URCDC-Depth: Uncertainty Rectified Cross-Distillation with CutFlip for Monocular Depth Estimation

Shuwei Shao, Zhongcai Pei, Weihai Chen*, Ran Li, Zhong Liu and Zhengguo Li, *Fellow, IEEE*

***Abstract*—This work aims to estimate a high-quality depth map from a single RGB image. Due to the lack of depth clues, making full use of the long-range correlation and local information is critical for accurate depth estimation. To this end, we introduce an uncertainty rectified cross-distillation between the Transformer and convolutional neural network (CNN) to achieve a comprehensive depth estimator. Specifically, we utilize the depth estimates from the Transformer branch and CNN branch as pseudo labels to teach each other. At the same time, the pixel-wise depth uncertainty is modeled to mitigate the negative impact of noisy pseudo labels. To avoid the large capacity gap induced by the strong Transformer branch deteriorating the cross-distillation, we transfer the feature maps from the Transformer to the CNN and develop coupling units to assist the weak CNN branch in leveraging the transferred features. Furthermore, we introduce CutFlip, a surprisingly simple yet highly effective data augmentation technique, which forces the model to focus on more valuable depth reasoning clues apart from the vertical image position. Extensive experiments demonstrate that our model, termed URCDC-Depth, exceeds in performance previous state-of-the-art approaches on the KITTI, NYU-Depth-v2 and SUN RGB-D datasets, with no additional computational burden in the evaluation phase. The source code will be publicly available upon acceptance.**

***Index Terms*—Monocular depth estimation, cross-distillation, Transformer, data augmentation.**

I. Introduction

Monocular depth estimation is a fundamental research topic in the computer vision community, with applications ranging from scene understanding and 3D reconstruction to image dehazing [1]–[3]. Benefiting from advances in convolutional neural networks (CNNs) [4]–[6], recent studies [7], [8] have achieved outstanding performance. Due to the lack of depth cues, fully exploiting the long-range correlation (i.e., inter-object distance relationship) and the local information (i.e., intra-object consistency), is crucial for accurate depth estimation [9]. However, the convolution operator with a limited receptive field has difficulty capturing the long-range correlation. This becomes a potential bottleneck of current CNN-based depth estimation methods [8].

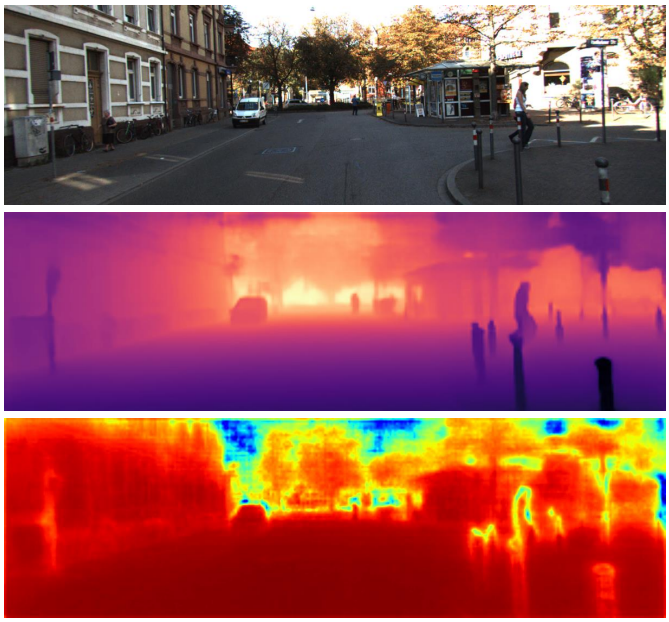

Fig. 1. **Illustration of the depth map and uncertainty map generated by the proposed URCDC-Depth**. *Top: input image; Middle: depth map; Bottom: pixel-wise depth uncertainty (red: low uncertainty; yellow/blue: high/highest uncertainty).*

There is extensive work dedicated to alleviating the above CNN limitation, which can be broadly categorized into two groups: manipulating the convolution operation and introducing the attention mechanism [10]. The former uses techniques such as atrous spatial pyramid pooling [11] and coarse-to-fine fusion [12] to enhance the efficacy of the convolution operator. The latter incorporates an attention module to establish the global interactions among pixels in the feature map [13], [14]. In addition, several general approaches leverage both of these strategies [8], [15]. Despite the considerable improvements in performance, limitations remain.

Recently, visual Transformer has been presented to be a promising alternative to the CNN [16]–[18]. Building upon the attention mechanism, the Transformer with a global receptive field is more proficient in capturing the long-range correlation. Nevertheless, local feature details are prone to be ignored by it, as observed in [18] and [19]. This inevitably results in

This work was supported in part by the National Natural Science Foundation of China under Grants 62333023 and U1909215, in part by the A*STAR Singapore underMTC Programmatic Funds under Grant M23L7b0021, in part by the Key Research and Development Program of Zhejiang Province under Grant 2021C03050, in part by the Scientific Research Project of Agriculture and Social Development of Hangzhou under Grant 20212013B11, and in part by the National Natural Science Foundation of China under Grants 61620106012 and 61573048.

Shuwei Shao, Zhongcai Pei, Weihai Chen, Ran Li and Zhong Liu are with the School of Automation Science and Electrical Engineering, Beihang University, Beijing, China. (email: swshao@buaa.edu.cn, peizc@buaa.edu.cn, whchen@buaa.edu.cn, rnlee1998@buaa.edu.cn and liuzhong@buaa.edu.cn)

Zhengguo Li is with the SRO department, Institute for Infocomm Research, 1 Fusionopolis Way, Singapore. (ezgli@i2r.a-star.edu.sg)

*(corresponding author: Weihai Chen.)

unsatisfactory performance. Several depth estimation methods try to overcome the drawback of Transformer by integrating an additional CNN branch [19], [20]. However, these frameworks also rely on the CNN branch in the evaluation phase which increases the computational cost at inference time.

In this work, we introduce a novel monocular depth estimation framework, termed URCDC-Depth, which boosts model performance by integrating the strengths of Transformer and CNN through **cross-distillation**. The core idea lies in that the Transformer branch establishes the long-range correlation, while the CNN branch focuses on the local information, and both global representations and local features are pivotal for accurate depth estimation. The cross-distillation between these two branches helps learn a comprehensive depth estimator with both properties without adding computational burden during evaluation. In addition, when compared to unidirectional CNN-to-Transformer distillation, cross-distillation enjoys the benefit of improving the CNN branch accuracy, allowing it to be a superior teacher to guide the Transformer branch.

To be specific, we leverage the derived depth estimates from these two branches as pseudo labels to teach their counterparts. Unfortunately, pseudo labels may contain heavy noise, for instance, induced by the inherent drawbacks of Transformer and CNN. This is harmful for cross-distillation. Hence, it is critical to be aware of possible errors in depth prediction. The **depth uncertainty** indicates the reliability and potential errors of depth prediction [21], [22], which is estimated to be associated with the depth map and used to alleviate the negative impact of noisy pseudo labels in cross-distillation. In addition, we transfer the feature maps from the Transformer to the CNN to bridge the large capacity gap caused by the strong Transformer branch. The coupling units are developed to assist the weak CNN branch in leveraging the transferred feature maps, which contribute to enhancing the cross-distillation as well. An illustration of the predicted depth and uncertainty maps is shown in Fig. 1.

Furthermore, we train URCDC-Depth with a very simple yet effective data augmentation technique termed **CutFlip**. It is based on the observation that the monocular depth estimation model relies heavily on the vertical image position to estimate depth, while other clues such as apparent sizes are ignored which deteriorates the model generalization ability [23]. Generally, the feature of the vertical image position is that the closer the projection on the image is to the lower boundary, the smaller the depth of the scene point. The traditional training mechanism allows the clue of vertical image position to exist in almost all training samples. In contrast, other cues are much less numerous. To resolve this, we vertically cut the training sample into upper and lower parts and flip these two parts along the vertical direction with a certain probability thereby weakening the relationship between depth and vertical image position. In such cases, the accuracy of the depth estimate is significantly improved.

To summarize, the main contributions of this work are as follows:

- We introduce a novel monocular depth estimation framework equipped with uncertainty rectified cross-distillation to exploit the long-range correlation and local information. As a significant benefit from the cross-distillation paradigm, the framework has no additional computational burden in the evaluation phase.
- We propose a surprisingly simple yet highly effective data augmentation technique to enforce the model focusing on more valuable depth reasoning cues and not just the vertical image position.
- Detailed experiments and analysis indicate the efficacy of our developed components. The proposed URCDC-Depth achieves state-of-the-art performance on the KITTI [24], NYU-Depth-v2 [25] and SUN RGB-D [26] datasets.

The rest of this paper is organized as follows: A brief review of related work is presented in Section II. The proposed method is detailed in Section III. Experimental results on standard benchmarks are reported in Section IV. The conclusions are summarized in Section V.

II. Related work

A. Monocular depth estimation

Monocular depth estimation attempts to derive the dense depth map from a single image. As one of the seminal works in the field, Saxena *et al.* [9] used the Markov random field to estimate depth. Subsequently, numerous studies have achieved consistent improvements [27]–[31], drawing on the well generalized CNN features across diverse tasks. More recently, Lee *et al.* [7] proposed multi-scale local planar layers to provide more effective guidance of encoded features to the desired depth prediction. Yin *et al.* [32] suggested the virtual surface normal and enforced 3D geometry-aware depth estimation. Song *et al.* [33] introduced the Laplacian pyramid to sharpen the depth boundaries. Bhat *et al.* [8] revisited the ordinal regression [28] and proposed to predict adaptive bins according to the image content. Patil *et al.* [34] developed the piecewise planarity prior and leveraged the displacement vector field to borrow information from coplanar pixels. Additionally, De *et al.* [35], Zhao *et al.* [36], Zhao *et al.* [37] and Metzger *et al.* [38] used graph regularization, discrete cosine transform, spherical space feature decomposition and deep anisotropic diffusion for guided depth map super-resolution, respectively.

