Supplementary Material (Learning to Prompt CLIP for Monocular Depth Estimation: Exploring the Limits of Human Language)

Dylan Auty
Imperial College London
dylan.auty12@imperial.ac.uk

Krystian Mikolajczyk
Imperial College London
k.mikolajczyk@imperial.ac.uk

A. Learned Tokens: Nearest Neighbours in Token Space

Following the observations detailed in section 4.5, we show the nearest-neighbours by Euclidean distance in the 512-dimensional token-space in table A.1. To create this table, the distances between the learned depth tokens along with the pretrained and frozen CLIP word tokens are measured. These tokens are what are used for tokenizing a sentence prior to running through the CLIP text transformer; table 8 shows the similarities between embeddings in the post-CLIP-transformer space.

It can be seen that, contrary to the CLIP embedding space, the learned depth tokens are not near to any depth-or-size related word tokens. In addition, there are many non-English words, nonsense tokens, and punctuation mark tokens present as nearest neighbours. The lower-valued depth tokens do appear to have some relationship to one another, but it is considerably weaker than it is in CLIP embedding space.

From this, we conclude that it is unlikely that the Euclidean distance is a useful metric to measure similarity of tokens in token-space. This is logical: the relationship between token embeddings in CLIP embedding space is strongly indicated when using Euclidean distance, and the CLIP text transformer is nonlinear. It follows that the tokens are not bound to lie in a contiguous region of space, even if their meanings (and therefore their CLIP embeddings) do lie in a similar region of latent space.

B. Comparison to State-Of-The-Art and DepthCLIP for MDE

We emphasise that **our work is not designed to compete with the state-of-the-art methods in MDE**, as the architecture does not use a dense feature decoder. This is to reduce confounding factors, and allow better understanding of the prompting process itself. The lack of a learned decoder naturally limits performance due to the low resolution, but without a learned decoder the learned tokens are forced to be as expressive as possible. This has the added effect of increased explainability.

With this in mind, we show table A.2 that compares some of our experiments to state-of-the-art methods. Table A.3 shows comparisons to DepthCLIP. We also include an upper bound on performance for the $\frac{1}{32}\times$ scale predictions that we generate: the ground-truth depth maps are downsampled to $\frac{1}{32}\times$ their original size to match the prediction generated from our method, then bilinearly upsampled and evaluated in the same way as our other experiments. To handle the invalid depth values in the ground truth, masking of invalid values is applied by logical ANDing the mask at different stages: the full-resolution ground-truth depth, the downsampled ground-truth depth (converted to floating point, upsampled bilinearly, then thresholded at 1.0 to convert back to Boolean), and the ground truth that has been both downsampled and re-upsampled.

We would like to emphasise again that there are only a few thousand learnable parameters in our method, and that the aim of our work is not to improve on SOTA MDE but to understand the way in which language encodes the information contained within CLIP, particularly as it relates to depth.

C. Qualitative Examples

Some qualitative examples are provided in figures A.1 and A.2. While the resolution is necessarily low due to our method deliberately excluding a decoder for the sake of interpretability, we note that recognisable depth is still obtainable, despite the limited number of paramters in use.

References

[1] Ashutosh Agarwal and Chetan Arora. Attention Attention Everywhere: Monocular Depth Prediction with Skip Attention, Oct. 2022. arXiv:2210.09071 [cs].

[2] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. AdaBins: Depth Estimation using Adaptive Bins. *arXiv:2011.14141 [cs]*, Nov. 2020. arXiv: 2011.14141.

