EDADepth: Enhanced Data Augmentation for Monocular Depth Estimation

Nischal Khanal and Shivanand Venakanna Sheshappanavar
Geometric Intelligence Research Lab, University of Wyoming

Fig. 1: Our EDADepth model takes single image **(top)** and estimates depth using a pre-trained U-Net model. It uses the BEiT semantic segmentation model to extract context for the generation of depth maps **(Middle)**. 3D point cloud (**Bottom**) is constructed from the estimated depth map and the respective input RGB image.

***Abstract*—Due to their text-to-image synthesis feature, diffusion models have recently seen a rise in visual perception tasks, such as depth estimation. The lack of good-quality datasets makes the extraction of a fine-grain semantic context challenging for the diffusion models. The semantic context with fewer details further worsens the process of creating effective text embeddings that will be used as input for diffusion models. In this paper, we propose a novel EDADepth, an enhanced data augmentation method to estimate monocular depth without using additional training data. We use Swin2SR, a super-resolution model, to enhance the quality of input images. We employ the BEiT pre-trained semantic segmentation model for better extraction of text embeddings. We introduce BLIP-2 tokenizer to generate tokens from these text embeddings. The novelty of our approach is the introduction of Swin2SR, the BEiT model, and the BLIP-2 tokenizer in the diffusion-based pipeline for the estimation of monocular depth. Our model achieves state-of-the-art results (SOTA) on the δ_3 metric on both NYUv2 and KITTI datasets. It also achieves results comparable to those of the SOTA models in the RMSE and REL metrics. Finally, we also show improvements in the visualization of the estimated depth in comparison to the SOTA diffusion-based monocular depth estimation models. Anonymous code repository: https://github.com/edadepthmde/EDADepth_ICMLA.**

***Index Terms*—Monocular Depth Estimation, Semantic Context, Text Embbedings, Tokenizer.**

I. Introduction

Depth estimation is an essential task in computer vision that measures the distance of each pixel relative to a camera. Depth is necessary for operations such as 3D reconstruction [1] (refer 1) and scene understanding [2]. One of the types of depth estimation is Monocular Depth Estimation (MDE) [3]. MDE is a task that estimates the depth of an object using a single-view image. Since single-view images do not have epipolar geometry [4], it is challenging to determine the depth of each pixel. Traditional methods for depth estimation used monocular cues [5] and shading [6]. However, such methods faced challenges, such as the varying image lighting and the need for precise camera calibration. Such limitations suggested a need for a technique to estimate depth value based on per-pixel regression[7], a task commonly used in deep learning [3]. Hence, deep learning methods have emerged as a reliable solution for depth estimation.

One of the most popular deep learning methods used in computer vision is Transformers [8]. Transformers employ self-attention mechanisms, making them a good choice for capturing long-range dependencies [9]. Long-range dependen-

cies are significant in MDE, as they capture contextual information from various regions in a single image. Hence, Transformers have successfully been applied to estimate monocular depth. Furthermore, transformers effectively create generative models such as the diffusion model [10]. Diffusion models have been widely used for text-to-image generation and image denoising. Additionally, data augmentation techniques have significantly improved the performance of Transformers on various datasets [11]. In this paper, we introduce a diffusion-based model called EDADepth, an enhanced data augmentation-based monocular depth estimation.

In EDADepth, we created a diffusion-based pipeline that does not use extra training data, following the footsteps of ECoDepth[12]. Our pipeline enhances the input image through data augmentation. The original indoor NYU–Depth V2 [13] dataset with low image quality has been fed to a pre-trained Swin2SR model [14] for obtaining the enhanced dataset. From the input data, to extract text-embeddings, pre-trained ViT [15] and CLIP [16] models are widely used. Our model introduces a novel idea of using a pre-trained BEiT semantic segmentation model [17] for extracting detailed text embeddings as summarized in Figure 2. For our experiments and visualization, we used both the indoor (NYU–Depth V2) [13] and outdoor (KITTI) [18] datasets. Following recent works [12], [19], the evaluation metrics are absolute relative error (REL) and root mean squared error (RMSE), the average error log_{10} between the predicted depth and the actual depth and threshold accuracy δ_n.

The key contributions of this work are three-fold:

- We propose a novel method to enhance the input images to improve the estimated depth map. The enhanced input is used for semantic context extraction.
- We adopt a BEiT semantic segmentation model to extract the semantic context for creating text embeddings. We employ the BLIP-2 tokenizer as a novel way to create text embedding tokens from the extracted semantic context.
- We provide both qualitative and quantitative evaluations on two popular datasets NYUv2[13] and KITTI [18] to demonstrate the effectiveness of our pipeline.

II. Related Works

A. Monocular Depth Estimation

Over the last decade, several methods [20], [21], [22], [23], [24], [25], [26] have been proposed addressing Monocular Depth Estimation (MDE). MDE using supervised [27] and self-supervised learning [28] are among the recent works. The first MDE challenge organized at WACV 2023 showcased the work of Spencer et al. [29]. Several participants at the challenge outperformed the baseline on the SYNS-Patches [30] dataset. Few among the teams that implemented self-supervised learning models were team OPDAI whose MDE model was based on ConvNeXt-B [31] and HRDepth [32] models. Team z.suri based their MDE model on ConvNeXt [31] and DiffNet [33] models and team MonoViT's [34] model was based on MPViT [35] model. ZoEDepth [36] introduced a generalized and robust method for MDE using zero-shot transfer knowledge.

B. Diffusion-based MDE models

Recently, diffusion model [10] has attracted more notable advances in estimating depth [37], [38], [39], [40], [41] because of its pre-trained features. Due to their nature of intentionally adding noise into the data during the forward process and trying to restore the original data during the reverse process, they have been extensively used to extract high-level features for MDE.

