Real (True) depth estimation from indoor scenes, given a model (DL tool) for virtual depth estimation (TPA 10)

Sujay S, *ED20B065,* Amruth R Vardhan G, *MM16B003.*

Abstract

This project presents a novel approach for true depth estimation employing ZoeDepth, a state-of-the-art neural network utilizing relative depth to predict metric depth centers. Leveraging the NYU V2 dataset for indoor imagery, our innovation lies in integrating a segmentation head into the core model during training, augmenting semantic learning and cross-talk distillation. Inspired by X-Distill, this augmentation enhances depth prediction accuracy, promising improved self-supervised monocular depth estimation.

I. Introduction

IN the realm of computer vision, accurate depth estimation remains pivotal for various applications. This project delves into advancing true depth estimation using a neural network framework, employing ZoeDepth[1], a cutting-edge model renowned for its utilization of relative depth to predict metric depth centers. Focused on indoor image inputs, the NYU V2[3] dataset was instrumental in training this model. A novel contribution of this work involves the introduction of a segmentation head into the core model architecture, facilitating the assimilation of semantic information alongside cross-talk distillation. While this augmentation occurs solely during training, its inspiration draws from X-Distill[2], a seminal paper enhancing self-supervised monocular depth estimation.

The report encompasses an in-depth algorithmic description, elucidating the integration of the segmentation head and the fine-tuning of hyperparameters to optimize model performance. Metrics utilized for evaluation and comparison are discussed, highlighting the efficacy of the proposed approach in contrast to existing methodologies. Results gleaned from experimentation underscore the potential advancements in accurate depth estimation achieved through this novel model configuration.

A. Single Image Depth Estimation:

Single-image depth estimation (SIDE) in computer vision has seen significant advancements, primarily diverging into two branches: metric depth estimation (MDE) and relative depth estimation (RDE). MDE, emphasized in recent works, focuses on estimating depth in absolute physical units like meters. Its advantage lies in practical applicability across various domains such as mapping, navigation, and 3D reconstruction. However, MDE models often struggle with generalization across datasets due to varying depth scales, impacting their performance and limiting applicability beyond specific contexts.

Conversely, RDE, highlighted in recent studies, addresses the challenge of depth scale variations across diverse environments by emphasizing relative depth. Here, disparity serves as supervision, eliminating the need for consistent camera parameters and metric scale across datasets. RDE methods enable training on diverse scenes, even spanning 3D movies, fostering model generalizability across domains. However, the trade-off involves predicted depths lacking metric significance, thereby limiting their applicability in certain contexts despite their enhanced generalizability.

B. True Depth Estimation (From Relative Depth Estimate):

The model introduced in this report builds upon the state-of-the-art ZoeDepth[1] architecture, renowned for leveraging relative depth estimation by the core network to predict true metric depth. ZoeDepth's prowess lies in its ability to infer metric depth centers using relative depth cues, providing a pathway to accurate metric depth estimation.

Furthermore, an innovative addition to this model involves the integration of segmentation distillation. During training, a segmentation head is incorporated into the core network, enabling the model to simultaneously learn semantic information while undergoing cross-talk distillation. This augmentation, inspired by recent advancements in cross-task distillation[2] methodologies, aims to enhance the model's depth estimation accuracy.

By combining ZoeDepth's relative depth-based metric estimation with segmentation distillation, this model seeks to achieve a holistic understanding of the scene, incorporating both depth-related information and semantic context. This integration aims to improve the model's ability to discern spatial relationships while inferring accurate metric depth, promising advancements in single-image depth estimation capabilities.

II. ALGORITHMIC DESCRIPTION

The model architecture integrates ZoeDepth's[1] cutting-edge design with an augmented depth-to-segmentation network, facilitating the transformation of predicted relative depth into segmentation. This innovation fosters a symbiotic relationship between the segmentation ground truth maps and the depth network during training, enabling seamless knowledge transfer across tasks.

Moreover, our approach involves a consolidation of semantic classes within the NYU V2[3] dataset, condensing them into a concise group of 14 classes (using the mapping from [7] for model training. Leveraging the feature maps derived from the MiDaS[5] core's decoder, both the segmentation head (U-Net)[6] and the ZoeDepth's bin module for metric depth are employed. This strategic fusion introduces additional semantic context from the input image to the MiDaS[5] core, thereby enhancing the efficacy of the training process.

Notably, this augmentation doesn't impact inference time as the segmentation component remains inactive during inference. The visualisation of overall model architecture is shown in Fig 1. The total number of parameters of the model is 105.3M. Subsequent sections will delve into an in-depth exploration of each core component, elucidating their roles and functionalities within the model framework.

