Semantic-MD: Infusing Monocular Depth with Semantic Signals

Sapar Charyyev[1,*] Ankita Ghosh[1,*] Oliver Lemke[1,*] Zuria Bauer[1] Mihai Dusmanu[2]
[1]ETH Zurich [2]Microsoft MR & AI Lab Zurich
{scharyyev, anghosh, olemke}@ethz.ch

Abstract

Monocular depth estimation (MDE) plays a crucial role in numerous computer vision tasks, such as 3D reconstruction, scene understanding, and augmented reality. However, as MDE is an ill-posed problem, the incorporation of additional semantic cues can facilitate improved depth estimation. In this paper, we explore different techniques based on deep learning to leverage the rich semantic details present in an image for monocular depth estimation. First, we explore different ways of integrating semantic signals to the input in the form of semantic maps and borders. Second, we jointly estimate depth and semantic maps to exploit the complementary nature of these tasks. We conduct extensive ablation studies for both of our approaches with different semantic signals and loss functions and compare them with our encoder-decoder based baseline architecture. We validate our results quantitatively through various evaluation metrics, and qualitatively on HyperSim dataset. The code is made available at https://github.com/charyyev/semantic_md.

1. Introduction

Depth estimation is a fundamental task for scene understanding with application in various robotics and autonomous system tasks [3]. Monocular depth estimation (MDE) is a technique that enables the prediction of depth information from a single image. Unlike stereo or multi-view depth estimation methods, MDE is convenient for portability and cost-effective for large-scale data processing. However, MDE is an ill-posed [1] problem since a single image does not provide enough information to uniquely determine depth. To tackle this issue, MDE algorithms heavily rely on learning scene priors [28], which are pre-existing assumptions about the structure and characteristics of the environment which help in inferring depth.

Depth estimation and semantic segmentation are considered to be correlated [22] as depth discontinuities often align with semantic edges. Semantic maps also contain valuable information about the scene and its constituent objects. Including semantic signals in an MDE task can provide the important details required to improve depth estimations. In this paper, we experiment with different methods of infusing semantic signals into our MDE pipeline. Our motivation, as shown in Figure 1, is to help the network learn the scene priors required to make correct predictions. This paper makes the following contributions:

- different methods of integrating semantics into the pipeline input.
- joint learning of MDE and semantic parsing to exploit correlated features.

Section 2 discusses the existing literature on the problem statement. In Section 3 we explain the deep learning architectures and loss function in our proposed approach, while in Section 4, we compare and evaluate the obtained results. We end with our conclusions in Section 5.

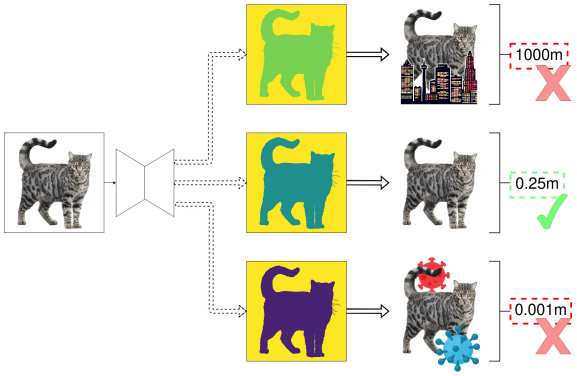


Figure 1: **Ambiguities of monocular depth estimation.** Monocular depth estimation is an inherently ill-posed problem. Infusing semantic signals might aid a neural network in accurately estimating the depth of a given object.

2. Related Work

Several advanced deep learning models [9, 19, 21, 12] have been able to produce effective results for MDE. Conditional random fields (CRF) [24, 40, 43], adversarial learn-