{depth_0}		{depth_1}		{depth_2}		{depth_3}		{depth_4}		{depth_5}		{depth_6}	
Token	Dist.	Token	Dist.	Token	Dist.	Token	Dist.	Token	Dist.	Token	Dist.	Token	Dist.
{depth_1}	1.256	*kids</w>*	1.245	*{depth_3}*	1.278	*presenter</w>*	1.240	*{depth_3}*	1.258	*assembly</w>*	1.276	*âĻ©âĻ</w>*	1.268
lestweforget</w>	1.272	*doesn</w>*	1.254	*spx</w>*	1.287	*{depth_4}*	1.258	*apparently</w>*	1.261	*crouch</w>*	1.282	*ãł*	1.270
concerned</w>	1.278	*{depth_0}*	1.256	*gm*	1.290	*staffs</w>*	1.270	*spears</w>*	1.262	*buch*	1.285	*ç¥*	1.280
—</w>	1.286	*watched</w>*	1.264	*ŁÍ´ŁÍ´</w>*	1.290	*....</w>*	1.271	*credible</w>*	1.276	*assemb*	1.289	*(;</w>*	1.283
approved</w>	1.286	*and</w>*	1.265	*ĭ*	1.291	*<—startoftext—>*	1.272	*wid</w>*	1.287	*surg*	1.292	*norms</w>*	1.284
oxford</w>	1.287	*breaks</w>*	1.266	*hobbit</w>*	1.292	*staff</w>*	1.274	*accepted</w>*	1.289	*spion*	1.292	*Łijĩ</w>*	1.285
british	1.288	*play</w>*	1.267	*ŁÍ¬*	1.292	*]</w>*	1.275	*Ê*	1.293	*scicom*	1.293	*Ł¤Ń</w>*	1.286
debate</w>	1.288	*but</w>*	1.269	*Łi*	1.295	*.....</w>*	1.277	*favour</w>*	1.294	*cence</w>*	1.293	*â¬Ĩ*	1.286
qs</w>	1.291	*does</w>*	1.270	*refurbi*	1.296	*visitors</w>*	1.278	*fad</w>*	1.295	*aug*	1.294	*Ã¤*	1.286
loud</w>	1.291	*doesnt</w>*	1.274	*talks</w>*	1.296	*{depth_2}*	1.278	*speake*	1.295	*willy*	1.295	*âÇĶâÇĶ</w>*	1.288
ŁĻı</w>	1.294	*dak</w>*	1.277	*âĨĴ</w>*	1.297	*...</w>*	1.283	*loo</w>*	1.295	*sexu*	1.296	*Ł§IJ</w>*	1.289
loose</w>	1.294	*welcomed</w>*	1.277	*ŁÍ*	1.298	*manager</w>*	1.283	*technological</w>*	1.296	*inaugu*	1.296	*Ł¤</w>*	1.292
..</w>	1.295	*starts</w>*	1.278	*hug*	1.298	*,</w>*	1.284	*friction</w>*	1.296	*incense</w>*	1.297	*quins</w>*	1.292
confuse</w>	1.296	*break</w>*	1.278	*asleep</w>*	1.298	*nesses</w>*	1.285	*positive</w>*	1.296	*muse*	1.297	*geta*	1.293

Table A.1. **Nearest-neighbours in token space** for each of the learned depth tokens. Does not include the human-language ordinal scales from table 1 because token-space embeddings can only capture a single token. Learned tokens from 7 evenly-distributed bins on NYUv2 using 'baseline' template from table 2. 'Distance' is Euclidean distance. Token 0 corresponds to a bin centre of approx. 0.714m, and token 6 to approx. 9.29m. It can be seen that most of the tokens do not correspond to any recognisably-useful English tokens, though bins 0, 1, 2, 3, and 4 do relate to one another. This is in contrast to table 8 of the main paper, in which the learned token's nearest neighbours in the CLIP embedding space are seen to be closely related to one another and to size and depth related words, and to be positioned approximately along a continuum of some kind.

Model	↓ Abs. Rel	↓ RMS	↓ Log10	δ^1	δ^2	δ^3
Eigen et al. [4]	0.158	0.641	-	0.769	0.950	0.988
Laina et al. [6]	0.127	0.573	0.055	0.811	0.953	0.988
DORN [5]	0.115	0.509	0.051	0.828	0.965	0.992
BTS [7]	0.110	0.392	0.047	0.885	0.978	0.994
AdaBins [2]	0.103	0.364	0.044	0.903	0.984	0.997
DepthFormer [8]	0.096	0.339	0.041	0.921	0.989	0.998
BinsFormer (Sw-L) [9]	0.094	0.330	0.040	0.925	0.989	0.997
PixelFormer [1]	0.090	0.322	0.039	0.929	0.991	0.998
LocalBins [3]	0.099	0.357	0.042	0.907	0.987	0.998
AiT (SwinV2-L) [10]	**0.076**	**0.279**	**0.033**	**0.953**	**0.993**	**0.999**
Low-res. prediction Upper Bound†	*0.023*	*0.146*	*0.010*	*0.986*	*0.998*	*1.000*
Ours (7 bins, even dist., learned depth tokens)	0.319	0.970	0.128	0.465	0.776	0.922
Ours (7 bins, even dist., learned depth tokens, ls4o4d learned context tokens)	0.317	0.955	0.126	0.474	0.782	0.925
Ours (256 bins, log dist., learned depth tokens)	0.298	0.933	0.122	0.485	0.796	0.934
Ours (256 bins, log dist., learned depth tokens, ls4o4d learned context tokens)	0.298	0.928	0.121	0.487	0.798	0.935

Table A.2. Comparison of SOTA methods on NYUv2 to our results. **Note that our method does not have a dense decoder and is therefore neither intended nor expected to be competitive with SOTA methods**. Our method predicts depth maps at $\frac{1}{32}\times$ resolution, directly from the output of the feature encoder. †: Upper bound given by downsampling ground-truth depth maps then bilinearly upsampling. Aggressive masking is applied to ensure that invalid pixels being interpolated into valid ranges by accident does not affect the final "prediction".