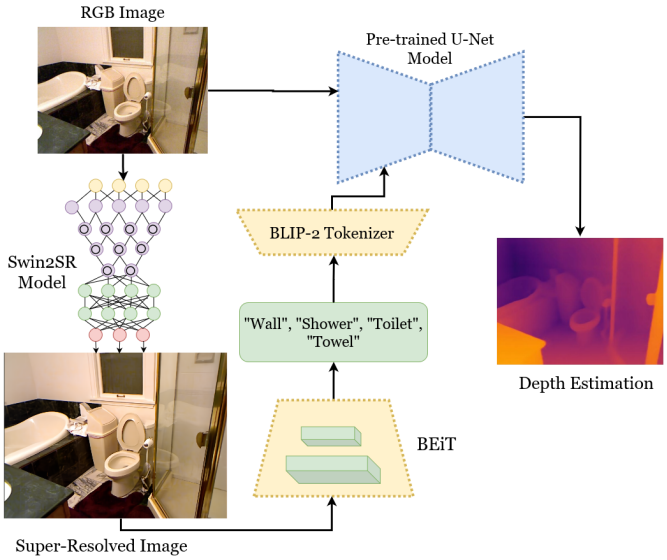

Fig. 2: In EDADepth, the raw RGB input image is enhanced using a Swin2SR model. BEiT model extracts detailed semantic context from the enhanced image and passes it to a BLIP-2 tokenizer for tokens. These text embedding tokens are fed to a pre-trained U-Net model to estimate depth.

VPD [37] uses the U-Net architecture for depth estimation and reference segmentation. EVP [42] and MetaPrompts [40] enhance the existing VPD model by adding layers to create effective text embeddings. Recently, ECoDepth [12] introduced the concept of using embeddings from a pre-trained ViT for detailed semantic information extraction. Existing MDE models [37], [12], [40], [42], [38] use CLIP [16] text tokenizer for generating text embedding tokens from semantic context. Our approach differs from traditional image-to-text caption generators such as CLIP [16] used in VPD [37], providing a more informative and precise representation of the input images. Likewise, PatchFusion [43] was the first to enhance low-quality input dataset as a data augmentation step in the MDE pipeline. In the same direction, we for the first time, use the Swin2SR model to enhance the input dataset. To effectively extract the semantic context from the enhanced dataset, we introduce a novel idea of using the BEiT semantic segmentation [17] model. Additionally, to extract tokens from the text embeddings effectively, we propose applying BLIP-2 [44] tokenizer.

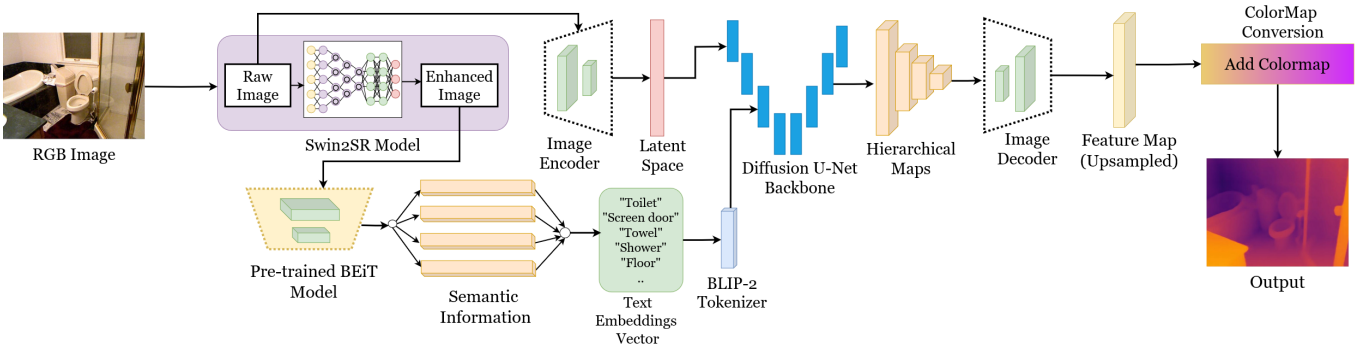

Fig. 3: **EDADepth model framework**. The architecture integrates a Swin2SR model to process raw RGB inputs, producing enhanced images for the text embedding module. It utilizes the BEiT semantic segmentation model for a segmentation-based, self-supervised text embedding process that generates a vector of text embeddings. These vectors are then fed into the U-Net model via the BLIP-2 tokenizer. The model follows a forward-reverse denoising process to generate an estimated depth map.

III. Methods

A. Diffusion Models Overview

Diffusion models [10] are generative models implemented by adding noise to the input and aiming to reconstruct the original input by learning the reverse denoising process. The diffusion model implemented in this project is stable diffusion, a text-to-image latent diffusion model [10]. The Stable Diffusion model consists of four key components: Encoder (E), Conditional denoising auto-encoder (ϵ_0), Language encoder (τ_θ), and decoder (D). The diffusion process is modeled as follows:

$$z_t \sim \mathcal{N}(\sqrt{\alpha_t} z_{t-1}, (1-\alpha_t)\boldsymbol{I}) \quad (1)$$

where z_t is random variable at time t, α_t is fixed coefficient representing the noise schedule, and $\mathcal{N}(z, \mu, \sigma)$ represents the normal distribution.

The encoder (E) and decoder (D) are trained before the ϵ_0, such that $D(E(x)) = \hat{x} \approx x$. ϵ_0 is implemented using U-Net as a pre-trained forward process (using the LAION-5B dataset [45]) and we train the reverse process for depth estimation. ϵ_0 of the latent diffusion model is trained to minimize the loss given by:

$$L_{LDM} := \mathbb{E}_{E(x),y,\epsilon\sim\mathcal{N}(0,1),t} ||\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))||_2^2 \quad (2)$$

where z_t is calculated using Equation 1.