*Authors have contributed equally to this work

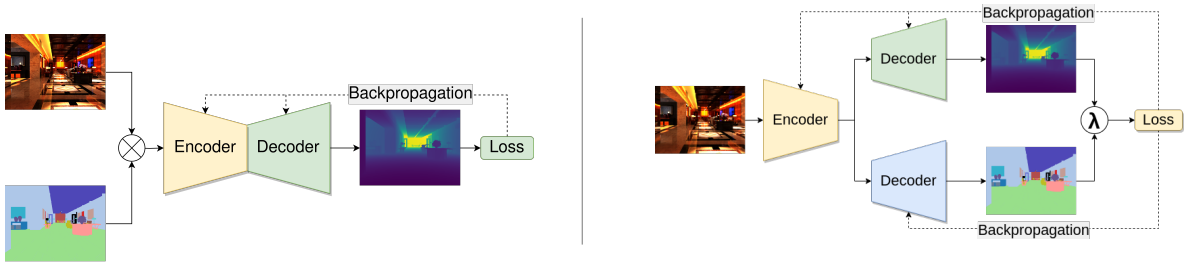


Figure 2: **Illustration of the compared architectures.** The first set of models are multi-input architectures (left), where $\otimes$ represents the combination functions used to infuse semantic signals to the input. The second set of models are multi-task models (right) with a shared backbone encoder and multiple auxiliary tasks predicted by their own decoder. The symbol λ denotes a weighted sum as explained in Section 3:Multi-Task Models.

ing [17, 25, 11], ordinal relationships [44] and application-focused pipelines for obstacle detection [26], real-time deployability [39] and recovering true depth from relative depth annotations [6] have also been explored.

Both depth maps and semantic masks have been used as additional input alongside RGB images [13, 29] to improve the accuracy of the counterpart model. Liu *et al.* [23] first performs semantic segmentation on the RGB images and then uses the semantic masks as input for depth estimation. SDC-Depth [37] decomposes an image into semantic segments and predicts a scale and shift invariant depth map for each segment in a canonical space. Kuga *et al.* [18] proposes an architecture that takes multi-modal input such as RGB images, depth, and semantic labels, and generates multi-modal outputs in a multi-task learning framework.

Various strategies have been explored for the multi-task learning of depth and semantics. These approaches use RGB image input and predict both depth and semantics via multi-scale architectures [8], cross-modal interactions [15], or designing more effective joint-optimization objective functions [16]. Some joint learning models also use CRF [27, 38] to refine their results. SOSD-Net [2] proposes the concept of semantic objectness which exploits the geometric relationship between the two tasks. PAD-Net [41] introduces a guided prediction-and-distillation network to predict a set of intermediate auxiliary tasks, which are then utilized as multi-modal input for the final tasks.

3. Method

This section describes the proposed approaches for infusing semantic signals for monocular depth estimation as demonstrated in Figure 2. We explore three different architecture styles. The first is a baseline approach, which focuses on direct depth regression. The next set of models aims to compare different ways of integrating semantic signals into the input, by applying various transformations to the semantic map before concatenating it to the image. Finally, the third category investigates multi-task learning solutions by predicting auxiliary tasks. All explored models use an encoder-decoder style architecture as presented in UNet [31]. In the following, we explore each category in more detail.

Baseline Model. We begin with a regression model, which predicts depth from only the RGB image as input. The model constitutes the baseline from which we can evaluate the performance of more complex architectures. We try two different loss functions for the regression model. We compare the ℓ_1 loss

$$\ell_1(d_i, \hat{d}_i) = |d_i - \hat{d}_i| \tag{1}$$

with the popular BerHu loss [45] which is typically formulated as

$$\mathcal{L}_{\text{berhu}}(d_i, \hat{d}_i) = \begin{cases} |d_i - \hat{d}_i| & \text{if} |d_i - \hat{d}_i| < c \\ \frac{(d_i - \hat{d}_i)^2 + c^2}{2c} & \text{otherwise} \end{cases}, \tag{2}$$

with $c = \frac{\max_i \{\hat{d}_i, d_i\}}{5}$ [20]. d_i and $\hat{d}_i$ represent the ground truth and predicted depth at pixel i respectively. The BerHu loss essentially models an ℓ_1 loss for predictions close to the ground truth and an ℓ_2 loss for less accurate predictions.

Multi-Input Models. This category focuses on comparing different approaches of combining the RGB image $\mathbf{x}_{\text{RGB}}$ with the semantic map $\mathbf{x}_{\text{semantic}}$ in the input. The semantic maps are feature maps with $W \times H$ dimension, wherein each cell is populated with an integer identifier representing the corresponding class. Formally, our input to the network, $\mathbf{x}$, is defined as

$$\mathbf{x} \doteq [\mathbf{x}_{\text{RGB}}, T(\mathbf{x}_{\text{semantic}})], \tag{3}$$

where $[\cdot, \cdot]$ represents concatenation along the channel axis, and $T(\mathbf{x})$ signifies some transformation on the input.

