ABSTRACT

Besides image classification, Contrastive Language-Image Pre-training (CLIP) has accomplished extraordinary success for a wide range of vision tasks, including object-level and 3D space understanding. However, it’s still challenging to transfer semantic knowledge learned from CLIP into more intricate tasks of quantified targets, such as depth estimation with geometric information. In this paper, we propose to apply CLIP for zero-shot monocular depth estimation, named DepthCLIP. We found that the patches of input image could respond to a certain semantic distance token and then be projected to a quantified depth bin for coarse estimation. Without any training, our DepthCLIP surpasses existing unsupervised methods and even approaches the early fully-supervised networks. To our best knowledge, we are the first to conduct zero-shot adaptation from the semantic language knowledge to quantified downstream tasks and perform zero-shot monocular depth estimation. We hope our work could cast a light on the future research. The code is available at https://github.com/Adonis-galaxy/DepthCLIP.

CCS CONCEPTS
• Computing methodologies → Computer vision; Natural language processing.

KEYWORDS
Contrastive language-image pre-training; monocular depth estimation; zero-shot transfer learning

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM Reference Format:
Renrui Zhang, Ziyao Zeng, Ziyu Guo, and Yafeng Li. 2022. Can Language Understand Depth?. In Proceedings of the 30th ACM International Conference on Multimedia (MM ’22), October 10–14, 2022, Lisboa, Portugal. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3503161.3549201

1 INTRODUCTION

Multi-modality learning has long been a fundamental problem, for which a lot of proposed models utilized language knowledge to assist vision tasks and showed inspiring outcomes. Therein, Contrastive Language-Image Pre-Training (CLIP) [19] exerted powerful transfer ability and achieved promising performance on zero/few-shot image classification [6, 27, 30, 32], object detection [20, 33], semantic segmentation [20, 31] and others [28]. Further, for those vision tasks, CLIP is only required to recognize visual signals from a high level, and how to apply its pre-trained semantic language knowledge to quantified vision tasks, e.g., depth estimation, has not been further explored.

Monocular depth estimation is a vital task in the industrial field, which serves as an essential component in various tasks like monocular 3D object detection [15, 24, 29] or point cloud reconstruction from images [25]. It normally relies on dense depth labels to train a network that extracts semantic relations within an image and regresses pixel-wise depth value. However, training networks from scratch supervised by dense labels severely hinders the efficient deployment for application. Also, it is quite costly to collect and annotate the large-scale datasets, such as NYUV2 [22]. Some of the unsupervised methods need extra data like single-view video [16, 26] to capture time-consistent constraint, or demand 3D priors [11] for better spatial modeling. We then ask the question: can we avoid the cost for both training models and collecting data by using semantic language knowledge learned by CLIP [19]?

For the first time, we explore the relationship between language and depth and apply CLIP [19] to monocular depth estimation. CLIP [19] trains an visual encoder and a textual encoder jointly by contrastive loss in the embedding space between both modalities, which narrows the cosine distance of paired image and text features while pushing away the others. The current transfer mode of CLIP only utilizes its pre-trained classification ability, such as classifying...
each localized object for detection and categorizing each pixel for segmentation. In contrast, during contrastive pre-training, CLIP is not likely to learn knowledge for regressing quantitative values. Thus, it’s impossible to directly output depth prediction from the CLIP model for depth estimation. Considering this, we convert the depth value regression to a distance classification task that the CLIP is only required to understand ‘the object is close/far’, but not ‘the object is 5 meters away’.

