We mainly conducted ablation experimental analysis on the blurring operation, dehazing and hazing losses, and orthogonality loss (seen in table 5 and table 6), which are represented by blur, H./DH., and L_{orth}, respectively. Basic represents the basic network of FODS-Net that does not include these three parts.

(a) **Blurring operation**: The blurring operation applies Gaussian blurring to the original input clear image when extracting exclusive domain features, thereby distorting its information to a certain extent. This allows the network model to learn more detailed information when extracting depth information from the common domain, making the obtained depth map more refined. Moreover, it increases the difficulty of learning in the image reconstruction process, making it more robust to learn features. From the results in Table 4 and 5, it can be seen that blurring significantly improves the estimation of depth maps.

(b) **Hazing and dehazing loss**: The hazing and dehazing losses are losses constructed based on the self-depth domain conversion module, mainly measuring the approximation between the network's domain conversion results of the

image using the self-estimated depth and the target domain. This loss can indirectly reflect the quality of depth map estimation, as it is directly related to the results of domain conversion. From the results in Table 4 and 5, it can be seen that the addition of hazing and dehazing losses has a great improvement effect on all indicators.

(c) **Orthogonality loss**: Orthogonality loss is mainly used to separate the features of an image, thereby decomposing the image information into common and exclusive domains, extracting common depth information as the final estimation result. The orthogonality loss calculates the orthogonality between the decomposed feature vectors, ensuring that the results of domain separation are complementary rather than intersecting. From the results in Table 4 and 5, it can be seen that the addition of orthogonality loss provides a clear directional division in the training process, allowing the network to better focus on extracting pure depth information.

5. Limition

Currently, mainstream public outdoor datasets for depth estimation research, such as KITTI [10], do not provide paired foggy and non-foggy versions. While the Cityscapes [7] dataset offers paired foggy and non-foggy data, the depth information is inaccurate and suffers from pixel depth loss, which is not generally used. Explicitly, the set of foggy and non-foggy image pairs is the beginning point of our framework, so we do not reveal the experimental tests under outdoor scenes. Our model is limited under indoor scenes and we will explore more possibilities for outdoor utilization in the future.

6. Conclusion

This article proposes an innovative unsupervised monocular depth estimation algorithm for foggy images with domain separation and self-depth domain conversion. By performing domain separation design on paired foggy and non-foggy images, we extracted common domain information (depth) and exclusive domain features (lighting, texture, color, etc.), achieving efficient depth estimation. By utilizing orthogonality loss to ensure the complementarity of features and applying blurring operations to fog-free images to increase learning difficulty, the model's ability to learn detailed depth information and the robustness of feature extraction are improved. The self-depth domain conversion part is based on the atmospheric scattering model, combined with the self-estimated depth map to achieve the conversion of images between foggy and non-foggy domains, further optimizing the accuracy of depth estimation. The experimental results show that our algorithm performs well in the depth estimation task of foggy images, effectively improving the accuracy and stability of depth estimation. Future work will be around further optimizing our network structure and loss function designs to cope with more complex practical application scenarios.

References

[1] F. Aleotti, F. Tosi, M. Poggi, and S. Mattoccia. Generative adversarial networks for unsupervised monocular depth prediction. In L. Leal-Taixé and S. Roth, editors, *Computer Vision - ECCV 2018 Workshops - Munich, Germany, September 8-14, 2018, Proceedings, Part I*, volume 11129 of *Lecture Notes in Computer Science*, pages 337–354. Springer, 2018. 2

[2] C. Ancuti, C. O. Ancuti, and C. D. Vleeschouwer. D-HAZY: A dataset to evaluate quantitatively dehazing algorithms. In *2016 IEEE International Conference on Image Processing, ICIP 2016, Phoenix, AZ, USA, September 25-28, 2016*, pages 2226–2230. IEEE, 2016. 5, 6

[3] S. F. Bhat, I. Alhashim, and P. Wonka. Adabins: Depth estimation using adaptive bins. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 4009–4018. Computer Vision Foundation / IEEE, 2021. 7, 9

[4] S. F. Bhat, I. Alhashim, and P. Wonka. Localbins: Improving depth estimation by learning local distributions. In S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part I*, volume 13661 of *Lecture Notes in Computer Science*, pages 480–496. Springer, 2022. 9