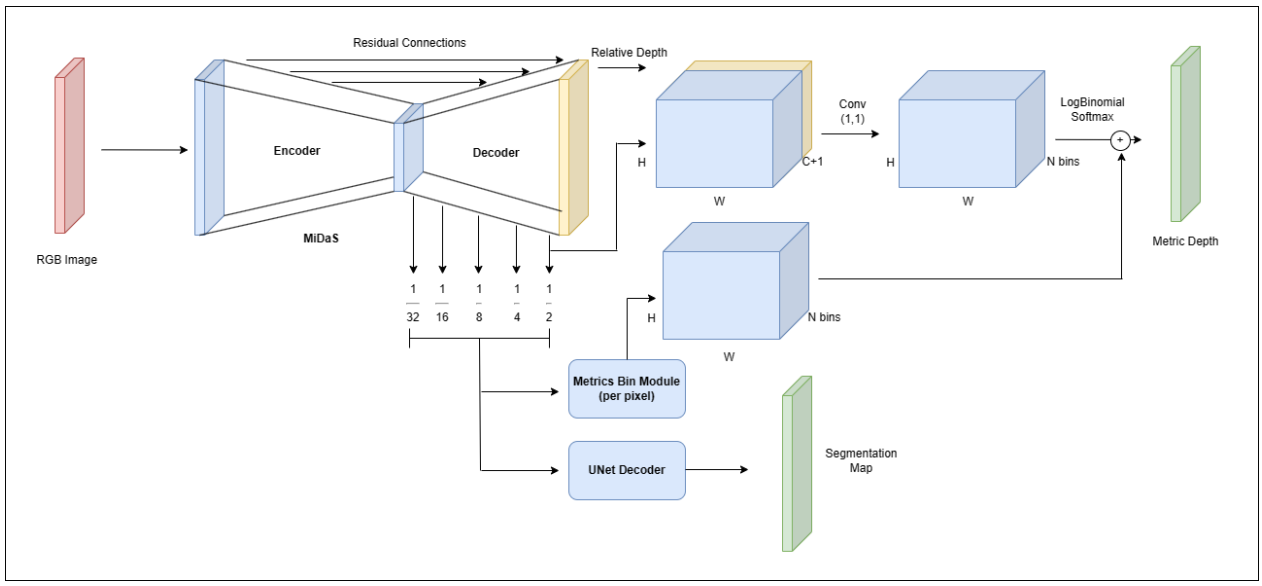

Fig. 1: Model Architecture

A. *About ZoeDepth:*

ZoeDepth[1], a state-of-the-art (SOTA) model, comprises essential components: the MiDaS[5] Core, responsible for extracting feature maps and predicting relative depth maps; a metric bins module, predicting bin centers corresponding to each pixel in the depth map; and a metric depth head module, utilizing log softmax probabilities from the bin centers to derive the final metric depth values for individual pixels. These pixel-wise probabilities serve as weightings for the predicted bin centers, contributing to the model's depth estimation process.

1) MiDaS Core: ZoeDepth adopts MiDaS's training approach for predicting relative depth. MiDaS[5] employs a scale and shift invariant loss function. The versatility of the MiDaS training strategy extends to various network architectures. The encoder-

decoder architecture serves a dual purpose: predicting relative depth maps and extracting feature maps from the input RGB image. In ZoeDepth, the authors leverage the DPT encoder-decoder structure[9] as the foundational model and however they replace the encoder with more recent transformer-based backbones[10]. They achieve the best results with large $BEiT_{384} - L$[10]. We choose the same DPT encoder-decoder structure, but however we opt for the $SwinV2_{384} - B$[8] transformer backbone for the encoder due to its commendable balance between performance and frames per second (FPS). The accompanying figure as shown in Fig 2 illustrates an overview of the different MiDaS models, with bubble sizes indicating parameter counts. Further specifics regarding the encoder-decoder are elucidated in Table I.