B. Transformer

Transformers have attracted widespread attention because of their effectiveness in natural language processing [10]. In the computer vision domain, Dosovitskiy *et al.* [16] introduced the Vision Transformer (ViT) and showed its feasibility for image classification. The success of ViT accelerated the application of Transformer to other tasks. Zheng *et al.* [39] developed one of the first attempts at dense prediction tasks by using ViT as the backbone.

There have been some attempts at applying the Transformer to monocular depth estimation. Bhat *et al.* [8] made use of a minimized ViT to generate bin width adaptively. [40]–[43] used the Transformer as an encoder to attain a global receptive field. As demonstrated in [19], a model with a Transformer encoder tends to lose local depth details. Li *et al.* [19] further combined the Transformer encoder with an additional CNN

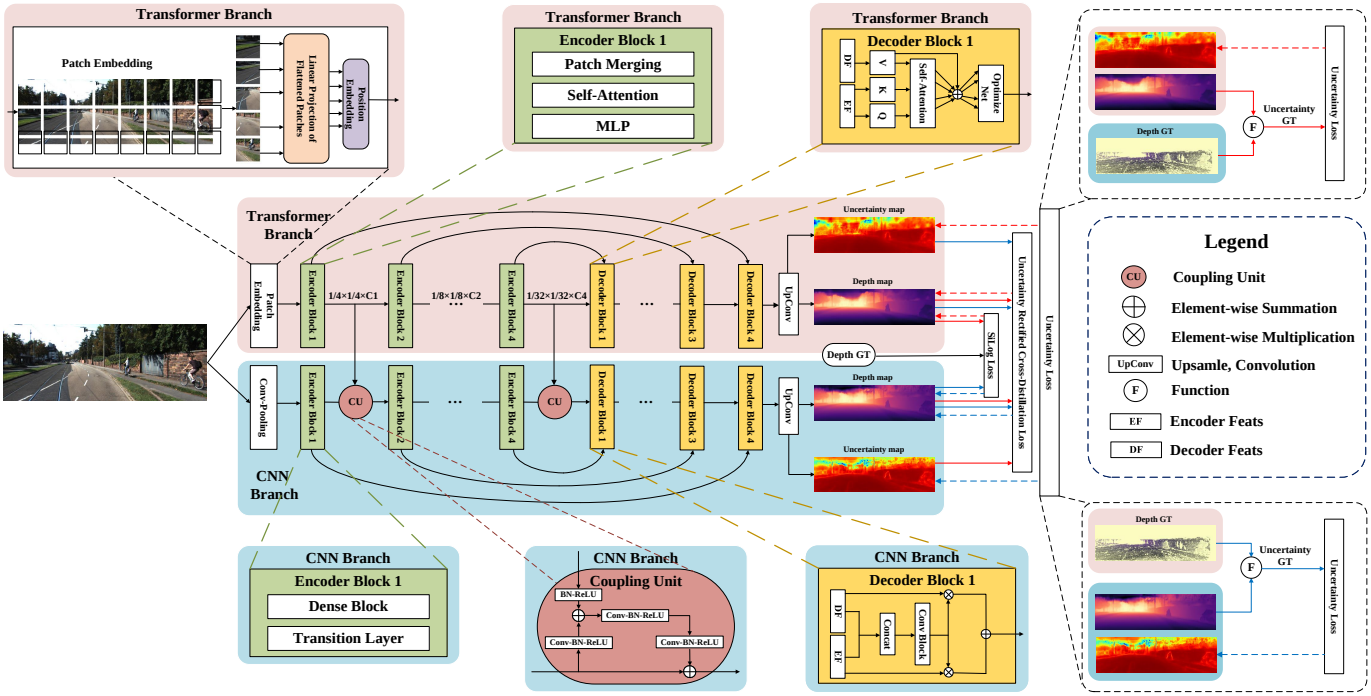

Fig. 2. **Overview of the proposed URCDC-Depth**. The URCDC-Depth in the training phase consists of two branches, a Transformer branch and a CNN branch. During the evaluation phase, only the Transformer branch is used to generate the depth map. The red and blue lines represent the data-loss correspondences for Transformer training and CNN training, respectively.

encoder so that the model can enjoy the desired properties from both networks. However, the ensembled model increases the computational complexity. In contrast, we introduce the cross-distillation between Transformer and CNN to construct a unified depth estimator, which allows our model to use only the Transformer branch at inference time with no additional computational burden.

C. Knowledge distillation

Knowledge distillation (KD) is a learning paradigm aimed at transferring learned knowledge from a teacher model to a student model. It was originally proposed in [44]. Since then, numerous KD variants have been proposed, either aiming to improve its efficacy [45]–[47] or applying it to other tasks [48], [49]. The cross-distillation is a special case of KD, where models can reach a consensus by simultaneously teaching each other, similar to the mutual learning [50]. A few studies have also applied KD to monocular depth estimation [51], [52]. Different from these methods, we introduce an uncertainty rectified cross-distillation for accurate depth estimation.

D. Data augmentation

Data augmentation is a powerful technique for mitigating overfitting risk by increasing the effective amount of training samples. Hence, common data augmentation techniques such as color jitter and rotation are widely adopted in various tasks to improve the model generalization ability. Recently, augmentation techniques tailored for monocular depth estimation have been proposed, such as CutDepth [53] and DataGrafting [18]. The CutDepth aims to shorten the distance between the RGB image and depth map in the latent space by replacing part of the RGB image with the corresponding depth ground-truth. The motivation of CutDepth differs significantly from our CutFlip. On the other hand, although the DataGrafting also aims to alleviate the overfitting risk for vertical image position, it relies on grafting together two training samples with different semantics. In contrast, our CutFlip is simpler and easier to implement because it requires only one training sample.

E. Uncertainty estimation

Estimating uncertainty (or complementary confidence) plays an essential role in a broad range of tasks. In general, there are two main uncertainty types: epistemic and aleatoric. Epistemic uncertainty stems from the model weights and can be reduced with more training data, while aleatoric uncertainty arises from noise in the input data [54]. Bootstrapped ensembles [55] and Monte Carlo Dropout [56] empirically estimate the epistemic uncertainty by sampling from the distribution of model weights. A scheme that accounts for aleatoric uncertainty is performed in a predictive manner, where a network is trained to learn a distribution with mean and variance [57]. Recent studies have incorporated uncertainty into dense prediction tasks, for example, semantic segmentation [58], depth estimation [22], [59] and optical flow [60]. We focus on a more targeted supervised uncertainty estimation setting and model the uncertainty ground-truth in the form of the probability density function of the Laplace distribution.

III. METHODOLOGY

In this section, we elaborate on the main contributions of this work, namely, uncertainty rectified cross-distillation and CutFlip. An overview of URCDC-Depth is presented in Fig. 2.

A. Uncertainty Rectified Cross-Distillation

Network architecture. The proposed URCDC-Depth contains a Transformer branch and a CNN branch in the training phase, which have an encoder-decoder architecture.

For the Transformer branch, the encoder follows the design of Swin Transformer [17]. First, patch embedding projects an input RGB image into a high-dimensional feature representation, which is then propagated into encoder blocks. Each block is composed of patch merging, multi-layer perception (MLP) and self-attention. In the decoder blocks, which are neural window fully-connected conditional random fields (CRFs) modules [43], the self-attention captures long-range correlation in the decoder, which the optimize net optimizes for a better prediction.

As for the CNN branch, the encoder uses DenseNet [5]. Each encoder block contains a dense block and transition layer. The decoder blocks are developed following [42], where the encoder and decoder features are multiplied with the features from the convolution block to emphasize the valuable information. The CNN branch is only used for complementary training and is discarded once the training process is complete.

Cross-distillation of Transformer and CNN. We leverage the efficient cross-distillation paradigm to build a comprehensive depth estimator that fully exploits the long-range correlation and local information. For an input RGB image $\mathbf{r}(\mathbf{p})$, where $\mathbf{p}$ denotes the pixel coordinate, our model generates two depth predictions,

$$\mathbf{d}^t(\mathbf{p}) = f^t(\mathbf{r}(\mathbf{p})); \mathbf{d}^c(\mathbf{p}) = f^c(\mathbf{r}(\mathbf{p})), \tag{1}$$

where $\mathbf{d}^t(\mathbf{p})$ and $\mathbf{d}^c(\mathbf{p})$ stand for the depth predictions from Transformer branch $f^t(\cdot)$ and CNN branch $f^c(\cdot)$, respectively. As mentioned before, Transformer and CNN are asymmetric learning networks, where Transformer relies on the long-range self-attention mechanism while CNN is built upon the local convolution operator. Therefore, $\mathbf{d}^t(\mathbf{p})$ and $\mathbf{d}^c(\mathbf{p})$ have inherently diverse properties and are used as pseudo labels to guide their counterparts toward the correct depth.