In the initial model, we directly concatenate the semantic map, that is, $T(\mathbf{x}) = \mathbf{x}$. The subsequent models expand on

this approach, but attempt to match domains with the RGB image by onehot encoding the classes along the channel dimensions instead. Formally, given pixel i with class c, we transform it according to

$$T(x_i) = \text{vec}_c(1), \tag{4}$$

where $\text{vec}_c(1)$ describes the onehot vector with the entry 1 at index c. Within our experimentation, we investigate different variations, encoding only the top 40, 20, or 3 classes based on their prevalence in the overall dataset, while assigning zero-vectors to represent the other classes. The 'semantic convolution' approach first feeds the semantic map to a convolutional neural network before concatenating. Specifically, we have chosen an MBConv-block [36], expanding the output channels to 3. Finally, the last variation extracts the semantic contours of the image by performing erosion ($\mathbf{E}$) and dilation ($\mathbf{D}$) on the semantic map

$$T(x_i) = \mathbf{D}(\mathbf{E}(x_i)). \tag{5}$$

Given the heavy correlation between semantic and depth discontinuities, this reduced approach could provide enough information for an improvement in the depth estimation accuracy.

Multi-Task Models. Within the multi-task architecture category, we experiment with the addition of different auxiliary tasks. This exploration aims to examine the model's capability in handling multiple tasks concurrently while investigating whether shared features could improve depth estimation. While these features might not be primarily aimed at depth estimation, they could indirectly enhance its performance by providing a broader knowledge base for the model to draw from. A defining characteristic of all architectures in this category is a shared backbone encoder to extract common features. However, each task has its own dedicated decoder, ensuring each task's specialized feature interpretation. The variations in this category are based on the number of tasks each version handles. The model referred to as '2-head', handles both depth regression and semantic parsing, while the '3-head' model handles both previous models, as well as predicting semantic contours. Both semantic losses are trained via the Cross-Entropy (CE) loss

$$\mathcal{L}_{\text{CE}}(\mathbf{y}_i, \hat{\mathbf{y}}_i) = -\sum_c y_{i,c} \log \hat{y}_{i,c}, \tag{6}$$

where $\hat{y}_{i,c}$ is the predicted probability for class c at pixel i, with the associated ground truth $y_{i,c}$. The different tasks are combined via the shared loss function which is a weighted sum of the associated losses,

$$\mathcal{L} = \mathcal{L}_{\text{depth}} + \sum_i \lambda^{(i)}_{\text{semantic}} \mathcal{L}^{(i)}_{\text{semantic}} \tag{7}$$

4. Experiments

To effectively compare the performance of our proposed architectures for infusing semantic signals, we conduct experiments on the HyperSim dataset [30], which features a comprehensive set of synthetic indoor scenarios. Below, we provide the specifics of our experimental assessment.

4.1. Experimental Setup

Datasets and Preprocessing. All images have a resolution of 1024×768, and are supplemented with fully labelled depth and semantic maps. Notably, the provided depth maps follow a log-normal distribution with a mean of 5.4 meters across the entire dataset, simulating real-world indoor environments. The initial dataset was subjected to a series of preprocessing steps to ready it for use in the model. Firstly, the raw data was converted to a Low Dynamic Range (LDR) format and normalized. Subsequently, the images were resized and center-cropped, ensuring a consistent resolution of 256×256 pixels. Finally, the depth information was clamped within the range of 0-15m and normalized, to focus on the depth ranges most relevant to indoor environments. We use 19,399 images from 362 scenes for training and 1203 images from 45 scenes for validation.

Evaluation Metrics. To facilitate quantitative comparisons, we utilize several depth metrics commonly applied in literature [42, 10, 20, 4]. Namely, root mean squared error (RMSE): $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{d}_i - d_i)^2}$, relative mean error (RME): $\frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{d}_i - d_i|}{\hat{d}_i}$, log10: $\frac{1}{N}\sum_{i=1}^{N} \| \log_{10}(\hat{d}_i) - \log_{10}(d_i)\|$, and δ_i, describing the percentage of pixels j, such that $\max(\frac{\hat{d}_j}{d_j}, \frac{d_j}{\hat{d}_j}) = \delta < t_i$, where $t_i = 1.25^i$. $\hat{d}_i$ is the prediction while d_i is the ground truth depth for pixel i. Pixels with `NaN` labels are ignored.