From equation 1, we know that the diffusion model is processed as Markov, making it a regression problem, which can be used to model the distribution $p(y|x)$, where y is the output depth and x is its corresponding input image. Since we already have a pre-trained model from stable diffusion, the model ϵ_0 can be used to predict the density function gradient, $\nabla_{z_t} log p(z_t|C)$. The distribution $p(y|x)$ can be further modeled as:

$$p_\lambda(y|x,\mathcal{T}) = p_{\lambda_4}(y|z_0) p_{\lambda_3}(z_0|z_t,\mathcal{C}) p_{\lambda_2}(z_t|x) p_{\lambda_1}(\mathcal{C},x) \quad (3)$$

where, $p_{\lambda_1}(\mathcal{C},x) = p(\mathcal{C}|\mathcal{T})p(\mathcal{T}|x)$.

$\mathcal{T}$ is used to denote the textual embeddings obtained from the BEiT model (discussed in the next section). The pre-trained transformer model implements the distribution $p(\mathcal{T}|x)$, and through the learnable embeddings from the BEiT model, $p(\mathcal{C}|\mathcal{T})$ is implemented. The distribution $p(z_0|x)$ is implemented using the encoder (E). Likewise, the distribution $p(z_0|z_t,\mathcal{C})$ is implemented via the U-Net model [46] to extract hierarchical feature maps. Finally, the distribution $p(y|z_0)$ generates the depth map from the hierarchical feature maps.

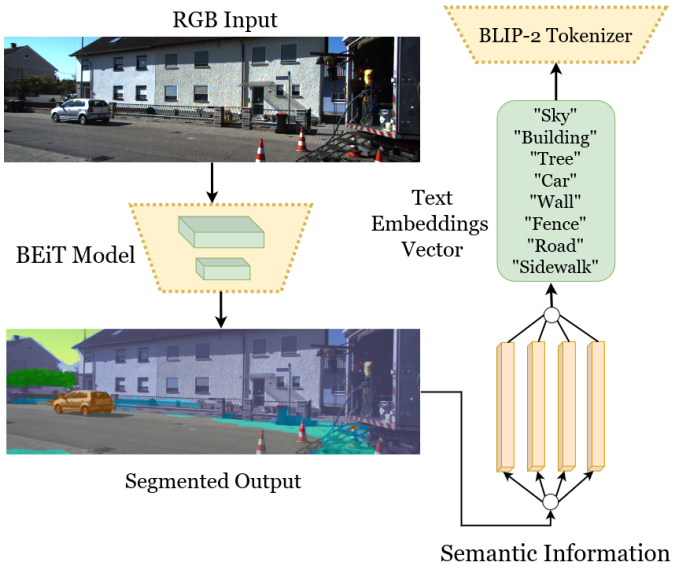

Fig. 4: Text-embedding extraction using BEiT model.

B. BEiT model for Text-Embeddings

We adopt the BEiT model, a self-supervised semantic segmentation model trained on ImageNet-21K from the ImageNet dataset [47] and fine-tuned on the ADE20K dataset [47] to extract semantic context from enhanced input images. Figure 4 describes the image semantic segmentation pipeline of the BEiT model that extracts the semantic context of the image into a 150-dimensional logit vector. The semantic context is fed to a multilayered perceptron equipped with GELU[48] to generate text embeddings (100-dimensional logit vectors). These vectors are now passed to the BLIP-2 tokenizer [44].

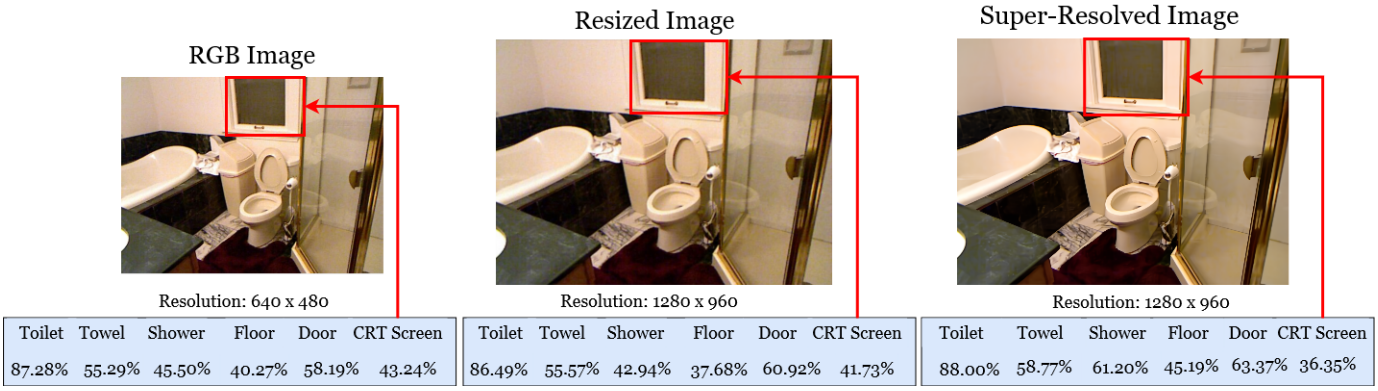

Fig. 5: Probabilities of the predicted semantic classes for the original, the resized, and the super-resolved images.

Our model then performs a forward-reverse denoising process to extract high-level knowledge based on the learnable embeddings.