Uncertainty-based rectification. Unfortunately, the pseudo labels may contain heavy noise, particularly at the beginning of training. This inevitably damages the cross-distillation and results in incorrect predictions. To alleviate the negative impact of noise, we introduce an uncertainty rectified cross-distillation loss, partially inspired by mutual learning [50] and defined as

$$\begin{aligned}\mathcal{L}_{urcd} = &\underbrace{\sum_{\mathbf{p}} \left(1 - \overline{\mathbf{u}^c}(\mathbf{p})\right) \odot \left|\mathbf{d}^t(\mathbf{p}) - \overline{\mathbf{d}^c}(\mathbf{p})\right|}_{Transformer\ term} \\ &+ \underbrace{\sum_{\mathbf{p}} \left(1 - \overline{\mathbf{u}^t}(\mathbf{p})\right) \odot \left|\mathbf{d}^c(\mathbf{p}) - \overline{\mathbf{d}^t}(\mathbf{p})\right|}_{CNN\ term},\end{aligned} \tag{2}$$

where $\bar{\cdot}$ is the gradient stopping operation, $\odot$ is the element-wise multiplication, and $\mathbf{u}^t(\mathbf{p})$ and $\mathbf{u}^c(\mathbf{p})$ denote the uncertainty predictions from the Transformer branch and CNN branch, respectively. More specifically, the gradient stopping stands for that gradient flowing to the target variable is prohibited during the backpropagation process so that the target variable is not affected. The Transformer term distills knowledge from CNN to Transformer while the CNN term distills knowledge from Transformer to CNN. The uncertainty predictions are used to downweight the relevant pixels to alleviate the negative impact on regions with high uncertainty. The values of uncertainty range from 0 to 1.

Since there is no ground-truth for the uncertainty, we model it with a function inspired by the probability density function of the Laplace distribution,

$$\mathbf{u}^{t*}(\mathbf{p}) = 1 - \exp\left(-\frac{\left|\mathbf{d}^t(\mathbf{p}) - \mathbf{d}^*(\mathbf{p})\right|}{b\left(\mathbf{d}^t(\mathbf{p}) + \mathbf{d}^*(\mathbf{p})\right)}\right), \mathbf{p} \in \mathbf{T}, \tag{3}$$

$$\mathbf{u}^{c*}(\mathbf{p}) = 1 - \exp\left(-\frac{\left|\mathbf{d}^c(\mathbf{p}) - \mathbf{d}^*(\mathbf{p})\right|}{b\left(\mathbf{d}^c(\mathbf{p}) + \mathbf{d}^*(\mathbf{p})\right)}\right), \mathbf{p} \in \mathbf{T}, \tag{4}$$

where $\mathbf{u}^{t*}(\mathbf{p})$ and $\mathbf{u}^{c*}(\mathbf{p})$ stand for the modeled uncertainty ground-truths for the Transformer and CNN branches, $\mathbf{d}^*(\mathbf{p})$ is the depth ground-truth, b is a coefficient that controls the tolerance for error and is set as 0.2 in this work, and $\mathbf{T}$ denotes a set of pixels with valid ground-truth depth values. Here, instead of adopting the absolute difference between depth prediction and ground-truth directly, we normalize it by their sum. This is because a unit depth difference, for example 1m, represents the different uncertainty between distant and nearby points in a scene, and it should be higher on the nearby points and less on the distant points. We apply $\mathcal{L}_u$ to encourage the uncertainty predictions to approximate $\mathbf{u}^{t*}(\mathbf{p})$ and $\mathbf{u}^{c*}(\mathbf{p})$,

$$\mathcal{L}_u = \sum_{\mathbf{p}} \left|\mathbf{u}^t(\mathbf{p}) - \mathbf{u}^{t*}(\mathbf{p})\right| + \sum_{\mathbf{p}} \left|\mathbf{u}^c(\mathbf{p}) - \mathbf{u}^{c*}(\mathbf{p})\right|, \mathbf{p} \in \mathbf{T}. \tag{5}$$

$\mathbf{u}^t(\mathbf{p})$ and $\mathbf{u}^c(\mathbf{p})$, apart from their efficacy in Eq. 2, enable our model to be deployed in safety-critical scenarios by indicating potential errors of depth prediction.

Coupling unit. As demonstrated in recent studies [7], [19], [43], the Swin Transformer-based models perform much better than CNN-based models for depth estimation. Unfortunately, the large capacity gap between teacher and student tends to cause poor performance of knowledge distillation [61]. To bridge this gap, we transfer the feature maps encoded by the Transformer branch to the CNN branch, and design **coupling units** to fuse these two types of features. First, we align the channel dimension of the feature maps via a 1×1 convolution and add them together. Second, we adopt a 3×3 convolution to adaptively fuse the added features. Finally, we adjust the feature channel dimension with a 1×1 convolution to form a residual connection. In addition, BatchNorm (BN) [62] and ReLU activation are leveraged to regularize features. We note that the feature transfer operation only exists in the encoder that mainly determines the model performance. Mathematically,

$$\mathcal{F}^{c'} = \mathrm{ReLU}\left(\mathrm{BN}\left(Conv_{1\times1}\left(\mathcal{F}^c\right)\right)\right), \tag{6}$$

$$\mathcal{F}^{t'} = \text{ReLU}\left(\text{BN}\left(\mathcal{F}^{t}\right)\right), \tag{7}$$

$$\mathcal{F}^{add} = \text{ReLU}\left(\text{BN}\left(Conv_{3\times3}\left(\mathcal{F}^{c'} + \mathcal{F}^{t'}\right)\right)\right), \tag{8}$$

$$\mathcal{F}^{out} = \text{ReLU}\left(\text{BN}\left(Conv_{1\times1}\left(\mathcal{F}^{add}\right)\right) + \mathcal{F}^{c}\right), \tag{9}$$

where $\mathcal{F}^{c}$ and $\mathcal{F}^{t}$ stand for the feature maps from CNN branch and Transformer branch, respectively, $\mathcal{F}^{add}$ is the added CNN and Transformer feature maps and $\mathcal{F}^{out}$ is the output feature maps.

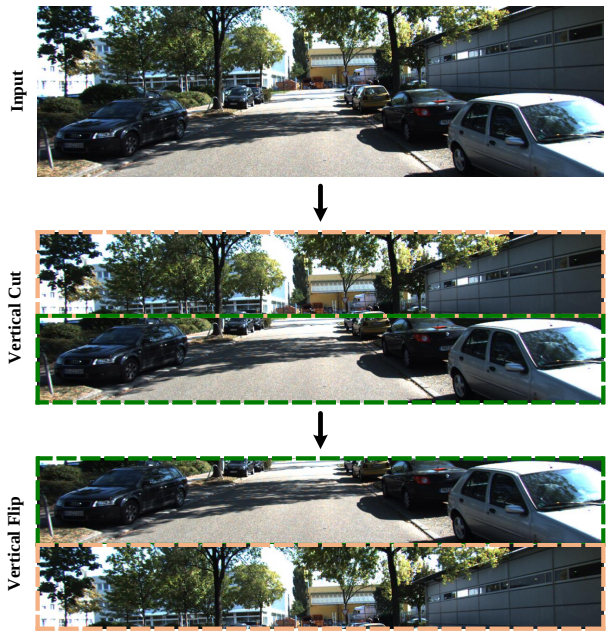

Fig. 3. **Illustration of the CutFlip technique**. The CutFilp contains two key steps: vertical cut and vertical filp.

Overall loss. The total optimization objective to train the URCDC-Depth is summarized as

$$\mathcal{L}_{total} = \mathcal{L}_{silog} + \lambda_1 \mathcal{L}_{urcd} + \lambda_2 \mathcal{L}_{u}, \tag{10}$$

with

$$\begin{aligned}\mathcal{L}_{silog} = \kappa\sqrt{\frac{1}{|\mathbf{T}|}\sum_{\mathbf{p}}\left(\mathbf{g}^{t}\left(\mathbf{p}\right)\right)^{2} - \frac{\eta}{|\mathbf{T}|^{2}}\left(\sum_{\mathbf{p}}\mathbf{g}^{t}\left(\mathbf{p}\right)\right)^{2}} \\ +\kappa\sqrt{\frac{1}{|\mathbf{T}|}\sum_{\mathbf{p}}\left(\mathbf{g}^{c}\left(\mathbf{p}\right)\right)^{2} - \frac{\eta}{|\mathbf{T}|^{2}}\left(\sum_{\mathbf{p}}\mathbf{g}^{c}\left(\mathbf{p}\right)\right)^{2}}, \mathbf{p} \in \mathbf{T},\end{aligned} \tag{11}$$

where $\mathcal{L}_{silog}$ denotes the scaled scale-invariant loss introduced by [7], $\mathbf{g}\left(\mathbf{p}\right) = \log \mathbf{d}\left(\mathbf{p}\right) - \log \mathbf{d}^{*}\left(\mathbf{p}\right)$, κ and η are set as 10 and 0.85 based on [7], and λ_1 and λ_2 are empirically set as 0.1 and 0.5.

B. CutFlip

The quantity and diversity of training data are both critical for deep learning-based models, which is difficult to ensure in depth estimation because data acquisition is extremely expensive and laborious. Lack of data reduces the model generalization ability, and one of the serious overfitting threats is the heavy reliance on the vertical image position [23].