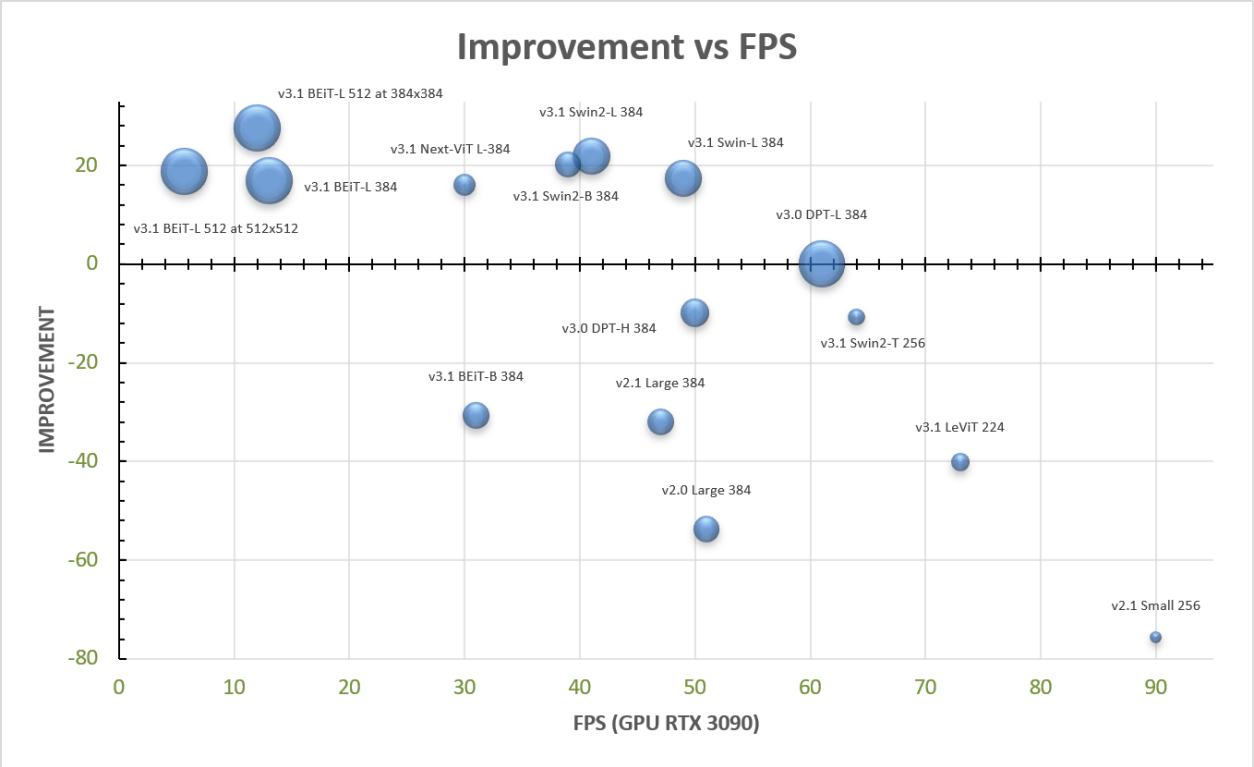

Fig. 2: Performance Improvement vs FPS

Model	Parameters	Image Resolution
SwinV2_B_384	102M	(384,384)

TABLE I: Core Backbone Details

2) Metrics Bin Module: The metric bins module within ZoeDepth[1] utilizes multiscale features from the MiDaS decoder as input, predicting bin centers crucial for metric depth estimation (refer to Fig 3). Initially, the module forecasts all bin centers at

the network bottleneck, refining them through subsequent decoder layers. This refinement occurs via attractor layers, fine-tuning the bin centers' positions within the depth interval. Their approach implements multiscale refinement by manipulating these centers, shifting them left or right along the depth spectrum. Leveraging the multiscale features, this predict a series of attractor points on the depth spectrum. Specifically, at the l-th decoder layer, an MLP processes pixel features, predicting a set of n_a attractor points $\{a_k : k = 1, ..., n_a\}$ for each pixel position. The adjusted bin center is denoted as $c_i' = c_i + \Delta c_i$, with the adjustment (Δc_i) calculated using the formula:

$$\Delta c_i = \sum_{k=1}^{n_a} \frac{a_k - c_i}{1 + \alpha |a_k - c_i|^\gamma}$$

Here, the hyperparameters α and γ dictate the strength of attraction. We adopt this variant of the attractor, known as the inverse attractor, based on its superior performance observed in ZoeDepth model creators' experiments. The default hyperparameters proved to be performing well and they are as follows as listed in Table II.