Implementation Details. With respect to the model architecture, we incorporate an EfficientNet-b4 [35] backbone after conducting comprehensive experiments against alternative models such as EfficientNet-b2, EfficientNet-b3, and various ResNet versions, due to its superior performance as shown in Table 2, 'Encoders' Category. The weighting coefficient in Eq. 7 is experimentally chosen to be $\lambda^{(i)}_{\text{sem}} = 0.07 \ \forall i$. We further investigate the use of different loss functions, but eventually settle on a combination of ℓ_1-loss for depth and Cross-Entropy (CE) loss for semantic, which exhibits the best (or comparable) results. All model configurations utilize an ImageNet [7] pretrained encoder, whereas the decoders are randomly initialized. The depth is continuously monitored throughout the experiments to validate the model's performance.

Category	Name	RMSE↓	RME↓	log10↓	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$
Baseline	ℓ_1	*1.246*	*0.311*	*0.109*	*0.596*	*0.838*	*0.919*
	BerHu	1.330	0.350	0.116	0.576	0.821	0.909
Multi-Input	concat	1.283	0.315	0.111	0.592	0.834	0.918
	1hot-40	**1.189**	0.285	0.101	0.623	0.863	**0.939**
	1hot-20	1.213	**0.271**	**0.099**	**0.628**	**0.864**	0.935
	1hot-3	1.209	0.284	0.106	0.602	0.850	0.931
	sem. conv.	1.228	0.297	0.106	0.614	0.842	0.923
	contour	1.233	0.311	0.109	0.598	0.838	0.920
Multi-Task	2-head	*1.243*	0.304	*0.108*	*0.607*	*0.840*	*0.920*
	3-head	1.289	*0.301*	0.110	0.601	0.839	0.918

Table 1: **Quantitative results for the depth estimation task on the HyperSim dataset.** Best results per category are italicized in blue, best overall results in bold. We can see that the multi-task models slightly outperform the baseline models, with the 2-head model receiving the better results. The best overall results are achieved by the multi-input architectures, where all models except for 'concat' outperform the other categories. The best performing models are 1hot-40 and 1hot-20.

4.2. Experimental Results

We compare our methods quantitatively as shown in Table 1 and provide qualitative results in Figures 3 and 4. These results confirm that semantic signals can play a substantial role in enhancing the performance of depth models. In fact, the introduction of semantic signals directly into the input appeared to be the most effective strategy, yielding the best performance across all experiments.

In the context of multi-task prediction, our findings indicate that this approach can provide some benefit. However, the degree of improvement was comparatively marginal. This approach's effectiveness may be considerably boosted with more explicit feature sharing between the tasks. Furthermore, the limited improvement from multi-task prediction hints at the necessity for more capable model architectures. Our results show that the UNet architecture simply hits a limit with regards to the semantic prediction task. As a performant semantic model is essential for effective feature sharing in multi-task models, it may be worthwhile to explore more sophisticated architectural design patterns.

Baseline Models. To establish a baseline model, two types of depth loss, ℓ_1 and BerHu, are compared on the baseline architecture. The resulting metrics show that ℓ_1-loss outperforms BerHu by a significant margin. Qualitatively, we confirm that BerHu underperforms at higher ranges of depth possibly because we use a normalized [0,1] depth range, leading to ℓ_2 actually penalizing less than ℓ_1.

Multi-Input Models. Focusing on multi-input representations, we find that the use of naive concatenation showed the least performance out of this set of approaches. This inefficiency may be attributable to domain mismatch issues between RGB images and semantic maps. Although the incorporation of semantic convolution alleviates some of the problems associated with concatenation, the overall model performance might be hindered due to capacity limitations.