Algorithm 1 CutFlip

Input: RGB image $\mathbf{r}$; Ground-truth depth map $\mathbf{d}^{*}$; Height of input h.
Output: Transformed $\mathbf{r}$; Transformed $\mathbf{d}^{*}$.

```
Random sampling p from the uniform distribution U (0, 1).
if p < 0.5 then
    return r; d*.
else
    Random sampling ς from [floor(0.2h), floor(0.8h)];
    r = r.copy(); d* = d*.copy();
    r[: h − ς, :, :] = r[ς :, :, :]; d*[: h − ς, :, :] = d*[ς :, :, :];
    r[h − ς :, :, :] = r[: ς, :, :]; d*[h − ς :, :, :] = d*[: ς, :, :];
    return r; d*.
end if
```

To enforce the model to focus on more valuable clues, we introduce a surprisingly simple yet highly effective data augmentation technique, namely, CutFlip. As illustrated in Fig. 3, we vertically cut the input sample into upper and lower parts, highlighted by the orange box and the green box, respectively, and flip these two parts along the vertical direction to weaken the relationship between depth and vertical image position. The details are demonstrated in Algorithm 1. The CutFlip is performed with a probability of 0.5, and the vertical position to cut is randomly sampled, which allows the model to greatly adapt to various types of data.

IV. Experiments

We conduct extensive experiments on the standard datasets, including KITTI [24], NYU-Depth-v2 [25] and SUN RGB-D [26]. In the following part, we first display a description of the relevant datasets, evaluation metrics and implementation details. We then show quantitative and qualitative comparisons to the state-of-the-art methods. Finally, we present generalization study and ablation study to perform a detailed analysis of URCDC-Depth.

A. Datasets and evaluation metrics

The KITTI dataset consists of images captured from outdoor scenes with equipment placed on a moving vehicle [24]. The image resolution is around 1241×376 pixels. We use two commonly used splits for monocular depth estimation. One is the Eigen split [27], including 23488 training image pairs and 697 testing images. The other one is the official split [24], including 42949 training image pairs, 1000 validation images and 500 testing images. The evaluation results on the official split are generated by the online server.

The NYU-Depth-v2 dataset provides RGB images and depth maps collected from indoor scenes at a resolution of 640×480 pixels [25]. Following prior works, we adopt the official split and the dataset processed by [7], which contains 24231 training image pairs and 654 testing images.

The SUN RGB-D dataset is collected using four sensors from multiple indoor scenes with high scene diversity, which consists of around 10K images. We use this dataset only for

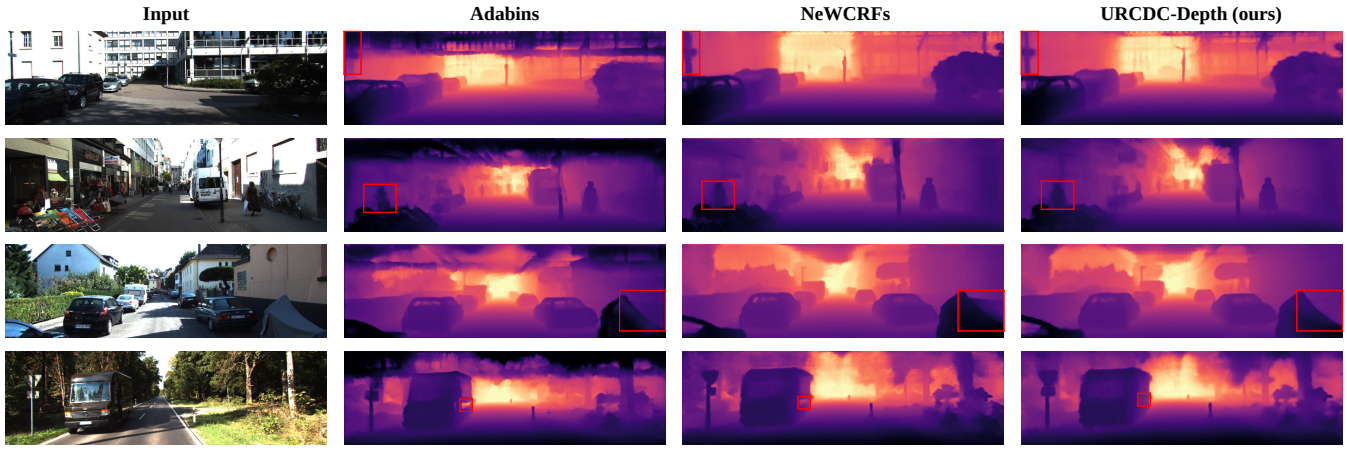

Fig. 4. **Qualitative depth results on the Eigen split of the KITTI dataset**. The red boxes indicate the regions to emphasize.

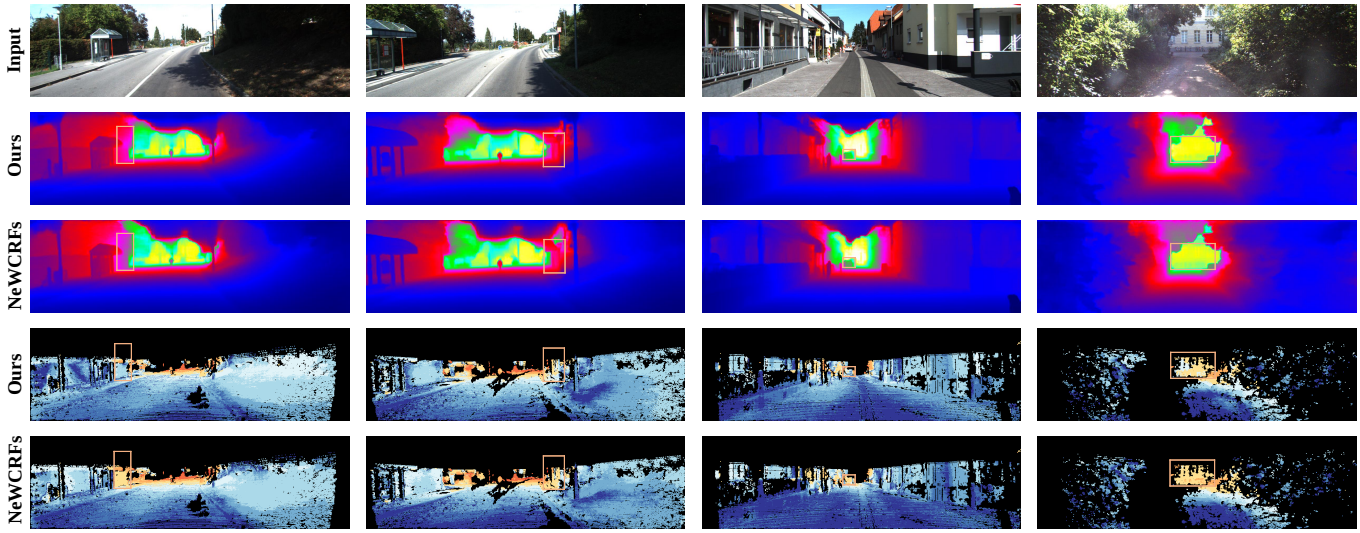

Fig. 5. **Qualitative depth results on the KITTI online benchmark**, generated by the online sever. The orange boxes show the regions to focus on. The second and third rows are depth predictions and the fourth and fifth rows are corresponding error maps, where the large errors are in orange or red.

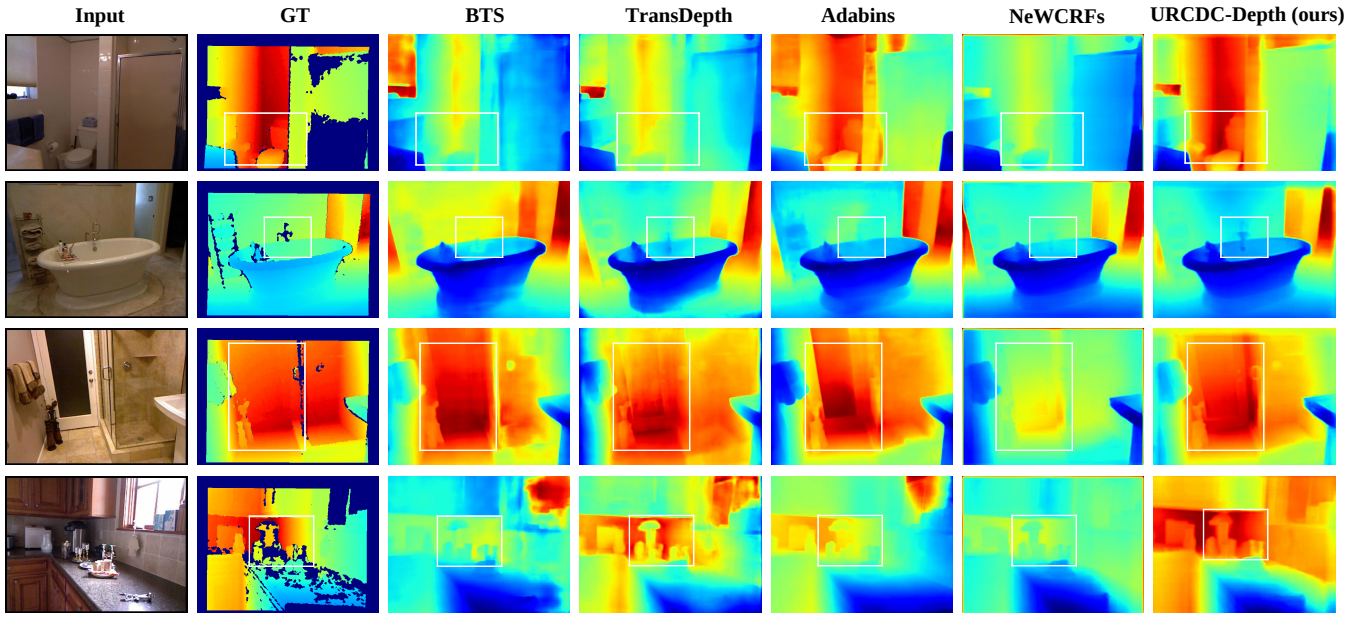

Fig. 6. **Qualitative depth results on the NYU-Depth-v2 dataset**. The white boxes indicate the regions to emphasize.