C. Enhanced Data Augmentation

Depth estimation models are sensitive to low-quality inputs [43], which can lead to loss of features. When a low-quality input image is supplied to the BEiT text embedding model, it results in incomplete semantic information or knowledge and improper estimation of depth for various objects. Hence, we propose an enhanced data augmentation through a pre-trained Swin2SR model [14] to enhance the quality of the input image. Figure 5 compares the probability of BEiT predicted semantic classes between the original images resized using bilinear interpolation and the superresolved images. There is a noticeable improvement in the probability of the super-resolved image. The model misclassifies the "window-pane" class as the "CRT screen" class in the red-annotated portions of the images, likely because of their similar visual appearances. However, the predictive probability of the superresolved image is lower than that of the original image, indicating that the superresolved input enhances the accuracy of semantic classification while reducing false positives. In addition, the probabilities for many components of the resized image are lower than those of the original.

D. Overall EDADepth Architecture

Figure 3 shows the overall architecture of EDADepth model.
Input: As shown in figure 3, the input RGB image is fed to an image encoder for conversion into a latent space [49], which is then processed through the denoising U-Net. The same image is also provided to the Swin2SR model, which enhances the image by super-resolving it. The enhanced image is passed through the pre-trained BEiT model for semantic context extraction. This step transforms the input image into a sequence of visual tokens [17] for semantic contextual information extraction to segment the image. This contextual information, after linear transformation [50] is passed through the U-Net diffusion backbone (stable diffusion).
Stable Diffusion: The stable diffusion model enables image diffusion into the latent space to learn latent embedding through a variational autoencoder (VAE) [51]. The VAE transforms the input image into latent space for the U-Net model which considers different features of the input image in various dimensions. Likewise, learnable text embeddings are fed into the U-Net. Based on the image denoising process and the text-to-image generation process from text embeddings in the U-Net model, multiscale hierarchical feature maps [52] are generated and sent to the upsampling decoder [53].
Decoder: The Decoder [53] is designed to perform convolution-deconvolution [54] to upsample the feature maps. The decoder has a regression model with two convolution layers that generate the depth map from the feature maps. This depth map is colorized to better visualize the metric depth estimate of the RGB input image.

IV. Experimental Results

A. Datasets and Evaluation

The monocular depth estimation (MDE) models are trained and evaluated using the NYU Depth v2 and KITTI data sets. NYU Depth v2 [13] dataset consists of video sequences in indoor scenes with 24,231 2D 640×480 RGB images and the corresponding infrared-based depth maps have a depth range of 0.1 to 10 meters. KITTI Eigen-Split [20] dataset consists of video sequences in outdoor scenes containing 23,158 2D 1240×375 RGB images and its corresponding infrared-based depth maps having a depth range of 0.1 to 80 meters. The results are measured using RMSE (root mean squared error) and REL (absolute relative error) as the primary evaluation metrics. The average error log_{10} between the predicted depth a and the actual depth d, and the threshold accuracy δ_n which measures the % of pixels that satisfy the $\max(a_i/d_i, d_i/a_i) < 1.25^n$, where $n = 1, 2, 3$, are other common metrics as in Tables I, II. We also provide qualitative results in figures 6,7.

B. Implementaton Details

Our proposed EDADepth model uses the HuggingFace Stable-Diffusion-v1-5 checkpoint as the U-Net diffusion backbone. For super-resolution, we utilized a Swin2SR model from HuggingFace to upscale the resolution by 2x. Additionally, we incorporated the BEiT-based model for semantic segmentation.

TABLE I: Comparison of recent models on the NYUv2 Dataset. The recent models are categorized into non-diffusion-based and diffusion-based monocular depth estimation (MDE) models. Diffusion-based MDE models are further divided into those trained with extra training data (ETD) and those without. **Bold metrics** indicate SOTA performance, and *italic metrics* indicate the second-best performance. The row with a light gray fill represents the performance of our model.

Method	Venue	RMSE↓	REL↓	log_{10}↓	δ_1↑	δ_2↑	δ_3↑	extra training data
Non-Diffusion-Based								
Eigen et al. [20]	NIPS'14	0.641	0.158	-	0.769	0.950	0.988	×
DORN [21]	CVPR'18	0.509	0.115	0.051	0.828	0.965	0.992	×
GeoNet [22]	TPAMI'20	0.569	0.128	0.057	0.834	0.960	0.990	×
SharpNet [55]	ICCVW'21	0.502	0.139	0.047	0.836	0.966	0.993	×
Yin et al.[56]	CVPR'21	0.416	0.108	0.048	0.875	0.976	0.994	×
BTS[23]	Arxiv'19	0.392	0.110	0.047	0.885	0.978	0.994	×
ASN [57]	ICCV'21	0.377	0.101	0.044	0.890	0.982	0.996	×
TransDepth [58]	ICCV'21	0.365	0.106	0.045	0.900	0.983	0.996	×
AdaBins[24]	CVPR'21	0.364	0.103	0.044	0.903	0.984	0.997	×
DPT [25]	ICCV'21	0.357	0.110	0.045	0.904	0.988	0.998	✓
P3Depth[59]	CVPR'22	0.356	0.104	0.043	0.898	0.981	0.996	×
NeWCRFs[26]	CVPR'22	0.334	0.095	0.041	0.922	0.992	0.998	×
Localbins [60]	ECCV'22	0.357	0.099	0.042	0.907	0.987	0.998	×
DepthFormer[61]	ArXiv'22	0.329	0.094	0.040	0.923	0.989	0.997	×
PixelFormer[62]	WACV'23	0.322	0.090	0.039	0.929	0.991	0.998	×
WorDepth [19]	CVPR'24	0.317	0.088	0.038	0.932	0.992	0.998	×
MIM [63]	CVPR'23	0.287	0.083	0.035	0.949	0.994	0.999	×
ZoeDepth [36]	ArXiv'23	0.270	0.075	0.032	0.955	0.995	0.999	✓
Diffusion-Based (with extra training data)								
VPD [37]	ICCV'23	0.254	0.069	0.030	0.964	0.995	0.999	✓
TADP [38]	CVPR'24	0.225	0.062	0.027	0.976	0.997	0.999	✓
Marigold [39]	CVPR'24	0.224	**0.055**	**0.024**	0.964	0.991	0.998	✓
MetaPrompts [40]	ArXiv'23	0.223	0.061	0.027	0.976	0.997	0.999	✓
Diffusion-Based (without extra training data)								
DDP [41]	ICCV'23	0.329	0.094	0.040	0.921	0.990	0.999	×
ECoDepth [12]	CVPR'24	**0.218**	*0.059*	0.026	**0.978**	0.997	0.999	×
EDADepth(ours)	ICMLA'24	*0.223*	0.061	*0.026*	*0.977*	**0.998**	**1.000**	×