In our experiments, we show that the pre-trained CLIP model is able to make responses to the pre-defined distance concepts (like close or remote) for different patches of an image. Then, we map each distance concept to a certain quantitative depth bin and linearly combine the multi-bin depth values weighted by language-patch similarities to produce the depth estimation for a patch. For simplicity, pixels within the same patch would be given the same predicted depth value. By this, our DepthCLIP could conduct zero-shot depth estimation using pre-trained CLIP without any fine-tuning or extra training data, which exceeds existing unsupervised transferred methods and only have a light gap with early fully-supervised models. For analysis, we visualize the distance response of image patches to show that DepthCLIP could coarsely capture the semantic distance concepts. Besides, we test class-dependent bins to verify that by setting such class-class-dependent bins, DepthCLIP could achieve further improvement for wider practical applications. We summarize our contributions as below:

- We propose DepthCLIP, which is the first to perform zero-shot adaptation from semantic language knowledge to monocular depth estimation task.
- We build the bridge between CLIP’s semantic knowledge and quantified depth value prediction, which casts a light on future research.
- To illustrate our effectiveness, we experiment DepthCLIP on NYU Depth v2 [22], which exceeds existing unsupervised methods and draws near to some fully-supervised models.

2 RELATED WORK

**Vision-Language Models.** Vision-Language learning has been widely studied and achieved remarkable outcomes to benefit visual representation learning, especially for zero-shot domain transfer in a wide range of downstream tasks. Contemporarily, driven by extensive image-text pairs emerging on the Internet, CLIP [19] and ALIGN [10] have raised a revolution in vision language learning, in which the former uses 400 million noisy image-text pairs and the latter uses 1.8 billion noisy image-text pairs.

**CLIP** [19] was originally designed for zero-shot image classification. During training, in each batch, CLIP encodes images and corresponding texts into feature space, then maximizes the similarity between corresponding pairs, while minimizing the rest. To conduct zero-shot classification during testing, prompts like “A photo of [class]” would be formed, and the class that has max similarity between prompt features and image features will serve as the prediction. After its remarkable success in zero-shot image classification, it has been applied to few-shot image classification by [6, 27, 30, 32], to image semantation by [20, 31], and to object detection by [20, 35]. However, solving those tasks requires only the recognition ability. Image classification requires learning generic representation for a single image. Segmentation requires distinguishing between various pixels belonging to different classes (semantic segmentation), objects (instance segmentation), or both (panoptic segmentation). As for object detection, CLIP is only responsible for object-level classification rather than localization. Therefore, no work has been done for CLIP to solve quantified tasks predicting continuous values.

**Monocular Depth Estimation.** Monocular depth estimation aims to predict pixel-wise depth value from a single input image. It is a challenging task due to the lack of stereo information that is the key to solve geometry constrain for depth estimation. As a result, the model could only utilize semantic relationships extracted from the input image and semantic prior leaned from training data to solve the problem.

**Fully-supervised Methods** could learn from ground-truth depth maps during training to master semantic prior and extract semantic relationship. DORN [5] proposes a deep convolutional neural network with a spacing-increasing depth discretization strategy and recasts depth learning as an ordinal regression problem. RPSF [18] introduces a differentiable physical model of the aperture mask and simulates the camera imaging pipeline precisely. Thanks to the recent booming of transformer in computer vision community [1, 13, 14, 17], some methods also adopt such architectures for monocular depth estimation. For example, ASTransformer [2] designs an Attention-based Up-sample Block to recompense the detailed texture features. DepthFormer [13] proposes a hierarchical aggregation and heterogeneous interaction modules to build up affinities between features and models. To sum up, all the fully-supervised methods have achieved astonishing performance on monocular depth estimation because of their dedicated designed structures and capacity to model semantics prior during training. However, gathering fully-annotated data would be costly and labor-consuming, which constrains its scalability.

**Unsupervised Methods** construct pre-text tasks delicately to teach models to discover the semantic affinity within a monocular image. Considering that videos could provide temporal constraints for geometry, [16, 26] learn depth knowledge from ego-motion of unlabelled monocular videos. Besides, [8] utilizes widely available binocular stereo images for self-training, which exploits epipolar geometry constraints to generate disparity images for depth prediction. Moreover, without using additional data from other modalities (video or stereo images), [11] introduces plane and line priors to enhance the unsupervised monocular depth estimation. Overall, all previous methods need additional vision-based data or geometry prior to help with depth estimation. Our work is the first to use language semantic information leaned by CLIP [19] to conduct monocular depth estimation, without any explicit geometry priors.