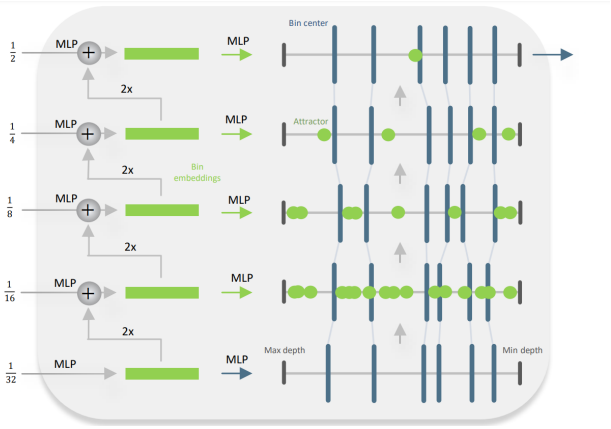

Fig. 3: Metric Bins Module

Hyperparameters for Bin Metric Module			
n_bins	**bin_embedding_dim**	**bin_centers_type**	**n_attractors**
64	128	softplus	[16, 8, 4, 1]
attractor_alpha	**attractor_gamma**	**attractor_kind**	**attractor_type**
1000	2	mean	inv

TABLE II: Hyperparameters for Bin Metric Module

3) Metric Depth Head: The calculation of the ultimate metric depth prediction involves a linear combination of bin centers, weighted by their respective probability scores. ZoeDepth[1] uses a binomial distribution, characterized by a single parameter

q that governs the mode's placement. Integrating the relative depth predictions with the decoder features, ZoeDepth predicts a 2-channel output (q-mode and t-temperature) from these decoder features to compute the probability score over the k-th bin center using the binomial probability. Apart from this, normalisation is carried out by taking logarithm of the probability to ensure numerical stability. Softmax gives the final normalised scores. Ultimately, these resulting probability scores, in conjunction with the bin centers obtained from the metric bins module, facilitate the derivation of the final depth.

B. Segmentation Cross-Talk Distillation

The feature outputs from the encoder-decoder network, scaled at different levels (1/32, 1/16, 1/8, 1/4, 1/2), serve as input features for the UNet[6] Decoder. The decoder network (expansive path) takes the feature map from the lower layer, up-converts it, crops and concatenates it with the encoder data of the same level, and then performs two 3×3 convolution layers followed by ReLU activation 1. We use UNet[6] encoder with 5 decoder layers. This decoder, encompassing approximately 3M parameters, generates the final logits for segmentation output. Upon computing the loss, Softmax is applied to derive class probabilities. The training involves feeding the decoder with inputs associated with 14 distinct classes. The following diagram as shown in Fig 4 illustrates the distillation task while training.