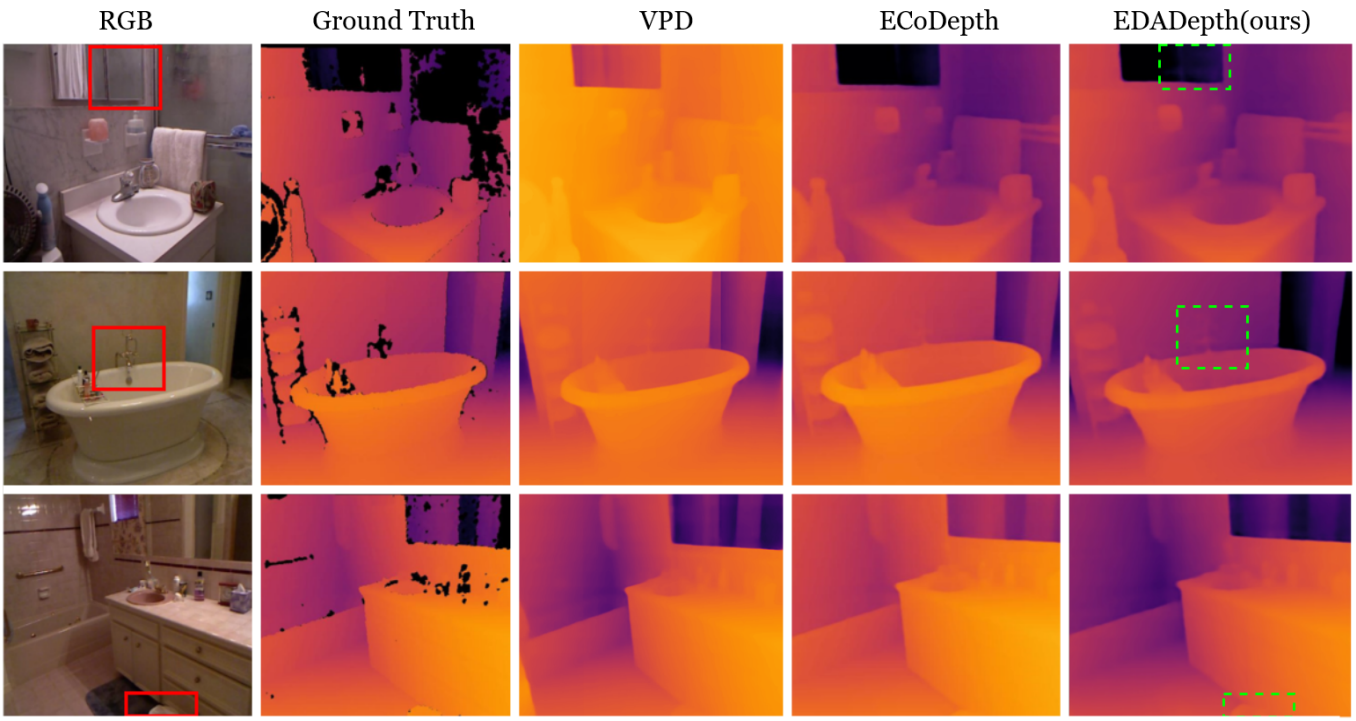

Fig. 6: Comparison of different diffusion-based MDE models with their test samples. The annotated green box denotes the area where the visualization of the output from our model outperforms the visualization of ECoDepth [12], the current diffusion-based SOTA. Zoom-in for better visibility.