3 METHOD

As shown in Figure 2, our model utilizes the pre-trained knowledge from CLIP to project the semantic response of each patch into a certain depth bin, and linearly combine those depth values to obtain the final prediction. We introduce and discuss three parts
Can Language Understand Depth? MM ’22, October 10–14, 2022, Lisboa, Portugal

This object is [CLASS]

Object

Texture

Features

Text Features

Visual Encoder

Input 2D RGB Image

Image Features

Textual Encoder

Figure 2: Pipeline of DepthCLIP.

respectively below: Image Encoding, Text Encoding, and Depth Projection and Combination.

3.1 Image Encoding

Given a monocular RGB image \( i \), we feed it into the visual encoder without the final pooling layer. Then we acquire its last-layer \( C \)-dimensional feature map \( F_{\text{img}} \), formulated as,

\[
F_{\text{img}} = \text{VisualEncoder}(i) \in \mathbb{R}^{HW \times C}
\]  

(1)

Each spot of \( F_{\text{img}} \) would be given a depth forecast based on its reaction toward semantic language tokens. Since the pre-trained visual encoder of CLIP [19] is received by molding a classification reaction toward semantic language tokens, those tokens \( T \) would grasp regional semantic information, and pooling would assemble the local knowledge to develop a global interpretation of the given image. As a result, each location of the feature map would preserve its surrounding semantic details, and be responded to by text token to obtain its depth approximation.

3.2 Text Encoding

Due to the contrastive pre-text task of CLIP, the textual encoder of CLIP [19] would project similar semantic tokens to the neighborhood of image features. In DepthCLIP, we utilize prompts like ‘This object is distance class’ and apply the semantic tokens close, far, remote that substitute distance class to form \( K \) semantic tokens. Those tokens \( T \) will pass through the textual encoder of CLIP [19] into latent space by

\[
F_{\text{text}} = \text{TextualEncoder}(T) \in \mathbb{R}^{K \times C},
\]  

(2)

where the dimension \( C \) equals \( F_{\text{img}} \). Then we calculate the cosine similarity between semantic tokens’ features \( F_{\text{text}} \) and image features \( F_{\text{img}} \) to acquire similarity scores by

\[
S = \frac{F_{\text{text}} \cdot F_{\text{img}}}{\|F_{\text{text}}\|\|F_{\text{img}}\|} \in \mathbb{R}^{K \times HW}
\]  

(3)

3.3 Depth Projection and Combination

After obtaining similarity scores \( S (\text{dim} = H \times W \times K) \), they ought to be cast to quantified depth to receive depth prediction. That is to say, “close” shall be projected to “1m”. We apply a temperature softmax function among such similarity scores \( S \) to obtained linear combination weight by:

\[
W = \frac{e^{S_{i,j}}}{\sum_{j=1}^{K} e^{S_{i,j}}}, \quad \text{for } i = 1, \ldots, K
\]  

(4)

Then, we employ such weights to linearly combined depth bins \( d \) (\( \text{dim} = K \times 1 \), like \( [d_1=1.00m, d_2=2.00m, d_3=3.00m, \ldots] \)) to obtain the final depth prediction \( D_{\text{pred}} \) (\( \text{dim} = H \times W \times 1 \)) for each location of the image feature map, which is aligned with patches in the image:

\[
D_{\text{pred}} = \sum_{i=1}^{K} W_{:,i} \ast d_i.
\]  

(5)

Where \( D_{\text{pred}}_{i,j} \) is the depth of the patch in \( i \) th row and \( j \) th column. All pixels belonging to such certain patch will be presented with the same depth prediction \( D_{\text{pred}}_{i,j} \), where \( i \in [1, H], j \in [1, W] \).

4 EXPERIMENTS

To demonstrate the effectiveness of our zero-shot training-free depth estimator, we conduct ample experiments examining different facets of our approach. After introducing the implementation details, we evaluate our methods on a mainstream challenging benchmark—NYU Depth v2 [22]. Ablation studies and visualization are provided to offer a more thorough investigation of our method.