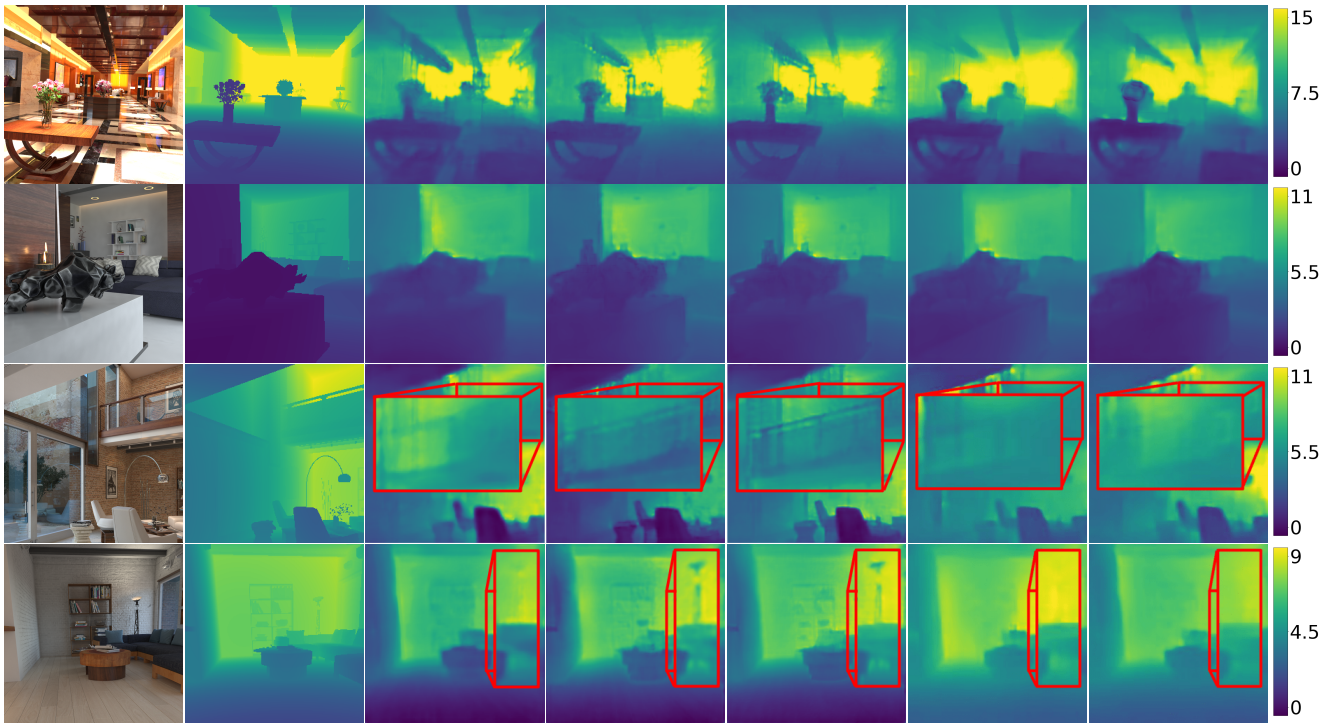
On the other hand, the onehot-encoding strategy rectifies the domain issues observed with concatenation. Particularly, the variant '1hot40' proves to be the most effective, likely due to it including the most information. Nevertheless, we do observe approaching returns from the '1hot20' variant. A possible explanation for this might be that incorporating more classes could result in an overwhelming number of channels when compared to the 3 channels allocated for RGB. The '1hot3' variant meanwhile offers a too limited capacity, resulting in sub-optimal performance. A considerable downside of these models consists in the face that they require semantic maps for inference.

Multi-Task Models. Finally, we discuss the multi-task models individually. Most importantly, we do not observe a significant improvement in the 3-head model over the 2-head version. In fact, the latter seems to outperform the former in most metrics. The reason behind this result might lie in the added complexity introduced by the additional prediction task, further pulling focus away from the depth estimation. Given that the semantic contour task could in theory be directly derived from the conventional semantic prediction, the value added by this additional regularization does not seem to outweigh the added difficulty.