Method	Cap	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
Eigen *et al.* [63]	0-80m	0.203	1.548	6.307	0.282	0.702	0.898	0.967
Fu *et al.* [28]	0-80m	0.072	0.307	2.727	0.120	0.932	0.984	0.994
VNL [32]	0-80m	0.072	-	3.258	0.117	0.938	0.990	0.998
BTS [7]	0-80m	0.060	0.249	2.798	0.096	0.955	0.993	0.998
PWA [64]	0-80m	0.060	0.221	2.604	0.093	0.958	0.994	**0.999**
TransDepth [40]	0-80m	0.064	0.252	2.755	0.098	0.956	0.994	**0.999**
Adabins [8]	0-80m	0.058	0.190	2.360	0.088	0.964	0.995	**0.999**
P3Depth [34]	0-80m	0.071	0.270	2.842	0.103	0.953	0.993	0.998
DepthFormer [19]	0-80m	0.052	0.158	2.143	0.079	0.975	**0.997**	**0.999**
NeWCRFs [43]	0-80m	0.052	0.155	2.129	0.079	0.974	**0.997**	**0.999**
URCDC-Depth (ours)	0-80m	**0.050**	**0.142**	**2.032**	**0.076**	**0.977**	**0.997**	**0.999**
Fu *et al.* [28]	0-50m	0.071	0.268	2.271	0.116	0.936	0.985	0.995
BTS [7]	0-50m	0.058	0.183	1.995	0.090	0.962	0.994	0.999
LapDepth [33]	0-50m	0.056	0.161	1.830	0.086	0.967	0.995	0.999
PWA [64]	0-50m	0.057	0.161	1.872	0.087	0.965	0.995	0.999
TransDepth [40]	0-50m	0.061	0.185	1.992	0.091	0.963	0.995	0.999
P3Depth [34]	0-50m	0.055	0.130	1.651	0.081	0.974	0.997	0.999
URCDC-Depth (ours)	0-50m	**0.049**	**0.108**	**1.528**	**0.072**	**0.981**	**0.998**	**1.000**

TABLE I
QUANTITATIVE DEPTH COMPARISON ON THE EIGEN SPLIT OF KITTI DATASET. NOTE THAT THE BACKBONES OF DEPTHFORMER, NEWCRFS AND OUR URCDC-DEPTH AT INFERENCE TIME ARE SWIN-LARGE AND RESNET-50, SWIN-LARGE AND SWIN-LARGE, RESPECTIVELY. "-" INDICATES NOT APPLICABLE. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Method	SILog ↓	Sq Rel ↓	Abs Rel ↓	iRMSE↓
P3Depth [34]	12.82	9.92	2.53	13.71
VNL [32]	12.65	10.15	2.46	13.02
GAC [65]	12.13	9.41	2.61	12.65
FPANet [66]	11.84	9.23	2.46	12.63
DLE [67]	11.81	9.09	2.22	12.49
Fu *et al.* [28]	11.77	8.78	2.23	12.98
BTS [7]	11.67	9.04	2.21	12.23
BA-Full [68]	11.61	9.38	2.29	12.23
PackNet-SAN [69]	11.54	9.12	2.35	12.38
PWA [64]	11.45	9.05	2.30	12.32
NeWCRFs [43]	10.39	8.37	1.83	11.03
URCDC-Depth (ours)	**10.03**	**8.24**	**1.74**	**10.71**

TABLE II
QUANTITATIVE DEPTH COMPARISON ON THE OFFICIAL SPLIT OF KITTI DATASET. THE RESULTS ARE AVAILABLE FROM THE ONLINE SERVER.

zero-shot evaluation and do not use it for training. The pre-trained models are evaluated on the official test set of 5050 images.

Evaluation metrics. Following previous works [7], [43], we adopt the standard evaluation metrics in our experiments.

- SILog: $\sqrt{\frac{1}{|\mathbf{T}|}\sum_{\mathbf{g}\in\mathbf{T}}(\mathbf{g})^2 - \frac{1}{|\mathbf{T}|^2}\left(\sum_{\mathbf{g}\in\mathbf{T}}\mathbf{g}\right)^2}$;
- Abs Rel: $\frac{1}{|\mathbf{T}|}\sum_{\mathbf{d}\in\mathbf{T}}|\mathbf{d}-\mathbf{d}^*|/\mathbf{d}^*$;
- Sq Rel (eigen split): $\frac{1}{|\mathbf{T}|}\sum_{\mathbf{d}\in\mathbf{T}}\|\mathbf{d}-\mathbf{d}^*\|^2/\mathbf{d}^*$;
- Sq Rel (official split): $\frac{1}{|\mathbf{T}|}\sum_{\mathbf{d}\in\mathbf{T}}\|\mathbf{d}-\mathbf{d}^*\|^2/\mathbf{d}^{*2}$;
- RMSE: $\sqrt{\frac{1}{|\mathbf{T}|}\sum_{\mathbf{d}\in\mathbf{T}}\|\mathbf{d}-\mathbf{d}^*\|^2}$;
- RMSE log: $\sqrt{\frac{1}{|\mathbf{T}|}\sum_{\mathbf{d}\in\mathbf{T}}\|\log\mathbf{d}-\log\mathbf{d}^*\|^2}$;
- iRMSE: $\sqrt{\frac{1}{|\mathbf{T}|}\sum_{\mathbf{d}\in\mathbf{T}}\|1/\mathbf{d}-\mathbf{1}/\mathbf{d}^*\|^2}$;
- $\log_{10}$: $\frac{1}{|\mathbf{T}|}\sum_{\mathbf{d}\in\mathbf{T}}\|\log_{10}\mathbf{d}-\log_{10}\mathbf{d}^*\|^2$;
- $\delta < thr$: % of $\mathbf{d}$ satisfies $\left(\max\left(\frac{\mathbf{d}}{\mathbf{d}^*},\frac{\mathbf{d}^*}{\mathbf{d}}\right)=\delta<thr\right)$ for $thr = 1.25, 1.25^2, 1.25^3$.

B. Implementation Details

The URCDC-Depth is implemented in the PyTorch [70] and trained on 4 NVIDIA RTX A5000 GPUs. We optimize it by using the Adam optimizer [71] with $\beta 1 = 0.9, \beta 2 = 0.999$. The training process runs a total number of 20 epochs with a batch size of 8 and a learning rate scheduled via polynomial decay from 2e-5 to 2e-6. Apart from the proposed CutFlip, we use the common data augmentation techniques as in previous works [7], [43]. The number of coupling units is 4.

C. Comparison to State-of-the-Arts

KITTI. We first conduct a comparison with the leading methods on the Eigen split. Table I presents the results. They indicate that the performance of our URCDC-Depth exceeds those of previous methods. It is worth noting that although URCDC-Depth and NeWCRFs share an almost identical network structure in the evaluation phase, it improves upon the NeWCRFs by 8.4% and 4.6% on the Sq Rel and RMSE for the 0-80m cap. When the maximum depth is capped at 50m, the performance of our URCDC-Depth far exceeds that of P3Depth. Fig. 4 demonstrates qualitative depth comparisons. It can be seen that the NeWCRFs struggles with thinner structures, such as difficult human and pole boundaries, while

Method	Cap	Abs Rel ↓	RMSE ↓	$\log_{10}$ ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
Eigen *et al.* [27]	0-10m	0.158	0.641	-	0.769	0.950	0.988
Fu *et al.* [28]	0-10m	0.115	0.509	0.051	0.828	0.965	0.992
VNL [32]	0-10m	0.108	0.416	0.048	0.875	0.976	0.994
LapDepth [33]	0-10m	0.110	0.393	0.047	0.885	0.979	0.995
BTS [7]	0-10m	0.110	0.392	0.047	0.885	0.978	0.994
DAV [15]	0-10m	0.108	0.412	-	0.882	0.980	0.996
PWA [64]	0-10m	0.105	0.374	0.045	0.892	0.985	0.997
Long *et al.* [72]	0-10m	0.101	0.377	0.044	0.890	0.982	0.996
TransDepth [40]	0-10m	0.106	0.365	0.045	0.900	0.983	0.996
Adabins [8]	0-10m	0.103	0.364	0.044	0.903	0.984	0.997
P3Depth [34]	0-10m	0.104	0.356	0.043	0.898	0.981	0.996
DepthFormer [19]	0-10m	0.096	0.339	0.041	0.921	0.989	0.998
NeWCRFs [43]	0-10m	0.095	0.334	0.041	0.922	**0.992**	**0.998**
URCDC-Depth (ours)	0-10m	**0.088**	**0.316**	**0.038**	**0.933**	**0.992**	**0.998**

TABLE III
QUANTITATIVE DEPTH COMPARISON ON THE NYU-DEPTH-V2 DATASET.

URCDC-Depth is capable of estimating these small details. This supports our standpoint that the cross-distillation helps learn a unified depth estimator by combining the desired properties of both Transformer and CNN.

Method	Abs Rel ↓	RMSE ↓	$\log_{10}$ ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑
Chen [73]	0.166	0.494	0.071	0.757	0.943
VNL [32]	0.183	0.541	0.082	0.696	0.912
BTS [7]	0.172	0.515	0.075	0.740	0.933
Adabins [8]	0.159	0.476	0.068	0.771	0.944
Ours	**0.136**	**0.408**	**0.060**	**0.819**	**0.971**

TABLE IV
GENERALIZATION TO THE SUN RGB-D DATASET. THE MODELS ARE TRAINED ON THE NYU-DEPTH-V2.