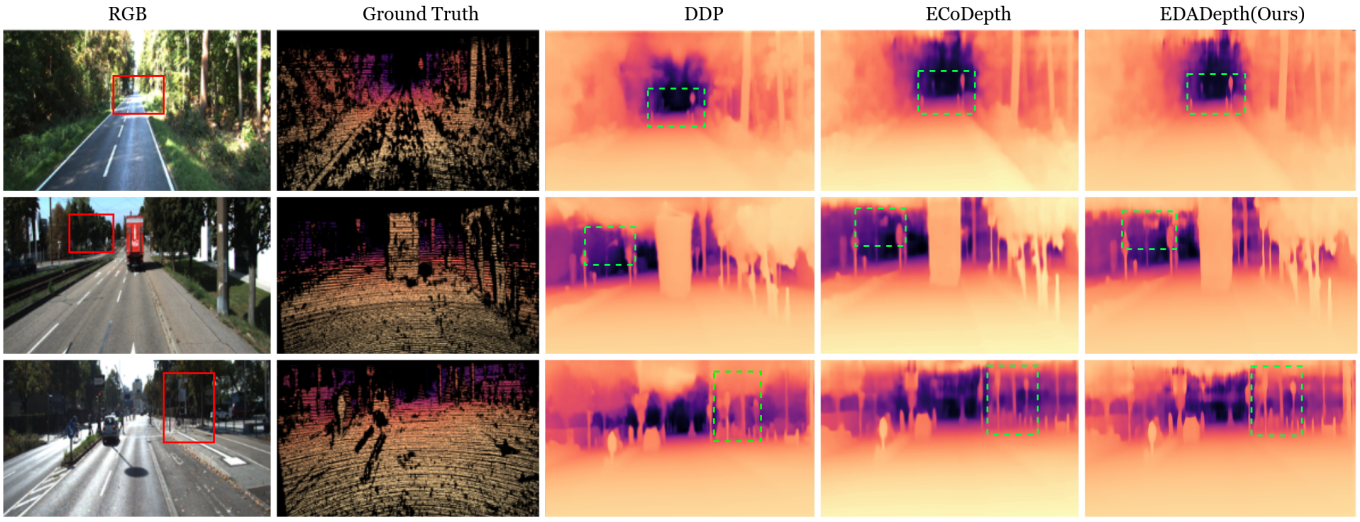

Fig. 7: Comparison of various diffusion-based monocular depth estimation (MDE) models on KITTI Eigen-Split [20] test samples, all trained without additional data. The annotated green box highlights the area where our model's depth estimation surpasses that of the other models. Zoom-in for better visibility.

TABLE II: Comparision of models on KITTI Eigen-Split [20] Dataset. The **bold metrics** represent SOTA, and the *italic metrics* represent the second-best performance. The row with a light gray fill represents the performance of our model, EDADepth.

Method	Venue	RMSE↓	REL↓	$RMSE_{log}$↓	δ_1↑	δ_2↑	δ_3↑	extra training data
Non-Diffusion-Based								
Eigen et al. [20]	NIPS'14	6.3041	0.203	0.282	0.702	0.898	0.967	×
DORN [21]	CVPR'18	2.727	0.072	0.120	0.932	0.984	0.994	×
Yin et al. [56]	CVPR'21	3.258	0.072	0.117	0.938	0.990	0.998	×
BTS[23]	Arxiv'19	2.756	0.059	0.096	0.885	0.978	0.994	×
TransDepth[58]	ICCV'21	2.755	0.064	0.098	0.956	0.994	0.999	×
AdaBins[24]	CVPR'21	2.960	0.067	0.088	0.949	0.992	0.998	×
DPT[25]	ICCV'21	2.573	0.060	0.092	0.959	0.995	0.996	✓
P3Depth[59]	CVPR'22	2.842	0.071	0.103	0.953	0.993	0.998	×
NeWCRFs[26]	CVPR'22	2.129	0.052	0.077	0.974	0.997	0.999	×
DepthFormer[61]	ArXiv'22	2.143	0.052	0.079	0.975	0.997	0.999	×
PixelFormer[62]	WACV'23	2.081	0.051	0.077	0.976	0.997	0.999	×
ZoeDepth [36]	ArXiv'23	2.440	0.054	0.083	0.970	0.996	0.999	✓
WorDepth [19]	CVPR'24	2.039	0.049	0.074	0.932	0.992	0.998	×
MIM [63]	CVPR'23	**1.966**	0.050	0.075	0.977	0.998	1.000	×
Diffusion-Based (with extra training data)								
Marigold [39]	CVPR'24	3.304	0.099	0.138	0.916	0.987	0.996	✓
MetaPrompts [40]	ArXiv'23	*1.928*	**0.047**	**0.071**	**0.981**	**0.998**	1.000	✓
Diffusion-Based (without extra training data)								
DDP [41]	ICCV'23	2.072	0.050	0.076	0.975	0.997	0.999	×
ECoDepth [12]	CVPR'24	2.039	*0.048*	*0.074*	*0.979*	0.998	1.000	×
EDADepth (ours)	ICMLA'24	2.070	0.051	0.077	0.978	*0.997*	**1.000**	×

We trained using 8 NVIDIA H100 GPUs [64] for 25 epochs, with a total batch size of 32. Our model did not use additional training data, but relied solely on the original NYUv2Depth [13] and KITTI Eigen-Split [20] datasets.

C. Quantitative Results

As shown in Table I for the NYUv2 test set, among the stable diffusion-based models (both with and without additional training data), our model achieved the second-best results in metrics such as RMSE, log10 and δ_1. Our model achieved SOTA results for δ_2 and δ_3 among diffusion-based models. For the KITTI Eigen-Split dataset, Table II shows that our model achieved SOTA for δ_3, indicating better visualizations through precise depth estimation than other diffusion-based models. These results support our model's ability to generate precise depth maps and improve visualization in outdoor datasets.

D. Qualitative Results

Figure 6 compares our model with recent SOTA models trained and evaluated on the NYUv2 Test Dataset. The three rows showcase our model's superior depth estimation. The red boxes in the first column highlight regions of interest in RGB images, while the green boxes in the last column show where our model outperforms existing diffusion-based methods. In the top row, a green dotted box marks a mirror accurately captured by our model. In the Our model correctly identifies a

"Faucet" in the second row depth map. Similarly, in the bottom row, the green dotted box highlights a "Towel" accurately included by our model.

Figure 7 illustrates our model performance in the KITTI eigen-split data set. We achieved results comparable to SOTA diffusion-based models. As demonstrated with the NYUv2 dataset, our model excels in depth estimation. The red boxes in the first column highlight objects where our model captures intricate details with greater precision. In the upper row, a red box indicates a street sign that is not visible in the original image because of its resolution, but our model accurately identifies it (last column). In the second row, a red box highlights a traffic light pole that our model captures with greater precision. Similarly, in the bottom row, our model more accurately represents street signs than other models.