4.3. Ablation Studies

Semantic Losses. To improve the sub-optimal results in semantic segmentation, as seen in Figure 4, we explore modified semantic losses. We experiment with various loss functions, including Dice loss [34]

$$\mathcal{L}_{\text{dice}} = 1 - \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FN} + \text{FP}}, \tag{8}$$



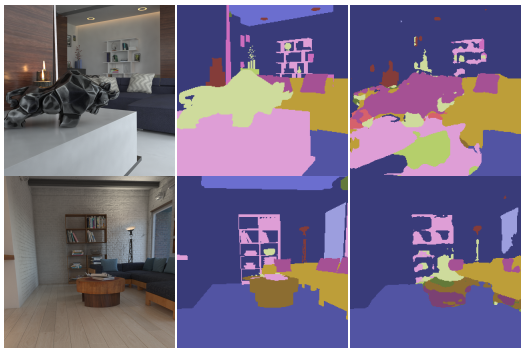
(a) RGB (b) Groundtruth (c) ℓ_1 (d) 1hot-20 (e) 1hot-40 (f) 2-head (g) 3-head

Figure 3: **Qualitative results across different architectures.** Multi-Input models (d) and (e) outperform the baseline version (c) especially with regards to small and thin objects (as shown in highlights) while producing sharper images in the process. Multi-Task models (f) and (g) produce comparable or improved results to (c) despite the added complexity in their architecture. Note especially the sharper object borders in these two models.

and Focal Tversky loss [32]

$$\mathcal{L}_{\text{ftl}} = \left(1 - \frac{\text{TP}}{\text{TP} + \alpha\text{FN} + \beta\text{FP}}\right)^{\gamma}, \tag{9}$$

where TP, FN and FP represent true positive, false negative and false positive predictions respectively. α, β and γ are hyperparameter values which are empirically chosen as 0.7, 0.3 and 1.5 respectively. $\mathcal{L}_{\text{dice}}$ and $\mathcal{L}_{\text{ftl}}$ when combined with CE loss have equally weighted contribution.



RGB Semantic GT Prediction

Figure 4: **Qualitative segmentation results (2-head model).** Note imprecise segmentations with often wrong classifications. Object boundaries are decently accurate.

Additionally, we investigate the correlation between semantic edges and depth discontinuities extracted via the Sobel Filter $\mathbf{S}$ [33] by implementing another variation of the 2-head model which predicts both the depth and the discontinuities, and implements a correlation loss $\mathcal{L}_{\text{sobel}}$ as

$$\mathcal{L}_{\text{sobel}}(c_i, \hat{d}_i) = c_i \log(\mathbf{S}(\hat{d}_i)) + (1 - c_i) \log(1 - \mathbf{S}(\hat{d}_i)) \tag{10}$$

where c_i and $\hat{d}_i$ indicate contour ground truth and depth prediction for the ith pixel respectively.

The results presented in Table 2, 'Losses' Category demonstrate no significant improvements in our model's performance. This lack of progress emphasizes the potential inadequacies inherent in our current architecture.

DeepLabV3. Following the previous results, we explore a more capable architecture: DeepLabV3 [5], retaining the EfficientNet-b4 architecture as the encoder backbone. The results are detailed in Table 2, 'DeepLab' Category. As expected, we observe an overall improvement in the results across all the implemented methods. The most notable progress can be seen in the performance of the 2-head model that employed a combination of Dice and CE loss. These results seem to hint at the possibility that a more ca-

Category	Name	RMSE↓	RME↓	log10↓	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$
Encoders	Resnet50 [14]	1.865	0.517	0.156	0.467	0.728	0.850
	EfficientNet-b2 [35]	1.741	0.459	0.143	0.508	0.752	0.862
	EfficientNet-b3 [35]	1.657	0.436	0.136	0.535	0.772	0.875
	EfficientNet-b4 [35]	1.462	0.391	0.121	0.581	0.817	0.9037
Losses	2-head with Dice	1.283	0.323	0.113	0.589	0.829	0.914
	2-head with Dice+CE	1.269	0.310	0.108	0.603	0.843	0.921
	2-head with FTL+CE	1.296	0.314	0.112	0.599	0.832	0.914
	2-head with Sobel	1.262	0.318	0.110	0.597	0.840	0.918
DeepLab	ℓ_1-baseline	1.208	0.292	0.107	0.604	0.840	0.923
	1hot-40	1.091	0.257	0.094	0.643	0.879	0.946
	2-head with CE	1.277	0.314	0.111	0.593	0.832	0.916
	2-head with Dice+CE	1.154	0.296	0.105	0.619	0.848	0.925

Table 2: **Quantitative ablation results.** The first section compares the baseline architecture with different encoders. The EfficientNet-b4 architecture outperforms the other models. In the second section we compare the 2-head model with different semantic loss functions, where we do not see any improvement over the CE loss. Finally, the last section shows the results of using the DeepLabV3 architecture instead of UNet. Significant enhancements are observed across all aspects, particularly in the case of the 2-head model utilizing the Dice+CE loss, which exhibits a substantial increase in quality.

pable architecture, when combined with a more comprehensive loss function, can indeed deliver significant improvements in the multi-task architecture framework.