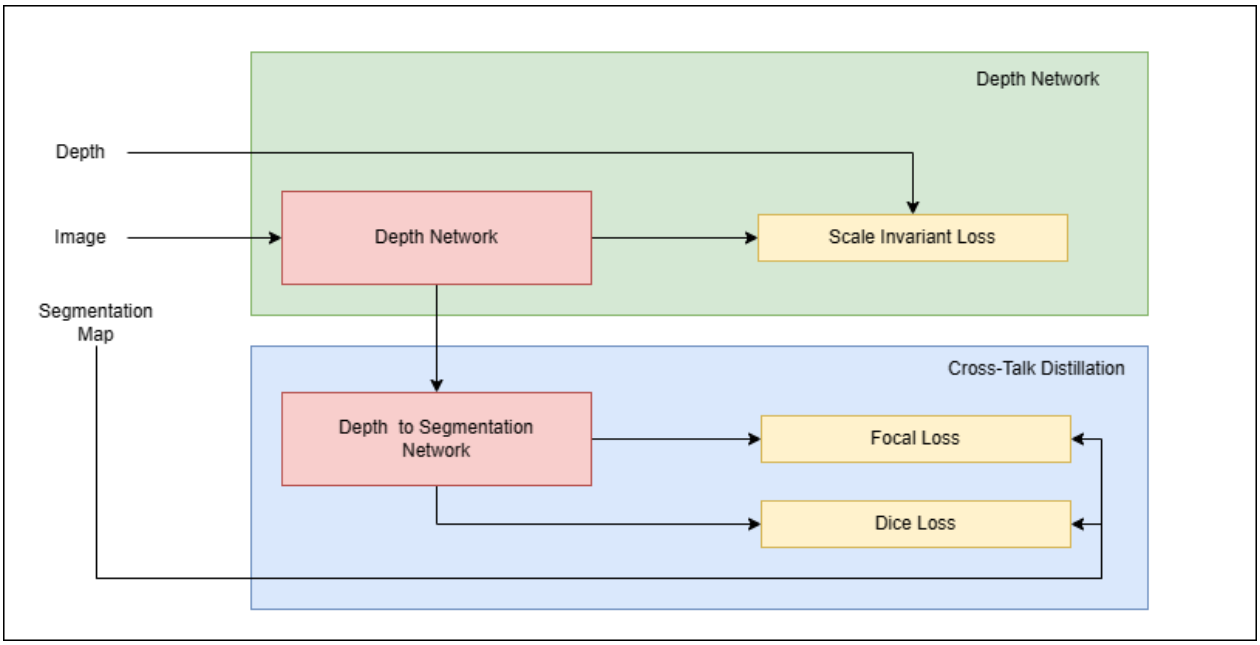

Fig. 4: Cross-Talk Distillation Network

C. Parameters Comparison

The following Table III shows the parameters comparison of our model with other SOTA models.

Model	Encoder	#Parameters
Adabins[11]	EfficientNet-B5	78M
NeWCRFs[12]	Swin-L	270M
ZoeD-M12-N (S2-B)[1]	Swin2-B	102M
ZoeD-M12-N (B-L)[1]	Beit-L	345M
Our model	**Swin2-B**	**105M**

TABLE III: Parameters comparison of our model with other SOTA models

D. Dataset

The datasets used in the experiments and the details are as lised in the Table IV. The choice of dataset for training, NYU V2[3], was deliberate due to its standing as the standard dataset for indoor images. This selection aligns with the project's exclusive focus on predicting metric depth for indoor scenes, where the maximum depth typically spans around 10 meters. NYU V2 presents an ideal range that encapsulates these indoor depth scenarios effectively. And iBims-1[4] is used for evaluation as unseen data along with NYU V2[3]. *The NYU-Depth V2 data set is comprised of video sequences from a variety of indoor scenes as recorded by both the RGB and Depth cameras from the Microsoft Kinect. It features 1449 densely labeled pairs of aligned RGB and depth images. iBims-1 is a top-tier RGB-D dataset crafted for testing single-image depth estimation (SIDE) methods. Captured using a DSLR and high-precision laser scanner, it offers high-quality images and accurate depth maps of indoor scenarios. Notable for minimal noise, clear depth transitions, no occlusion, and extensive depth ranges, iBims-1 contains 100 high and low-resolution RGB-D image pairs showcasing diverse interior scenes.*

Dataset	**Domain**	**Type**	**Seen in Training?**	**Train Samples**	**Eval Samples**	**Eval Depth (m)**	
						Min	**Max**
NYU V2[3]	Indoor	Real	Yes	1158	291	1e-3	10
iBims-1[4]	Indoor	Real	Yes	-	100	1e-3	10

TABLE IV: Overview of datasets used in training the model and evaluation of our model architecture

E. Loss

The training of the model involves two primary loss components: Depth loss and Segmentation loss. For depth supervision at the pixel level, we employ the scale-invariant log loss (L_{pixel}). On the other hand, the Segmentation loss is calculated by combining the dice loss and focal loss. We introduce a new hyperparameter, λ_{dis}, that allows us to adjust the weight on the segmentation loss. Our experiments suggest that the best choice of the hyperparameter λ_{dis} is 0.1.

Here are the mathematical definitions of the loss functions:

Depth loss (L_{pixel}):

$$g = \log(input + \alpha) - \log(target + \alpha)$$

$$Dg = Var(g) + beta \cdot (mean(g))^2$$

$$SILoss = 10 \cdot \sqrt{Dg}$$

Segmentation loss:

$$L_{segmentation} = \alpha \cdot DiceLoss + (1 - \alpha) \cdot FocalLoss$$

Final loss:

$$L_{final} = \lambda_{dis} \cdot L_{segmentation} + L_{pixel}$$

F. Training

The input image is resized to (384,384) since the input image size to the transformer backbone is 384 x 384. We use the pre-trained MiDaS[5] core as the starting point of our training process for relative depth predictions and then train the entire model (inclusing MiDaS core instead of freezing the weights). The model was trained for 100 epochs (which was split as 2 runs due to google collab runtime issue: first run to be 85 epochs and the final run to be 15 epochs) with a batch size of 3 owing to compute resource constraints. Adam optimiser along with cyclic learning rate scheduler is employed to aid the training process. For the AdamW optimizer, the learning rate is set to 0.000161 with a weight decay of 0.01. The cyclic learning rate scheduler is configured with parameters: div_factor = 1, final_div_factor = 10000, pct_start = 0.7, three_phase = false, and cycle_momentum = true. The labelled dataset NYU V2[3] containing 1449 samples was split into train and validation with 80-20% split. Validation was carried out every 25% of an epoch on validation data to ensure that model doesn't overfit. The best performing weights on the validation set is saved in .pt as model checkpoint. The final metric depth map is interpolated to the target size in order to compute the loss for training. Training was carried out on 1 X NVIDIA T4 GPU. The training rate was on an average 3.0 seconds per iteration.