V. Ablation Study

Extraction of text embeddings: We experimented using SOTA backbones to extract semantic context and generate text embedding vectors on the NYUv2 dataset. As shown in Table III, the BEiT-Base model backbone outperforms other models by providing better results on the given metrics.

TABLE III: Comparison of various models performing different tasks for obtaining semantic context to generate text embeddings. BEiT-Base performs better for the provided metrics. **ImgC: Image Classification**, **SSeg: Semantic Segmentation**

Model	Task	RMSE↓	REL↓	log_{10}↓
SwinV2-Base [65]	ImgC	0.227	0.062	0.027
SegFormer-Base [66]	SSeg	0.225	0.062	0.027
BEiT-Base [17]	SSeg	**0.223**	**0.061**	**0.026**

VI. Conclusion

In this paper, we proposed **EDADepth**, a novel method for monocular depth estimation using the enhanced super-resolution data enhancement technique. Firstly, we focused on superresolving the input data using a pre-trained Swin2SR model to improve the extraction of textual embeddings and the denoising process in the U-Net framework. Secondly, we employed a pre-trained BEiT Semantic Segmentation model to generate text embeddings to capture the semantic context from the input images. Third, we introduce the use of BLIP-2 as a tokenizer. Finally, we conducted extensive experiments on the NYUDepthv2 and KITTI Eigen-split datasets, demonstrating the effectiveness of our method. Our quantitative results show that our model achieves RMSE and REL values comparable to the current SOTA models while achieving SOTA on δ_3 values. From our qualitative results, it is evident that EDADepth competes closely with diffusion-based (both with and without using extra training data) SOTA models, particularly in enhancing the generation of visually detailed depth estimation.

References

[1] Michael Zollhöfer, Patrick Stotko, Andreas Görlitz, Christian Theobalt, Matthias Nießner, Reinhard Klein, and Andreas Kolb. State of the art on 3d reconstruction with rgb-d cameras. In *Computer graphics forum*, pages 625–652. Wiley Online Library, 2018. 1

[2] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Antonio Torralba, and Aude Oliva. Places: An image database for deep scene understanding. *arXiv preprint arXiv:1610.02055*, 2016. 1

[3] Yue Ming, Xuyang Meng, Chunxiao Fan, and Hui Yu. Deep learning for monocular depth estimation: A review. *Neurocomputing*, 438:14–33, 2021. 1

[4] Zhengyou Zhang. Determining the epipolar geometry and its uncertainty: A review. *International journal of computer vision*, 27:161–195, 1998. 1

[5] Ashutosh Saxena, Jamie Schulte, Andrew Y Ng, et al. Depth estimation using monocular and stereo cues. In *IJCAI*, volume 7, pages 2197–2203, 2007. 1

[6] Yuan Tian and Xiaodong Hu. Monocular depth estimation based on a single image: a literature review. In Zhigeng Pan and Xinhong Hei, editors, *Twelfth International Conference on Graphics and Image Processing (ICGIP 2020)*, volume 11720, page 117201Z. International Society for Optics and Photonics, SPIE, 2021. 1

[7] Vasileios Arampatzakis, George Pavlidis, Nikolaos Mitianoudis, and Nikos Papamarkos. Monocular depth estimation: A thorough review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2396–2414, 2024. 1

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[9] Shuangquan Zuo, Yun Xiao, Xiaojun Chang, and Xuanhong Wang. Vision transformers for dense prediction: A survey. *Knowledge-Based Systems*, 253:109552, 2022. 1

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2, 3

[11] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers, 2022. 2

[12] Suraj Patni, Aradhye Agarwal, and Chetan Arora. Ecodepth: Effective conditioning of diffusion models for monocular depth estimation. *arXiv preprint arXiv:2403.18807*, 2024. 2, 5, 6

[13] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 2, 4, 6

[14] Marcos V. Conde, Ui-Jin Choi, Maxime Burchi, and Radu Timofte. Swin2sr: Swinv2 transformer for compressed image super-resolution and restoration, 2022. 2, 4

[15] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision, 2020. 2

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. 2

[17] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers, 2022. 2, 4, 7

[18] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 2

[19] Ziyao Zeng, Daniel Wang, Fengyu Yang, Hyoungseob Park, Yangchao Wu, Stefano Soatto, Byung-Woo Hong, Dong Lao, and Alex Wong. Wordepth: Variational language prior for monocular depth estimation, 2024. 2, 5, 6

[20] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network, 2014. 2, 4, 5, 6

[21] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation, 2018. 2, 5, 6

[22] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose, 2018. 2, 5

[23] 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, 2021. 2, 5, 6

[24] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *2021 IEEE/CVF Conference*

on Computer Vision and Pattern Recognition (CVPR). IEEE, June 2021. 2, 5, 6

[25] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction, 2021. 2, 5, 6

[26] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. New crfs: Neural window fully-connected crfs for monocular depth estimation, 2022. 2, 5, 6

[27] Zhenyu Li, Zehui Chen, Xianming Liu, and Junjun Jiang. Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation. *Machine Intelligence Research*, 20(6):837–854, 2023. 2

[28] Stefano Gasperini, Nils Morbitzer, HyunJun Jung, Nassir Navab, and Federico Tombari. Robust monocular depth estimation under challenging conditions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8177–8186, October 2023. 2

[29] Jaime Spencer, C. Stella Qian, Chris Russell, Simon Hadfield, Erich Graf, Wendy Adams, Andrew J. Schofield, James H. Elder, Richard Bowden, Heng Cong, Stefano Mattoccia, Matteo Poggi, Zeeshan Khan Suri, Yang Tang, Fabio Tosi, Hao Wang, Youmin Zhang, Yusheng Zhang, and Chaoqiang Zhao. The monocular depth estimation challenge. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 623–632, January 2023. 2