4.1 Datasets

We evaluate our method on NYU Depth v2 [22]. The dataset consists of 120K pairs of RGB and depth images, in which all image pairs are taken by the Microsoft Kinect sensor under 464 indoor
scenarios with a resolution of 480×640. Depth range for each pixel is 0-10m. The same training/testing split configuration are applied following [4, 12]. The training set covers 36,253 images from 249 scenes, while the testing set contains 654 images from the remaining 215 scenes. To remove frames, all samples are cropped to the resolution of 416×512 with the identical configuration in [23].

4.2 Implementation Details
We implement our model with the PyTorch framework. Our image and textual encoders employ the pre-trained ResNet-50 [9] of CLIP [19]. For semantic prompts, we tested various kinds of hand-craft prompts, and pick “This object is [distance class]”. For semantic distance classes, we investigated diverse combinations and select [‘giant’, ‘extremely close’, ‘close’, ‘not in distance’, ‘a little remote’, ‘far’, ‘unseen’], 7 semantic bins in total. Each of which aligns with a depth bin of [1.00, 1.50, 2.00, 2.25, 2.50, 2.75, 3.00]. We set this setting for our main experiments since the range of indoor depth could be properly captured under such proper numbers of semantic and depth bins. The temperature of the final softmax function is set to 0.1.

4.3 Evaluation Metrics
The evaluation metrics are following previous works [2]. We compare our method quantitatively with metrics listed below: mean absolute relative error (rel), root mean square error (rmse), absolute error in log space (log10), and threshold accuracy (δ). The formula of each metric has been itemized below:

\[
\text{rel} = \frac{1}{n} \sum_{p} \frac{|y_p - \hat{y}_p|}{\hat{y}_p}, \quad \text{rmse} = \sqrt{\frac{1}{n} \sum_{p} (y_p - \hat{y}_p)^2}
\]

\[
\text{log10} = \frac{1}{n} \sum_{p} |\log_{10}(y_p) - \log_{10}(\hat{y}_p)|
\]

\[
\delta = \% \text{ of } \hat{y}_p \text{ s.t. } \max \left( \frac{y_p}{\hat{y}_p}, \frac{\hat{y}_p}{y_p} \right) = \delta < \text{thr for thr = 1.25, 1.25^2, 1.25^3} \quad (6)
\]

4.4 Quantified Results
Table 1 shows our results compared with other monocular depth estimation methods. The table is divided by different supervisions and pre-training datasets. The lower bound is obtained by randomly making predication for each pixel within the depth range 0-10m. It could be noticed that despite the gap between our method with fully-supervised methods, DepthCLIP exceeds the lower bound by a large margin, even surpassing other zero-shot transferring methods that are pre-trained on the dataset especially prepared for monocular depth estimation (unsupervised KITTI monocular video [7]). Besides, the performance of our DepthCLIP has been highly close to some fully-supervised methods like Make3D [21], even exceeding it in terms of rmse(1.186 of ours compared with 1.214 of Make3D [21]).

4.5 Class-dependent Depth Bin
Our DepthCLIP requires attaching a quantified depth bin to each semantic language token, then linearly combined to obtain final predication. Shown as Table 3, images from different classes possess different depth distribution. CLIP [19] can capture semantic distance relationships within one image, but patches holding different distances of various scenes could be mapped to the same semantic concept. In other words, the same semantic token should be mapped to different class-dependent quantified depth bins in different scenes, to achieve better performance. Due to the limit of time, we remain the same bin partition for all classes, and conduct ablation to examine the effectiveness of using class-dependent bin partitioning.

Shown as Table 2, we select the bathroom class, the classroom class, and all classes of NYU Depth v2 [22] to test class-dependent depth bin partitioning. Each depth bin partition is also tested with all classes of NYU Depth v2 [22] to serve as a comparison. We could notice that by depth bin partition based on scene class of test image, compared with evaluating with original shared by all classes for NYU Depth v2 [22] used in Table 1, we could achieve a remarkable performance gain. On the other hand, different classes have different best bin partitions. That is to say, we could learn a class-dependent bin partition, and conduct zero-shot classification using CLIP [19] to match images with corresponding depth bin partitions to achieve better results, which would be a worthy future research direction.