III. Output

We observe that the segmentation distillation process seems to be effective in improving the prediction results as expected, and the model achieves SOTA results on NYU V2[3] and iBims-1[4]. We evaluate the metrics on both NYU V2[3] and iBims-1[4] and also plot the predicted depth maps for both NYU V2[3] and iBims[4] in the following sections for visualisation. The inference rate was on an average 1.3 seconds per iteration.

A. Evaluation Metrics

We evaluate in metric depth space **d** by computing the absolute relative error (REL) $= \frac{1}{M}\sum_{i=1}^{M}\frac{|d_i-\hat{d}_i|}{d_i}$, the root mean squared error (RMSE) $= \left[\frac{1}{M}\sum_{i=1}^{M}(d_i-\hat{d}_i)^2\right]^{\frac{1}{2}}$, the average $\log_{10}$ error $= \frac{1}{M}\sum_{i=1}^{M}|\log_{10} d_i - \log_{10}\hat{d}_i|$, and the threshold accuracy δ_n = % of pixels s.t. max $\left(\frac{d_i}{\hat{d}_i}, \frac{\hat{d}_i}{d_i}\right) < 1.25^n$ for n = 1, 2, 3, where d_i and $\hat{d}_i$ refer to ground truth and predicted depth at pixel i, respectively, and M is the total number of pixels in the image. We cap the evaluation depth at 10m for indoor datasets. The Table V depicts the results evaluated on NYU Depth V2[3] along with quantitative comparisons. And the Table VI shows the results evaluated on iBims-1[2] along with quantitative comparisons.

Model	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	REL $\downarrow$	RMSE $\downarrow$	$\log_{10} \downarrow$
Adabins[11]	0.903	0.984	0.997	0.103	0.364	0.044
NeWCRFs[12]	0.922	0.992	0.998	0.095	0.334	0.041
ZoeDepth (S2-B)[1]	0.927	0.992	0.999	0.090	0.313	0.038
ZoeDepth (B-L)[1]	0.955	0.995	0.999	0.075	0.270	0.032
Our model	**0.980**	**0.998**	**0.999**	**0.048**	**0.2**	**0.021**

TABLE V: Results on NYU Depth V2 (Quantitative Comparison)

Model	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	REL $\downarrow$	RMSE $\downarrow$	$\log_{10} \downarrow$
Adabins[11]	0.555	0.873	0.960	0.212	0.901	0.107
NeWCRFs[12]	0.548	0.884	0.979	0.206	0.861	0.102
ZoeDepth (B-L)[1]	0.658	0.947	0.985	0.169	0.711	0.083
Our model	**0.669**	**0.924**	**0.972**	**0.182**	**0.922**	**0.086**

TABLE VI: Results on iBims-1 (Quantitative Comparison)

B. Plots

The predicted metric depth maps from our model for the datasets NYU V2[3], iBims-1[4] are plotted and visualised in the form of inferno maps where the depth is in the range 0-10 metres. The relative depth predictions are obtained by a pretrained MiDas[5] core.