Method	Abs	RMS	log10	δ_1	δ_2	δ_3
Ours (7-bin even dist. 'Baseline')	0.319	0.970	0.128	0.465	0.776	0.922
Ours (256-bin log dist. 4o4d)	**0.298**	**0.930**	**0.121**	**0.487**	**0.799**	**0.935**
DepthCLIP	0.388	1.167	0.156	0.394	0.683	0.851

Table A.3. **Comparison to DepthCLIP** on NYUv2. We show improved performance across all metrics with only 7 learned tokens (3584 params). Comparison to SOTA is provided in table A.2.

[3] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. LocalBins: Improving Depth Estimation by Learning Local Distributions, Mar. 2022. arXiv:2203.15132 [cs].

[4] David Eigen, Christian Puhrsch, and Rob Fergus. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. In *NIPS*, 2014.

[5] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep Ordinal Regression Network for Monocular Depth Estimation. In *CVPR*, 2018.

[6] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper Depth Prediction with Fully Convolutional Residual Networks. In *3DV*, 2016. arXiv: 1606.00373v2.

[7] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From Big to Small: Multi-Scale Local Planar Guidance for Monocular Depth Estimation. *arXiv:1907.10326 [cs]*, Aug. 2019. arXiv: 1907.10326.

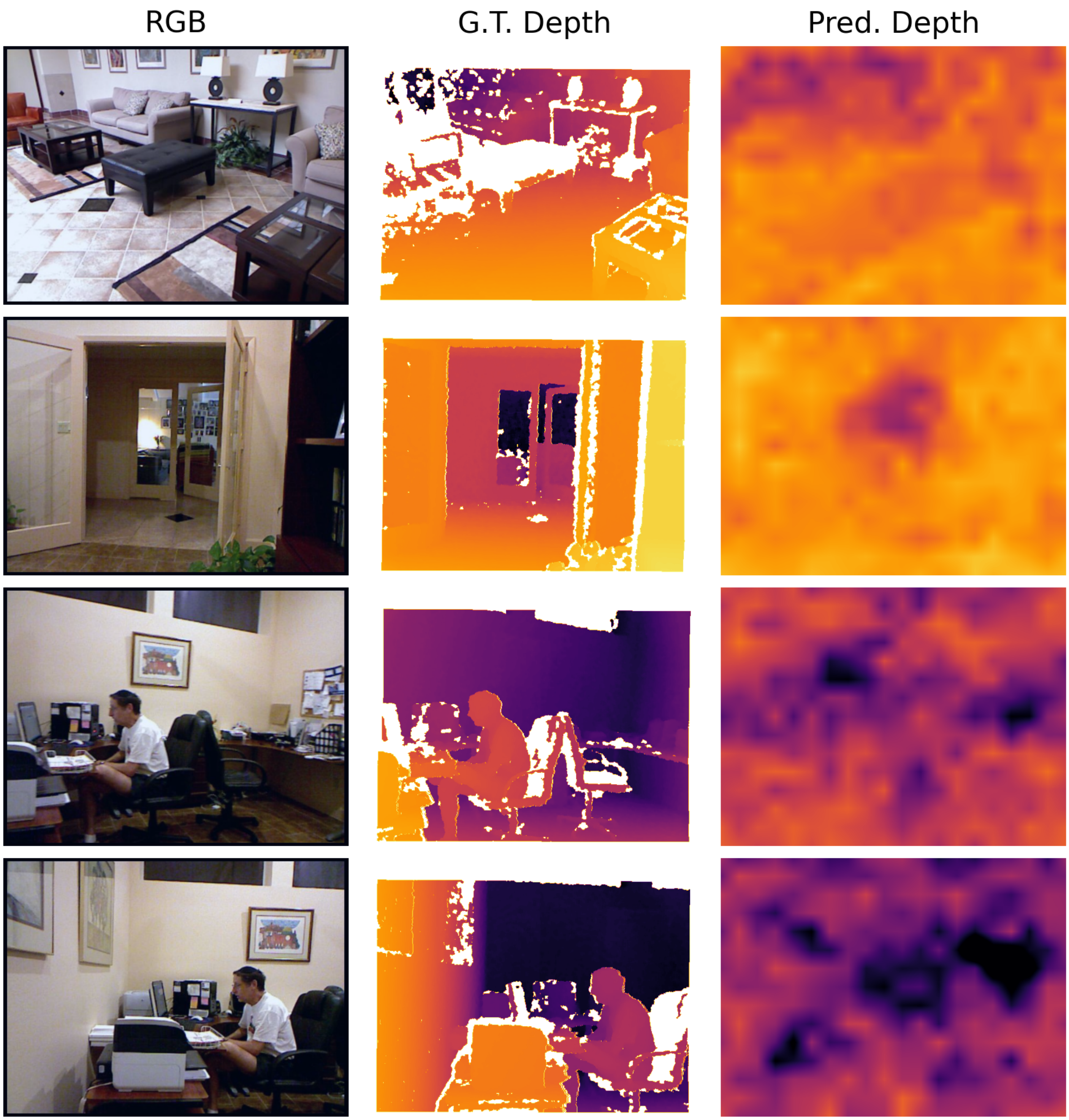

Figure A.1. **Qualitative samples from NYUv2**, using 256 learnable depth tokens with 256 log-distributed depth bins. Also used were 8 total learnable prompt context tokens (ls4o4d).

[8] Zhenyu Li, Zehui Chen, Xianming Liu, and Junjun Jiang. DepthFormer: Exploiting Long-Range Correlation and Local Information for Accurate Monocular Depth Estimation. *arXiv:2203.14211 [cs]*, Mar. 2022. arXiv: 2203.14211.

[9] Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. BinsFormer: Revisiting Adaptive Bins for Monocular Depth Estimation. page 21, 2022.

[10] Jia Ning, Chen Li, Zheng Zhang, Zigang Geng, Qi Dai, Kun He, and Han Hu. All in Tokens: Unifying Output Space of Visual Tasks via Soft Token, Jan. 2023. arXiv:2301.02229 [cs].

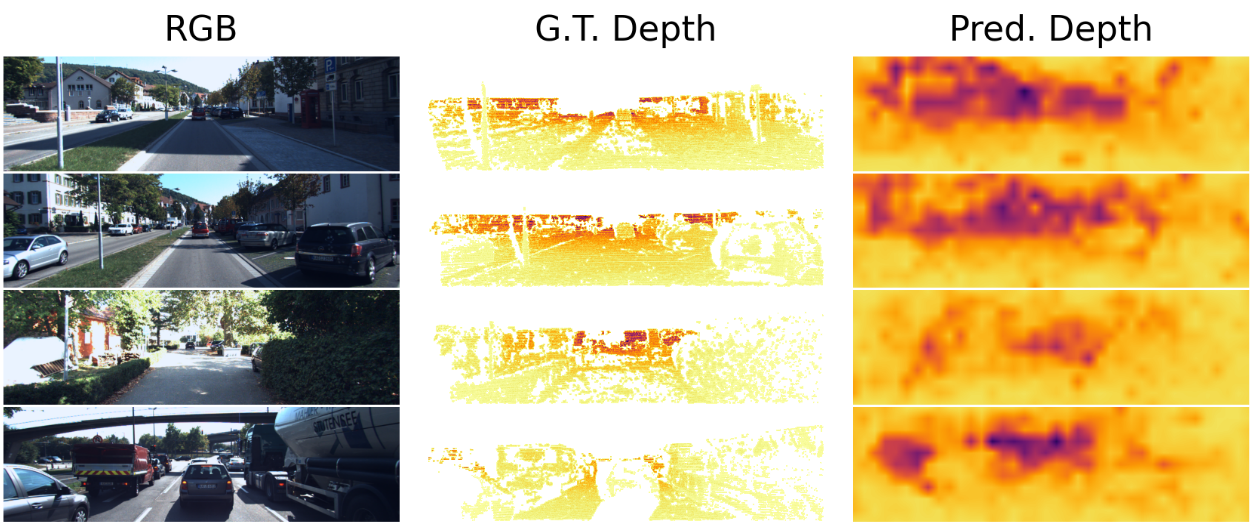

Figure A.2. **Qualitative samples from KITTI**, using 7 learnable depth tokens with 7 evenly-distributed depth bins. Also used were 2 total learnable prompt context tokens (ls1o1d).