ID	CD	UR	CU	CF	Abs Rel ↓	Sq Rel ↓	RMSE ↓	$\delta < 1.25$ ↑
1					0.052	0.155	2.129	0.974
2	✓				0.055	0.155	2.086	0.975
3		✓			0.052	0.156	2.122	0.974
4	✓	✓			0.051	0.147	2.076	0.976
5	✓		✓		0.051	0.148	2.078	0.975
6	✓	✓	✓		0.051	0.147	2.062	0.976
7				✓	0.051	0.144	2.056	**0.977**
8	✓	✓	✓	✓	**0.050**	**0.142**	**2.032**	**0.977**

TABLE V
ABLATION STUDY OF THE PROPOSED URCDC-DEPTH ON THE KITTI DATASET. CD: CROSS-DISTILLATION; UR: UNCERTAINTY-BASED RECTIFICATION; CU: COUPLING UNIT; CF: CUTFLIP.

We then compare our URCDC-Depth against the competing methods on the official split. The results are generated by the online server and reported in Table II. Here, we can see that URCDC-Depth outperforms previous methods again. A notable phenomenon is that the main ranking metric SILog, from the model proposed by Fu *et al.* [28] to the PWA [64], has only increased by 2.7%. With the advent of the Swin Transformer, the NeWCRFs makes a significant breakthrough on this metric through the designed neural CRFs. Our URCDC-Depth further improves the NeWCRFs by 3.5% on the SILog, with no additional computational burden at inference time. The colored visualizations of the depth predictions from the online test are shown in Fig. 5.

ID	CD	UR	CU	CF	Abs Rel ↓	RMSE ↓	$\log_{10}$ ↓	$\delta < 1.25$ ↑
1					0.095	0.334	0.041	0.922
2	✓				0.095	0.329	0.040	0.923
3		✓			0.095	0.335	0.040	0.922
4	✓	✓			0.091	0.323	0.039	0.928
5	✓		✓		0.091	0.326	0.039	0.926
6	✓	✓	✓		0.089	0.319	**0.038**	0.931
7				✓	0.091	0.322	0.039	**0.934**
8	✓	✓	✓	✓	**0.088**	**0.316**	**0.038**	0.933

TABLE VI
ABLATION STUDY OF THE PROPOSED URCDC-DEPTH ON THE NYU-DEPTH-V2 DATASET.

NYU-Depth-v2. To show the effectiveness of URCDC-Depth in the indoor scenario, we evaluate it on the NYU-Depth-v2 dataset. The results are reported in Table III, which indicate that our method has greatly boosted the performance on most metrics, *e.g.*, Abs Rel and RMSE. This emphasizes our contributions in improving the results. We display qualitative depth comparisons in Fig. 6. As can be seen, our URCDC-Depth preserves small details *e.g.*, handle.

We further convert depth predictions into point clouds and present qualitative point cloud comparisons on the KITTI and NYU-Depth-v2 datasets in Fig. 7 and Fig. 8, respectively. The point clouds of our URCDC-Depth show fewer distortions and are with prominent geometric features.

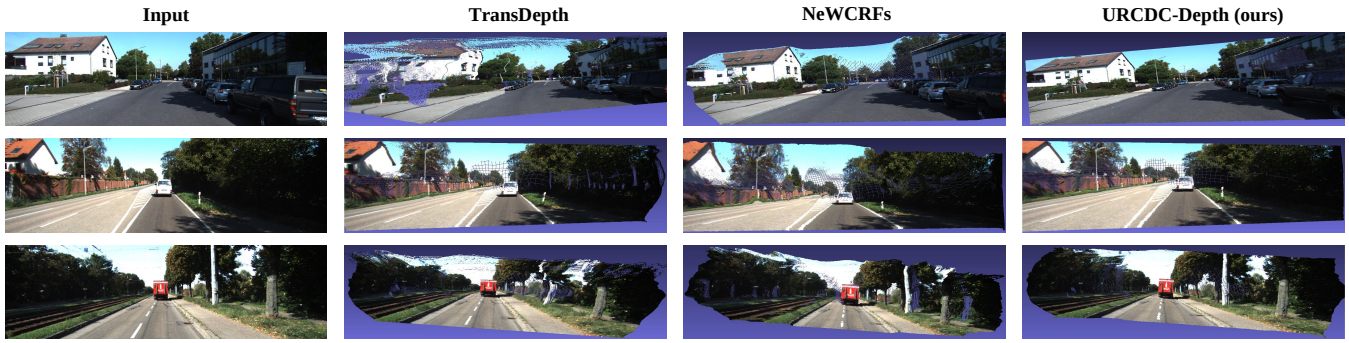

Fig. 7. **Qualitative point cloud results on the KITTI dataset.**

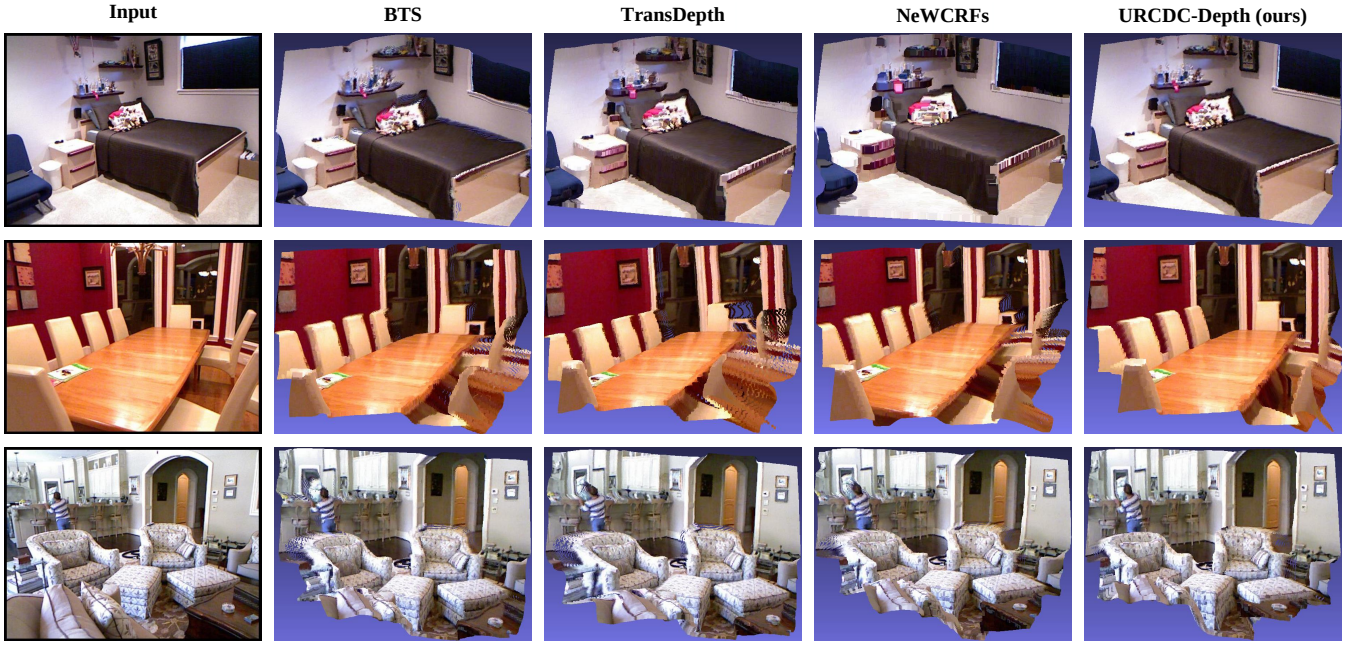

Fig. 8. **Qualitative point cloud results on the NYU-Depth-v2 dataset.**

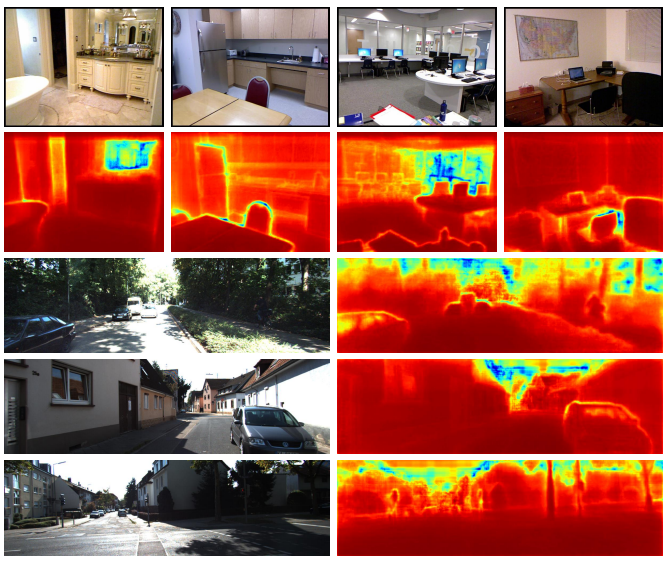

Fig. 9. **Pixel-wise depth uncertainty from our URCDC-Depth**. Red indicates areas with low uncertainty, and yellow/blue indicates areas with high/highest uncertainty.

D. Generalization Ability

As listed in Table IV, we investigate the generalization ability of URCDC-Depth in a zero-shot setting. The models are trained on NYU-Depth-v2 but evaluated on SUN RGB-D to validate that our model does learn transferable discriminative features rather than simply memorizing training data. URCDC-Depth achieves the best performance on all metrics, which is an indicator of its excellent generalizability across different cameras.