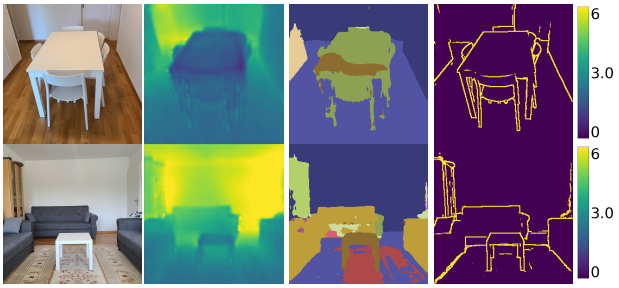


Figure 5: **Qualitative results from the 3-head model on real-life images.** The distances to the tables are 1.6m and 1.8m in the first and second image respectively. The corresponding predicted distances are between 1.50-1.76m, and 2.5m. For the center of the second back wall, the distance is about 4.35m, while our model predicts 5m. Segmentation predictions remain inaccurate, while contour predictions and object boundaries are visually precise.

Real Life Scenes. Finally, we endeavour to examine the simulation-to-real generalization capability of our models. The findings from our 3-head model on two real-world images are illustrated in Figure 5. We observe that the model manages to generalize well, producing accurate spatial layouts, relatively sharp object boundaries, and good gradients on walls and floors. Especially the semantic contour predictions are visually precise, hinting at decent semantic capabilities. However, depth predictions vary in precision. This could be attributed to changing camera intrinsics (such as lenses) or post-processing.

5. Conclusions

In this work, we have proposed two ways of incorporating semantic signals into a monocular depth estimation deep learning pipeline — by introducing semantics in the model input and by simultaneous learning of depth and semantics. Our extensive experiments on the HyperSim dataset demonstrate that infusing semantics does aid in depth estimation by achieving better results across all metrics, particularly a 12.86% and 9.17% improvement in RME and log10 metric respectively. Our ablation studies demonstrate the impact of different loss functions and model architectures on the results, providing potential avenues for future work. It would be especially interesting to experiment with recent state-of-the-art models, that can more accurately capture both semantics and depth information. We also visualize our work on real-world data, which shows our model's potential to perform well on non-synthetic datasets in the future.

Work Distribution.

- Charyyev: data cleaning, baseline setup, multi-input: concat, contour, visualizations
- Lemke: codebase infrastructure & logging setup, data preprocessing, baseline, multi-input: sem-conv, multi-task models
- Ghosh: euler integration, data preprocessing, baseline, multi-input: 1hot methods, multi-task ablations
- All: training, approaches, experiments, ppts & report

References

[1] Deep learning for monocular depth estimation: A review. *Neurocomputing*, 438:14–33, 2021. 1

[2] Sosd-net: Joint semantic object segmentation and depth estimation from monocular images. *Neurocomputing*, 440:251–263, 2021. 2

[3] Towards real-time monocular depth estimation for robotics: A survey. *arXiv: Robotics*, 2021. 1

[4] S. F. Bhat, I. Alhashim, and P. Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021. 3

[5] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 5

[6] W. Chen, Z. Fu, D. Yang, and J. Deng. Single-image depth perception in the wild. *Advances in neural information processing systems*, 29, 2016. 2

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. 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

[8] D. Eigen and R. 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. 2

[9] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 1

[10] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 3

[11] T. Feng and D. Gu. Sganvo: Unsupervised deep visual odometry and depth estimation with stacked generative adversarial networks. *IEEE Robotics and Automation Letters*, 4(4):4431–4437, 2019. 2

[12] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. 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. 1

[13] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik. Learning rich features from rgb-d images for object detection and segmentation. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*, pages 345–360. Springer, 2014. 2

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[15] O. H. Jafari, O. Groth, A. Kirillov, M. Y. Yang, and C. Rother. Analyzing modular cnn architectures for joint depth prediction and semantic segmentation. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4620–4627. IEEE, 2017. 2