[5] P. Chen, A. H. Liu, Y. Liu, and Y. F. Wang. Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2624–2632. Computer Vision Foundation / IEEE, 2019. 1, 2, 3

[6] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 3213–3223. IEEE Computer Society, 2016. 1

[7] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016. 11

[8] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao. Deep ordinal regression network for monocular depth estimation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 2002–2011. Computer Vision Foundation / IEEE Computer Society, 2018. 7

[9] S. Gasperini, N. Morbitzer, H. Jung, N. Navab, and F. Tombari. Robust monocular depth estimation under challenging conditions. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 8143–8152. IEEE, 2023. 3

[10] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 11

[11] C. Godard, O. M. Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6602–6611. IEEE Computer Society, 2017. 2

[12] C. Godard, O. M. Aodha, M. Firman, and G. J. Brostow. Digging into self-supervised monocular depth estimation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 3827–3837. IEEE, 2019. 2, 5, 7

[13] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014. 2

[14] X. Guo, H. Zhao, S. Shao, X. Li, and B. Zhang. F2depth: Self-supervised indoor monocular depth estimation via optical flow consistency and feature map synthesis. *Eng. Appl. Artif. Intell.*, 133:108391, 2024. 3

[15] C. Han, C. Lv, Q. Kou, H. Jiang, and D. Cheng. Dcl-depth: monocular depth estimation network based on iam and depth consistency loss. *Multimedia Tools and Applications*, 2024. 3, 7

[16] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. B. Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 15979–15988. IEEE, 2022. 4

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 5

[18] T. Hui. Rm-depth: Unsupervised learning of recurrent monocular depth in dynamic scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 1665–1674. IEEE, 2022. 2, 3

[19] Q. Kang, J. Gao, K. Li, and Q. Lao. Deblurring masked autoencoder is better recipe for ultrasound image recognition. In H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T. F. Syeda-Mahmood, and R. H. Taylor, editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2023 - 26th International Conference, Vancouver, BC, Canada, October 8-12, 2023, Proceedings, Part I*, volume 14220 of *Lecture Notes in Computer Science*, pages 352–362. Springer, 2023. 4

[20] H. Koschmieder. Theorie der horizontalen sichtweite. *Beitrage zur Physik der freien Atmosphäre*, 12:33–55, 1924. 4, 5

[21] B. Lee, K. Lee, J. Oh, and I. S. Kweon. Cnn-based simultaneous dehazing and depth estimation. In *2020 IEEE International Conference on Robotics and Automation, ICRA 2020, Paris, France, May 31 - August 31, 2020*, pages 9722–9728. IEEE, 2020. 1

[22] J. H. Lee, M. Han, D. W. Ko, and I. H. Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *CoRR*, abs/1907.10326, 2019. 7, 8, 9

[23] S. Lee, J. Lee, B. Kim, E. Yi, and J. Kim. Patch-wise attention network for monocular depth estimation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 1873–1881. AAAI Press, 2021. 7

[24] Z. Liang, T. Fang, Y. Hu, and Y. Wang. Sparse depth densification for monocular depth estimation. *Multim. Tools Appl.*, 83(5):14821–14838, 2024. 3

[25] Z. Liang, T. Fang, Y. Hu, and Y. Wang. Sparse depth densification for monocular depth estimation. *Multim. Tools Appl.*, 83(5):14821–14838, 2024. 3

[26] M. Lin, G. Li, and Y. Hao. Bridging local and global representations for self-supervised monocular depth estimation. *Eng. Appl. Artif. Intell.*, 133:108277, 2024. 2, 3

[27] L. Liu, X. Song, M. Wang, Y. Liu, and L. Zhang. Self-supervised monocular depth estimation for all day images using domain separation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 12717–12726. IEEE, 2021. 3

[28] A. Masoumian, H. A. Rashwan, S. Abdulwahab, J. Cristiano, M. S. Asif, and D. Puig. Gcndepth: Self-supervised monocular depth estimation based on graph convolutional network. *Neurocomputing*, 517:81–92, 2023. 2, 3

[29] V. Patil, C. Sakaridis, A. Liniger, and L. V. Gool. P3depth: Monocular depth estimation with a piecewise planarity prior. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 1600–1611. IEEE, 2022. 1

[30] L. Piccinelli, Y. Yang, C. Sakaridis, M. Segù, S. Li, L. V. Gool, and F. Yu. Unidepth: Universal monocular metric depth estimation. *CoRR*, abs/2403.18913, 2024. 3