E. Ablation Study

To better understand how the components in URCDC-Depth affect the performance, we conduct detailed ablation studies on the KITTI and NYU-Depth-v2 datasets. The results are presented in Tables V, VI, VII, VIII, IX, X and Figs. 9, 10, 11.

Cross-distillation (Tables V and VI). We start from the baseline NeWCRFs (ID 1). By directly introducing cross-distillation, we observe a slight performance improvement on most metrics (ID 2). The cross-distillation suffers from the heavy depth noise from pseudo labels. In addition, the large

Method	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
BTS [7]	0.060	0.249	2.798	0.096	0.955	0.993	0.998
+ CutFlip	**0.059**	**0.198**	**2.366**	**0.091**	**0.959**	**0.995**	**0.999**
TransDepth [40]	**0.064**	0.252	2.755	0.098	0.956	0.994	**0.999**
+ CutFlip	**0.064**	**0.222**	**2.457**	**0.095**	**0.958**	**0.995**	**0.999**
Adabins [8]	0.058	0.190	2.360	0.088	0.964	0.995	**0.999**
+ CutFlip	**0.056**	**0.174**	**2.236**	**0.086**	**0.968**	**0.996**	**0.999**

TABLE VII
EFFECT OF THE PROPOSED CUTFLIP FOR DIFFERENT ALGORITHMS ON THE KITTI DATASET, USING BTS, TRANSDEPTH AND ADABINS AS CASES.

capacity gap between the Transformer branch and CNN branch limits the performance gain from cross-distillation.

Uncertainty-based rectification (Tables V and VI, Fig. 9). Adding an additional uncertainty estimation head has little impact on performance (ID 3). However, leveraging the estimated uncertainty in cross-distillation can significantly boost the performance (ID 4). Especially notable are the results on the Sq Rel, which is sensitive to large depth errors. The rectification is able to mitigate the negative impact of depth noise in the cross-distillation, hence resulting in a markedly lower Sq Rel result.

Augmentation method	Abs Rel ↓	Sq Rel ↓	RMSE ↓	$\delta < 1.25 \uparrow$
CutMix [74]	0.054	0.154	2.093	0.974
CutOut [75]	0.052	0.148	2.078	0.975
CutDepth [53]	0.052	0.150	2.076	0.975
DataGrafting [18]	0.051	0.146	2.051	0.976
CutFlip	**0.050**	**0.142**	**2.032**	**0.977**

TABLE VIII
COMPARISON OF CUTFLIP AGAINST CUTMIX, CUTOUT, CUTDEPTH AND DATAGRAFTING ON THE KITTI DATASET.

Method	Abs Rel ↓	Sq Rel ↓	RMSE ↓	$\delta < 1.25 \uparrow$
Base.	**0.051**	0.149	2.067	0.975
Base. w/ Color Jitter	**0.051**	0.148	2.059	**0.976**
Base. w/ Random Rotation	**0.051**	0.149	2.064	**0.976**
Base. w/ CutFlip	**0.051**	**0.145**	**2.040**	**0.976**

TABLE IX
COMPARISON OF CUTFLIP AGAINST CONVENTIONAL COLOR JITTER AND RANDOM ROTATION ON THE KITTI DATASET. BASE.: BASELINE.

Method	Abs Rel ↓	Sq Rel ↓	RMSE ↓	$\delta < 1.25 \uparrow$
Different Encoders in the CNN Branch				
ResNet101 [4]	**0.050**	0.145	2.040	**0.977**
ResNeXt50 [76]	**0.050**	0.146	2.045	**0.977**
DenseNet161 [5]	**0.050**	**0.142**	**2.032**	**0.977**
Different Decoders in the CNN Branch				
BTS [7]	**0.050**	0.144	2.043	**0.977**
NeWCRFs [43]	0.051	0.147	2.049	0.976
GLPDepth [42]	**0.050**	**0.142**	**2.032**	**0.977**

TABLE X
EFFECT OF DIFFERENT ENCODERS AND DECODERS IN THE CNN BRANCH FOR CROSS-DISTILLATION ON THE KITTI DATASET.

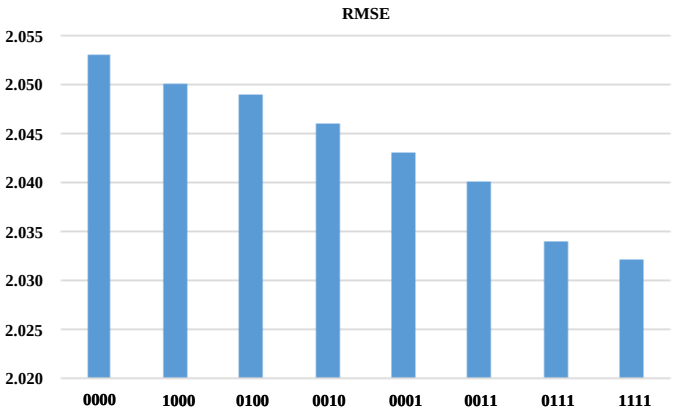

Fig. 10. **Effect of coupling unit position and number on the KITTI dataset.** There are four encoder blocks in the CNN branch. 1 indicates that the coupling unit is used to enhance the features output by the encoder block, and 0 is the opposite.

Fig. 9 displays the visualization of uncertainty maps. The existing method (Fig. 6 in [14]) is also capable of estimating uncertainty maps by showing a trend where the uncertainty increases with distance. In contrast, our uncertainty maps pay attention to the difficult regions, *e.g.*, object boundaries and vanishing points. We attribute this to the normalization operation when modeling the uncertainty ground-truth.

Coupling unit (Tables V and VI, Fig. 10). To bridge the large capacity gap between the Transformer branch and CNN branch, we transfer the feature maps and develop the coupling units to fuse the transferred features from the Transformer branch and the features in the CNN branch. The results after adding the coupling units are in IDs 5 and 6, which contribute to the performance.

We further display the effect of the coupling unit position and number in Fig. 10. When there is only one coupling unit, placing it behind the last encoder block works better, i.e., the abstract features transferred by the Transformer branch are more conducive to compensating for the capabilities between the Transformer and CNN branches. As the number of coupling units increases, the accuracy will be higher.

CutFlip (Tables V, VI, VII, VIII and IX). With the CutFlip technique, we can see that the performance is largely improved (IDs 7 and 8).

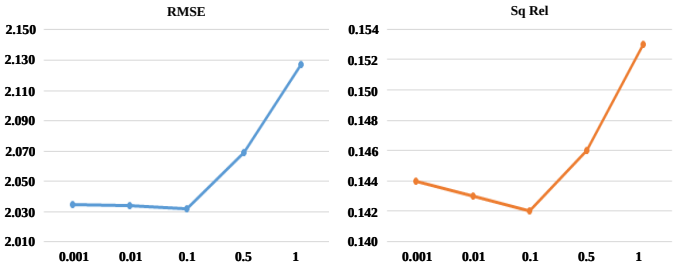

Fig. 11. **Effect of varied uncertainty-rectified cross-distillation loss weight**.

We explore the effect of CutFlip for different depth estimation algorithms in Table VII, such as BTS, TransDepth and Adabins. As we can see, our CutFlip is capable of consistently improving all three methods on most metrics.

Besides, we compare CutFlip against other similar augmentation methods. The results are shown in Table VIII. To make a fair comparison, we remain all other configurations the same except for these specially crafted data augmentation techniques. The inferior performance of CutMix, CutOut and CutDepth may lie in the lack of constraint on vertical image position. DataGrafting takes into consideration the overfitting threat of vertical image position. However, grafting together two training samples with different semantics increases the learning burden of the network.

Furthermore, we compare the CutFlip against conventional augmentation methods in Table IX. The baseline is constructed by removing color jitter, random rotation and CutFilp from our framework. We add these three techniques to the baseline and see that our CutFlip is superior to color jitter and random rotation.

Uncertainty rectified cross-distillation loss (Fig. 11). We notice that 0.1 is the most suitable loss weight, and reducing or increasing it from 0.1 results in performance degradation. This is because reducing the weight will weaken the influence of cross-distillation, while increasing the weight will hinder the learning of the depth ground-truth by the Transformer and CNN branches.

Different encoders and decoders in the CNN branch (Table X). For encoders, we choose widely used ResNet101, ResNeXt50 and DenseNet161 considering the balance of parameters and FLOPs. In terms of decoders, we choose well-behaved BTS, NeWCRFs and GLPDepth. As can be seen, different encoders have little influence on the cross-distillation. Besides, the decoders of GLPDepth and BTS achieve a similar cross-distillation performance, while the decoder of NeWCRFs performs worse, which is caused by deploying the same decoder for both CNN and Transformer branches thereby deteriorating the complementary training process.

V. Conclusion

In this work, we introduce a novel monocular depth estimation framework called URCDC-Depth, which leverages uncertainty rectified cross-distillation to fully exploit the long-range correlation and local information. This paradigm enables our framework to generate a high-quality depth prediction with no additional computational burden at inference time. In addition, we propose a simple yet effective data augmentation technique named CutFlip to encourage the model to emphasize more valuable depth reasoning clues beyond the vertical image position. We conduct extensive experiments on the KITTI, NYU-Depth-v2 and SUN RGB-D datasets, and the results verify the efficacy of the proposed URCDC-Depth.

References

[1] Po-Yi Chen, Alexander H Liu, Yen-Cheng Liu, and Yu-Chiang Frank Wang. Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2624–2632, 2019.