[30] Jaime Spencer, Chris Russell, Simon Hadfield, and Richard Bowden. Deconstructing self-supervised monocular reconstruction: The design decisions that matter. *arXiv preprint arXiv:2208.01489*, 2022. 2

[31] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s, 2022. 2

[32] Xiaoyang Lyu, Liang Liu, Mengmeng Wang, Xin Kong, Lina Liu, Yong Liu, Xinxin Chen, and Yi Yuan. Hr-depth: High resolution self-supervised monocular depth estimation, 2020. 2

[33] Juhung Park, Woojin Jung, Eun-Jung Choi, Se-Hong Oh, Dongmyung Shin, Hongjun An, and Jongho Lee. Diffnet: Diffusion parameter mapping network generalized for input diffusion gradient schemes and bvalues, 2021. 2

[34] Chaoqiang Zhao, Youmin Zhang, Matteo Poggi, Fabio Tosi, Xianda Guo, Zheng Zhu, Guan Huang, Yang Tang, and Stefano Mattoccia. Monovit: Self-supervised monocular depth estimation with a vision transformer. In *2022 International Conference on 3D Vision (3DV)*. IEEE, September 2022. 2

[35] Youngwan Lee, Jonghee Kim, Jeff Willette, and Sung Ju Hwang. Mpvit: Multi-path vision transformer for dense prediction, 2021. 2

[36] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth, 2023. 2, 5, 6

[37] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception, 2023. 2, 5

[38] Neehar Kondapaneni, Markus Marks, Manuel Knott, Rogerio Guimaraes, and Pietro Perona. Text-image alignment for diffusion-based perception. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13883–13893, June 2024. 2, 5

[39] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 5, 6

[40] Qiang Wan, Zilong Huang, Bingyi Kang, Jiashi Feng, and Li Zhang. Harnessing diffusion models for visual perception with meta prompts. *arXiv preprint arXiv:2312.14733*, 2023. 2, 5, 6

[41] Yuanfeng Ji, Zhe Chen, Enze Xie, Lanqing Hong, Xihui Liu, Zhaoqiang Liu, Tong Lu, Zhenguo Li, and Ping Luo. Ddp: Diffusion model for dense visual prediction, 2023. 2, 5, 6

[42] Mykola Lavreniuk, Shariq Farooq Bhat, Matthias Müller, and Peter Wonka. Evp: Enhanced visual perception using inverse multi-attentive feature refinement and regularized image-text alignment, 2023. 2

[43] Zhenyu Li, Shariq Farooq Bhat, and Peter Wonka. Patchfusion: An end-to-end tile-based framework for high-resolution monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10016–10025, 2024. 2, 4

[44] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 2, 3

[45] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. 3

[46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 3

[47] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[48] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), 2023. 3

[49] Yang Liu, Eunice Jun, Qisheng Li, and Jeffrey Heer. Latent space cartography: Visual analysis of vector space embeddings. In *Computer graphics forum*, pages 67–78. Wiley Online Library, 2019. 4

[50] Yujia Bao and Theofanis Karaletsos. Contextual vision transformers for robust representation learning. *arXiv preprint arXiv:2305.19402*, 2023. 4

[51] Carl Doersch. Tutorial on variational autoencoders, 2021. 4

[52] Koray Kavukcuoglu, Pierre Sermanet, Y-Lan Boureau, Karol Gregor, Michaël Mathieu, Yann Cun, et al. Learning convolutional feature hierarchies for visual recognition. *Advances in neural information processing systems*, 23, 2010. 4

[53] Guangli Ren, Wenjie Geng, Peiyu Guan, Zhiqiang Cao, and Junzhi Yu. Pixel-wise grasp detection via twin deconvolution and multi-dimensional attention. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 4

[54] Hu Chen, Yi Zhang, Mannudeep K Kalra, Feng Lin, Yang Chen, Peixi Liao, Jiliu Zhou, and Ge Wang. Low-dose ct with a residual encoder-decoder convolutional neural network. *IEEE transactions on medical imaging*, 36(12):2524–2535, 2017. 4

[55] Michaël Ramamonjisoa and Vincent Lepetit. Sharpnet: Fast and accurate recovery of occluding contours in monocular depth estimation, 2019. 5

[56] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image, 2020. 5, 6

[57] Xiaoxiao Long, Cheng Lin, Lingjie Liu, Wei Li, Christian Theobalt, Ruigang Yang, and Wenping Wang. Adaptive surface normal constraint for depth estimation, 2021. 5

[58] Guanglei Yang, Hao Tang, Mingli Ding, Nicu Sebe, and Elisa Ricci. Transformer-based attention networks for continuous pixel-wise prediction, 2021. 5, 6

[59] Vaishakh Patil, Christos Sakaridis, Alexander Liniger, and Luc Van Gool. P3depth: Monocular depth estimation with a piecewise planarity prior, 2022. 5, 6

[60] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Localbins: Improving depth estimation by learning local distributions, 2022. 5

[61] Zhenyu Li, Zehui Chen, Xianming Liu, and Junjun Jiang. Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation. *Machine Intelligence Research*, 20(6):837–854, September 2023. 5, 6

[62] Ashutosh Agarwal and Chetan Arora. Attention attention everywhere: Monocular depth prediction with skip attention, 2022. 5, 6

[63] Zhenda Xie, Zigang Geng, Jingcheng Hu, Zheng Zhang, Han Hu, and Yue Cao. Revealing the dark secrets of masked image modeling, 2022. 5, 6

[64] Jack Choquette. Nvidia hopper h100 gpu: Scaling performance. *IEEE Micro*, 43(3):9–17, 2023. 6

[65] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 7

[66] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers, 2021. 7