[31] A. Pilzer, S. Lathuilière, N. Sebe, and E. Ricci. Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 9768–9777. Computer Vision Foundation / IEEE, 2019. 3

[32] M. Poggi, F. Aleotti, F. Tosi, and S. Mattoccia. Towards real-time unsupervised monocular depth estimation on CPU. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2018, Madrid, Spain, October 1-5, 2018*, pages 5848–5854. IEEE, 2018. 2

[33] M. Poggi, F. Tosi, and S. Mattoccia. Learning monocular depth estimation with unsupervised trinocular assumptions. In *2018 International Conference on 3D Vision, 3DV 2018,*

Verona, Italy, September 5-8, 2018, pages 324–333. IEEE Computer Society, 2018. 2

[34] W. Ren, L. Wang, Y. Piao, M. Zhang, H. Lu, and T. Liu. Adaptive co-teaching for unsupervised monocular depth estimation. In S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part I*, volume 13661 of *Lecture Notes in Computer Science*, pages 89–105. Springer, 2022. 2, 3

[35] S. Saxena, C. Herrmann, J. Hur, A. Kar, M. Norouzi, D. Sun, and D. J. Fleet. The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 2, 3

[36] S. Shao, Z. Pei, X. Wu, Z. Liu, W. Chen, and Z. Li. Iebins: Iterative elastic bins for monocular depth estimation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 3, 7, 8, 9

[37] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from RGBD images. In A. W. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V*, volume 7576 of *Lecture Notes in Computer Science*, pages 746–760. Springer, 2012. 6

[38] S. Song, S. P. Lichtenberg, and J. Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 567–576. IEEE Computer Society, 2015. 6

[39] F. Tosi, F. Aleotti, M. Poggi, and S. Mattoccia. Learning monocular depth estimation infusing traditional stereo knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 9799–9809. Computer Vision Foundation / IEEE, 2019. 1

[40] X. Wang, J. Sun, H. Qin, Y. Yuan, J. Yu, Y. Su, and Z. Sun. Accurate unsupervised monocular depth estimation for ill-posed region. In *Frontiers of Physics*, 2023. 2, 3

[41] G. Yang, H. Tang, M. Ding, N. Sebe, and E. Ricci. Transformer-based attention networks for continuous pixel-wise prediction. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 16249–16259. IEEE, 2021. 7, 8, 9

[42] W. Yin, Y. Liu, C. Shen, and Y. Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 5683–5692. IEEE, 2019. 7, 9

[43] W. Yuan, X. Gu, Z. Dai, S. Zhu, and P. Tan. Neural window fully-connected crfs for monocular depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 3906–3915. IEEE, 2022. 7, 9

[44] H. Zhan, R. Garg, C. S. Weerasekera, K. Li, H. Agarwal, and I. D. Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 340–349. Computer Vision Foundation / IEEE Computer Society, 2018. 2

[45] Y. Zhang, M. Gong, J. Li, M. Zhang, F. Jiang, and H. Zhao. Self-supervised monocular depth estimation with multiscale perception. *IEEE Trans. Image Process.*, 31:3251–3266, 2022. 2, 3

[46] C. Zhao, Y. Tang, and Q. Sun. Unsupervised monocular depth estimation in highly complex environments. *IEEE Trans. Emerg. Top. Comput. Intell.*, 6(5):1237–1246, 2022. 3

[47] C. Zhao, Y. Zhang, M. Poggi, F. Tosi, X. Guo, Z. Zhu, G. Huang, Y. Tang, and S. Mattoccia. Monovit: Self-supervised monocular depth estimation with a vision transformer. In *International Conference on 3D Vision, 3DV 2022, Prague, Czech Republic, September 12-16, 2022*, pages 668–678. IEEE, 2022. 2, 3

[48] W. Zhao, S. Liu, Y. Shu, and Y. Liu. Towards better generalization: Joint depth-pose learning without posenet. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9148–9158. Computer Vision Foundation / IEEE, 2020. 7

[49] J. Zhu, L. Liu, Y. Liu, W. Li, F. Wen, and H. Zhang. Fg-depth: Flow-guided unsupervised monocular depth estimation. In *IEEE International Conference on Robotics and Automation, ICRA 2023, London, UK, May 29 - June 2, 2023*, pages 4924–4930. IEEE, 2023. 3