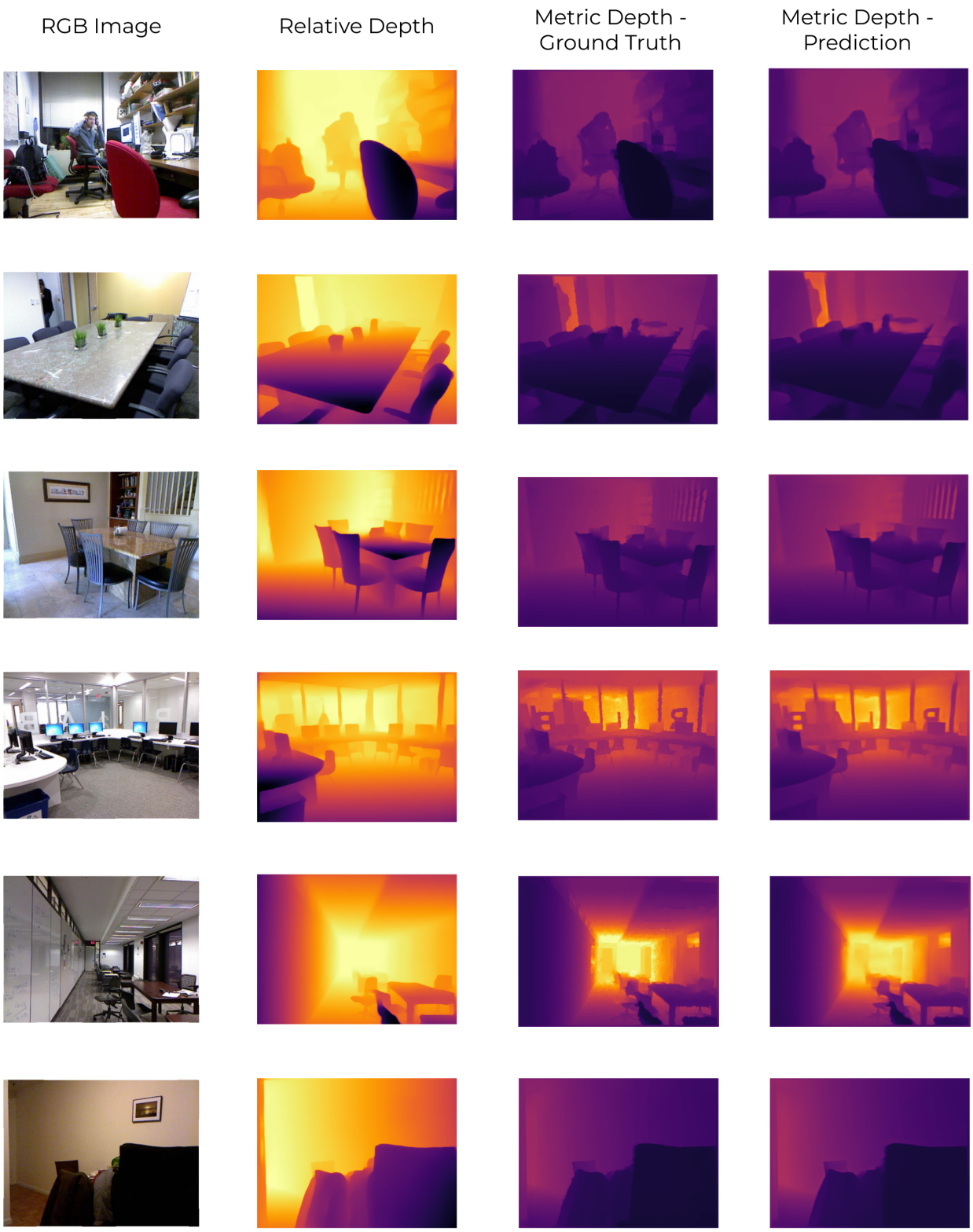

Fig. 5: RGB Image, Relative Depth, Metric Depth-Ground Truth, Metric Depth-Prediction (from left to right) for NYU V2 dataset