[2] Shuwei Shao, Zhongcai Pei, Weihai Chen, Wentao Zhu, Xingming Wu, Dianmin Sun, and Baochang Zhang. Self-supervised monocular depth and ego-motion estimation in endoscopy: Appearance flow to the rescue. *Medical Image Analysis*, 77:102338, 2022.

[3] Zhengguo Li, Chaobing Zheng, Haiyan Shu, and Shiqian Wu. Dual-scale single image dehazing via neural augmentation. *IEEE Transactions on Image Processing*, 31:6213–6223, 2022.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.

[5] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

[6] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[7] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019.

[8] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021.

[9] Ashutosh Saxena, Sung H Chung, Andrew Y Ng, et al. Learning depth from single monocular images. In *Advances in Neural Information Processing Systems*, volume 18, pages 1–8, 2005.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[11] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017.

[12] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.

[13] Junsheng Zhou, Yuwang Wang, Kaihuai Qin, and Wenjun Zeng. Unsupervised high-resolution depth learning from videos with dual networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6872–6881, 2019.

[14] Adrian Johnston and Gustavo Carneiro. Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4756–4765, 2020.

[15] Lam Huynh, Phong Nguyen-Ha, Jiri Matas, Esa Rahtu, and Janne Heikkilä. Guiding monocular depth estimation using depth-attention volume. In *European Conference on Computer Vision*, pages 581–597. Springer, 2020.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2021.

[17] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision

transformer using shifted windows. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10012–10022, October 2021.

[18] Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. Conformer: local features coupling global representations for visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 367–376, October 2021.

[19] Zhenyu Li, Zehui Chen, Xianming Liu, and Junjun Jiang. Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation. *arXiv preprint arXiv:2203.14211*, 2022.

[20] Shuwei Shao, Ran Li, Zhongcai Pei, Zhong Liu, Weihai Chen, Wentao Zhu, Xingming Wu, and Baochang Zhang. Towards comprehensive monocular depth estimation: Multiple heads are better than one. *IEEE Transactions on Multimedia*, 2022.

[21] Dongting Hu, Liuhua Peng, Tingjin Chu, Xiaoxing Zhang, Yinian Mao, Howard Bondell, and Mingming Gong. Uncertainty quantification in depth estimation via constrained ordinal regression. In *Proceedings of the European Conference on Computer Vision*, pages 237–256. Springer, 2022.

[22] Julia Hornauer and Vasileios Belagiannis. Gradient-based uncertainty for monocular depth estimation. In *Proceedings of the European Conference on Computer Vision*, pages 613–630. Springer, 2022.

[23] Tom van Dijk and Guido de Croon. How do neural networks see depth in single images? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2183–2191, 2019.

[24] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[25] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, pages 746–760. Springer, 2012.

[26] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 567–576, 2015.

[27] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems*, pages 2366–2374, 2014.

[28] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2002–2011, 2018.

[29] Baoliang Chen, Cheolkon Jung, and Zhendong Zhang. Variational fusion of time-of-flight and stereo data for depth estimation using edge-selective joint filtering. *IEEE Transactions on Multimedia*, 20(11):2882–2890, 2018.

[30] Xin Yang, Yang Gao, Hongcheng Luo, Chunyuan Liao, and Kwang-Ting Cheng. Bayesian denet: Monocular depth prediction and frame-wise fusion with synchronized uncertainty. *IEEE Transactions on Multimedia*, 21(11):2701–2713, 2019.

[31] Wenfeng Song, Shuai Li, Ji Liu, Aimin Hao, Qinping Zhao, and Hong Qin. Contextualized cnn for scene-aware depth estimation from single rgb image. *IEEE Transactions on Multimedia*, 22(5):1220–1233, 2019.

[32] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5684–5693, 2019.

[33] Minsoo Song, Seokjae Lim, and Wonjun Kim. Monocular depth estimation using laplacian pyramid-based depth residuals. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(11):4381–4393, 2021.

[34] Vaishakh Patil, Christos Sakaridis, Alexander Liniger, and Luc Van Gool. P3depth: Monocular depth estimation with a piecewise planarity prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1610–1621, 2022.

[35] Riccardo De Lutio, Alexander Becker, Stefano D'Aronco, Stefania Russo, Jan D Wegner, and Konrad Schindler. Learning graph regularisation for guided super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1979–1988, 2022.

[36] Zixiang Zhao, Jiangshe Zhang, Shuang Xu, Zudi Lin, and Hanspeter Pfister. Discrete cosine transform network for guided depth map super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5697–5707, 2022.

[37] Zixiang Zhao, Jiangshe Zhang, Xiang Gu, Chengli Tan, Shuang Xu, Yulun Zhang, Radu Timofte, and Luc Van Gool. Spherical space feature decomposition for guided depth map super-resolution. *arXiv preprint arXiv:2303.08942*, 2023.

[38] Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Guided depth super-resolution by deep anisotropic diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18237–18246, 2023.

[39] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6881–6890, 2021.

[40] Guanglei Yang, Hao Tang, Mingli Ding, Nicu Sebe, and Elisa Ricci. Transformer-based attention networks for continuous pixel-wise prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 16269–16279, October 2021.

[41] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 12179–12188, 2021.

[42] Doyeon Kim, Woonghyun Ga, Pyungwhan Ahn, Donggyu Joo, Sehwan Chun, and Junmo Kim. Global-local path networks for monocular depth estimation with vertical cutdepth. *arXiv preprint arXiv:2201.07436*, 2022.

[43] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. Neural window fully-connected crfs for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3916–3925, June 2022.

[44] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

[45] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4133–4141, 2017.

[46] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2019.

[47] Dawei Sun, Anbang Yao, Aojun Zhou, and Hao Zhao. Deeply-supervised knowledge synergy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6997–7006, 2019.

[48] Nuno C Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *Proceedings of the European Conference on Computer Vision*, pages 103–118, 2018.

[49] Frank Hafner, Amran Bhuiyan, Julian FP Kooij, and Eric Granger. A cross-modal distillation network for person re-identification in rgb-depth. *arXiv preprint arXiv:1810.11641*, 2018.

[50] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4320–4328, 2018.

[51] Andrea Pilzer, Stephane Lathuiliere, Nicu Sebe, and Elisa Ricci. Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9768–9777, 2019.

[52] Filippo Aleotti, Giulio Zaccaroni, Luca Bartolomei, Matteo Poggi, Fabio Tosi, and Stefano Mattoccia. Real-time single image depth perception in the wild with handheld devices. *Sensors*, 21(1):15, 2020.

[53] Yasunori Ishii and Takayoshi Yamashita. Cutdepth: edge-aware data augmentation in depth estimation. *arXiv preprint arXiv:2107.07684*, 2021.

[54] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30, 2017.

[55] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 2017.

[56] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the International Conference on Machine Learning*, pages 1050–1059. PMLR, 2016.

[57] Maria Klodt and Andrea Vedaldi. Supervising the new with the old: learning sfm from sfm. In *Proceedings of the European Conference on Computer Vision*, pages 698–713, 2018.

[58] Fredrik K Gustafsson, Martin Danelljan, and Thomas B Schon. Evaluating scalable bayesian deep learning methods for robust computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops*, pages 318–319, 2020.

[59] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. On the uncertainty of self-supervised monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3227–3237, 2020.

[60] Anne S Wannenwetsch, Margret Keuper, and Stefan Roth. Probflow: Joint optical flow and uncertainty estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1173–1182, 2017.

[61] Junjie Hu, Chenyou Fan, Hualie Jiang, Xiyue Guo, Yuan Gao, Xiangyong Lu, and Tin Lun Lam. Boosting light-weight depth estimation via knowledge distillation. *arXiv preprint arXiv:2105.06143*, 2021.

[62] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456. PMLR, 2015.

[63] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2650–2658, 2015.

[64] Sihaeng Lee, Janghyeon Lee, Byungju Kim, Eojindl Yi, and Junmo Kim. Patch-wise attention network for monocular depth estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1873–1881, 2021.

[65] Yuxuan Liu, Yuan Yixuan, and Ming Liu. Ground-aware monocular 3d object detection for autonomous driving. *IEEE Robotics and Automation Letters*, 6(2):919–926, 2021.

[66] Yifang Xu, Chenglei Peng, Ming Li, Yang Li, and Sidan Du. Pyramid feature attention network for monocular depth prediction. In *2021 IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE, 2021.

[67] Ce Liu, Shuhang Gu, Luc Van Gool, and Radu Timofte. Deep line encoding for monocular 3d object detection and depth prediction. In *British Machine Vision Conference*, page 354. BMVA Press, 2021.

[68] Shubhra Aich, Jean Marie Uwabeza Vianney, Md Amirul Islam, and Mannat Kaur Bingbing Liu. Bidirectional attention network for monocular depth estimation. In *2021 IEEE International Conference on Robotics and Automation*, pages 11746–11752. IEEE, 2021.

[69] Vitor Guizilini, Rares Ambrus, Wolfram Burgard, and Adrien Gaidon. Sparse auxiliary networks for unified monocular depth prediction and completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11078–11088, 2021.

[70] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *Advances in Neural Information Processing Systems Workshop Autodiff*, 2017.

[71] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[72] Xiaoxiao Long, Cheng Lin, Lingjie Liu, Wei Li, Christian Theobalt, Ruigang Yang, and Wenping Wang. Adaptive surface normal constraint for depth estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 12849–12858, October 2021.

[73] Xiaotian Chen, Xuejin Chen, and Zheng-Jun Zha. Structure-aware residual pyramid network for monocular depth estimation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 694–700, 2019.

[74] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6023–6032, 2019.

[75] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[76] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1492–1500, 2017.