[16] J. Jiao, Y. Cao, Y. Song, and R. Lau. Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. In *Proceedings of the European conference on computer vision (ECCV)*, pages 53–69, 2018. 2

[17] H. Jung, Y. Kim, D. Min, C. Oh, and K. Sohn. Depth prediction from a single image with conditional adversarial networks. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1717–1721. IEEE, 2017. 2

[18] R. Kuga, A. Kanezaki, M. Samejima, Y. Sugano, and Y. Matsushita. Multi-task learning using multi-modal encoder-decoder networks with shared skip connections. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 403–411, 2017. 2

[19] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016. 1

[20] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016. 2, 3

[21] J. Li, R. Klein, and A. Yao. A two-streamed network for estimating fine-scaled depth maps from single rgb images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3372–3380, 2017. 1

[22] X. Lin, D. Sánchez-Escobedo, J. R. Casas, and M. Pardàs. Depth estimation and semantic segmentation from a single rgb image using a hybrid convolutional neural network. *Sensors*, 19(8), 2019. 1

[23] B. Liu, S. Gould, and D. Koller. Single image depth estimation from predicted semantic labels. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 1253–1260. IEEE, 2010. 2

[24] F. Liu, C. Shen, and G. Lin. Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5162–5170, 2015. 1

[25] K. G. Lore, K. Reddy, M. Giering, and E. A. Bernal. Generative adversarial networks for depth map estimation from rgb video. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1258–12588. IEEE, 2018. 2

[26] M. Mancini, G. Costante, P. Valigi, and T. A. Ciarfuglia. Fast robust monocular depth estimation for obstacle detection with fully convolutional networks. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4296–4303. IEEE, 2016. 2

[27] A. Mousavian, H. Pirsiavash, and J. Košecká. Joint semantic segmentation and depth estimation with deep convolutional networks. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 611–619. IEEE, 2016. 2

[28] V. Patil. *Improving depth learning with scene priors*. PhD thesis, ETH Zurich, 2022. 1

[29] X. Ren, L. Bo, and D. Fox. Rgb-(d) scene labeling: Features and algorithms. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2759–2766. IEEE, 2012. 2

[30] M. Roberts, J. Ramapuram, A. Ranjan, A. Kumar, M. A. Bautista, N. Paczan, R. Webb, and J. M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *ICCV*, 2021. 3

[31] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2

[32] S. S. M. Salehi, D. Erdogmus, and A. Gholipour. Tversky loss function for image segmentation using 3d fully convolutional deep networks. In *Machine Learning in Medical Imaging: 8th International Workshop, MLMI 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 10, 2017, Proceedings 8*, pages 379–387. Springer, 2017. 5

[33] I. Sobel. An isotropic 3x3 image gradient operator. *Presentation at Stanford A.I. Project 1968*, 02 2014. 5

[34] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 240–248. Springer, 2017. 4

[35] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 3, 6

[36] M. Tan and Q. Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021. 3

[37] L. Wang, J. Zhang, O. Wang, Z. Lin, and H. Lu. Sdc-depth: Semantic divide-and-conquer network for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[38] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. L. Yuille. Towards unified depth and semantic prediction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2800–2809, 2015. 2

[39] D. Wofk, F. Ma, T.-J. Yang, S. Karaman, and V. Sze. Fastdepth: Fast monocular depth estimation on embedded systems. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6101–6108. IEEE, 2019. 2

[40] D. Xu, W. Ouyang, X. Alameda-Pineda, E. Ricci, X. Wang, and N. Sebe. Learning deep structured multi-scale features using attention-gated crfs for contour prediction. *Advances in neural information processing systems*, 30, 2017. 1

[41] D. Xu, W. Ouyang, X. Wang, and N. Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 675–684, 2018. 2

[42] D. Xu, W. Ouyang, X. Wang, and N. Sebe. Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 675–684, 2018. 3

[43] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5354–5362, 2017. 1

[44] D. Zoran, P. Isola, D. Krishnan, and W. T. Freeman. Learning ordinal relationships for mid-level vision. In *Proceedings of the IEEE international conference on computer vision*, pages 388–396, 2015. 2

[45] L. Zwald and S. Lambert-Lacroix. The berhu penalty and the grouped effect. *arXiv preprint arXiv:1207.6868*, 2012. 2