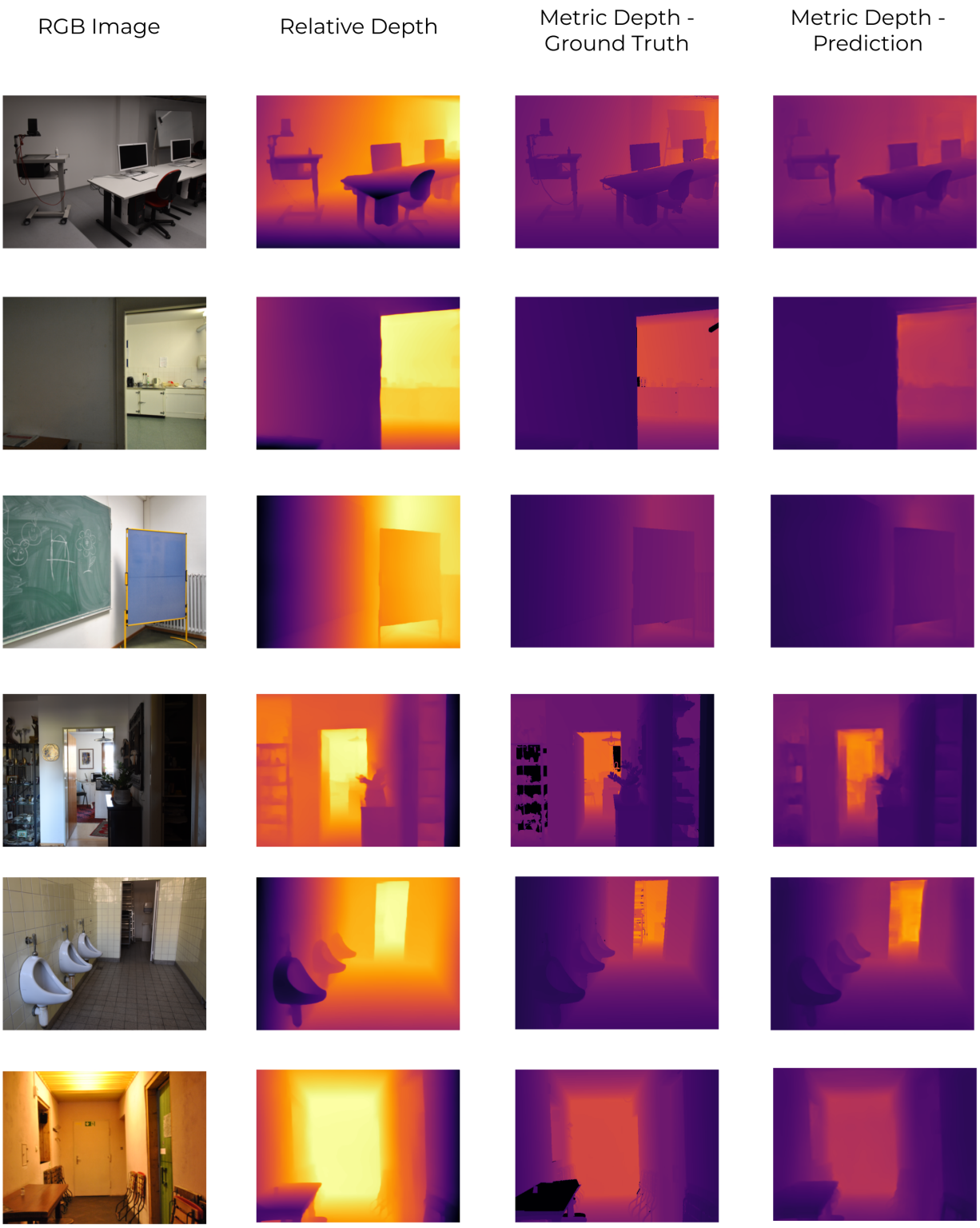

Fig. 6: RGB Image, Relative Depth, Metric Depth-Ground Truth, Metric Depth-Prediction (from left to right) for iBims dataset

IV. Conclusion

In conclusion, the integration of segmentation distillation into the ZoeDepth[1] model architecture showcased substantial enhancements in training efficiency and overall performance, achieving state-of-the-art results. Leveraging the Swin2-B[9] backbone with 102M parameters yielded promising outcomes, yet the potential for further improvements exists by exploring even more robust encoders like BeiT-L[10], paving the way for future advancements. The findings strongly support the notion that enriching the depth network with valuable semantic information significantly enriches the training process. Future endeavors could delve into meaningful class grouping for depth estimation, presenting an exciting avenue for exploration and refinement.

References

[1] S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller, *ZoeDepth: Zero-shot Transfer by Combining Relative and Metric Depth*, ArXiv preprint, arXiv:2302.12288, 2023.

[2] H. Cai, J. Matai, S. Borse, Y. Zhang, A. Ansari, and F. Porikli, *X-Distill: Improving Self-Supervised Monocular Depth via Cross-Task Distillation*, ArXiv preprint, arXiv:2110.12516, 2021.

[3] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus, *Indoor Segmentation and Support Inference from RGBD Images*, *ECCV*, 2012.

[4] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Korner, *Evaluation of CNN-based Single-Image Depth Estimation Methods*, In *Proceedings ECCV 2018 Workshops*, 2019.

[5] Rene Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun, *Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer*, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, *U-Net: Convolutional Networks for Biomedical Image Segmentation*, ArXiv preprint, arXiv:1505.04597, 2015.

[7] Ankur Handa, *nyuv2-meta-data*, GitHub repository, https://github.com/ankurhanda/nyuv2-meta-data, 2017.

[8] Rene Ranftl, Alexey Bochkovskiy, and Vladlen Koltun, *Vision Transformers for Dense Prediction*, ArXiv preprint, arXiv:2103.13413, 2021.

[9] 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.

[10] Hangbo Bao, Li Dong, and Furu Wei, *Beit: BERT pretraining of image transformers*, CoRR, abs/2106.08254, 2021.

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

[12] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan, *New CRFs: Neural window fully-connected CRFs for monocular depth estimation*, arXiv preprint arXiv:2203.01502, 2022.