4.6 Prompt Design
We evaluate our method under different prompt designs on all classes of NYU Depth v2 [22]. Shown as Table 3, our method achieves the best performance under the original prompt design, which is used for the results we report for Table 1. It is worth noticing that, the performance gap among different prompt settings does not vary too much, especially not as sensitive as the depth bin partition shown in Table 2. This outcome is caused by similar semantic meanings inhabiting within different semantic tokens, like “giant” and “extremely close” express a similar meaning, and a patch containing a close object would be projected to both tokens, with only response intensity varying a little for different prompt designs. In this way, the major part of the semantic response score
Table 1: Performance of Monocular Depth Estimation on NYU Depth v2 [22]. The table is divided by different supervisions and pre-training datasets. Lower bound is obtained by randomly making predication for each pixel within depth range 0-10m.

| Method             | Pre-training | Supervision | \(\delta < 1.25\) | \(\delta < 1.25^2\) | \(\delta < 1.25^3\) | \(\text{rel}\) | \(\log_{10}\) | \(\text{rmse}\) |
|--------------------|--------------|-------------|-------------------|-------------------|-------------------|----------------|-----------|------------|
| Make3D [21]        | -            | depth       | 0.447             | 0.745             | 0.897             | 0.349         | -         | 1.214      |
| DORN [5]           | -            | depth       | 0.828             | 0.965             | 0.992             | 0.115         | 0.051     | 0.509      |
| ASFormer [2]       | -            | depth       | 0.902             | 0.985             | 0.997             | 0.103         | 0.044     | 0.374      |
| DepthFormer [13]   | -            | depth       | 0.921             | 0.989             | \textbf{0.998}    | 0.096         | 0.041     | 0.339      |
| RPSF [18]          | -            | depth       | \textbf{0.952}    | \textbf{0.989}    | 0.997             | \textbf{0.072} | 0.029     | \textbf{0.267} |
| Lower Bound        | -            | -           | 0.140             | 0.297             | 0.471             | 1.327         | 0.323     | 2.934      |
| vid2depth [16]     | KITTI video  | 0-shot      | 0.268             | 0.507             | 0.695             | 0.572         | -         | 1.637      |
| Zhang et al. [26]  | KITTI video  | 0-shot      | 0.350             | 0.617             | 0.799             | 0.513         | 0.529     | 1.457      |
| DepthCLIP          | CLIP [19]    | 0-shot      | \textbf{0.394}    | \textbf{0.683}    | \textbf{0.851}    | \textbf{0.388} | \textbf{0.156} | \textbf{1.167} |

Table 2: Ablations of class-dependent depth bin. Perform in bathroom, classroom, and all classes of NYU Depth v2 [22]. The uppermost table shows partition details of each depth bin partition, the lower three tables show the performance of each partition in the bathroom, classroom, and all classes. The original bin is the one used for Table 1 to evaluate the entire dataset among all classes. Performance gain between the original bin and best class-dependent bin is listed in the last row, while the best class-dependent bin is bold.

Table 4.1: Performance of Monocular Depth Estimation on NYU Depth v2 [22]. The table is divided by different supervisions and pre-training datasets. Lower bound is obtained by randomly making predication for each pixel within depth range 0-10m.

| Bin partition | Depth bin partition details (in meters) |
|---------------|----------------------------------------|
| Original bin  | [1.00, 1.50, 2.00, 2.25, 2.50, 2.75, 3.00] |
| Class-dependent 1 | [1.00, 2.00, 2.25, 2.50, 2.75, 3.00, 4.00] |
| Class-dependent 2 | [1.00, 1.50, 2.00, 2.50, 3.00, 3.50, 4.00] |
| Class-dependent 3 | [1.00, 1.25, 1.50, 1.75, 2.00, 2.25, 2.50] |
| Class-dependent 4 | [2.00, 2.50, 3.00, 3.25, 3.50, 3.75, 4.00] |

Figure 4: Semantic bin responses in different patches of an image for DepthCLIP.

Table 4.2: Semantic Bin Responses

4.7 Semantic Bin Responses

Table 4: Performance of Monocular Depth Estimation on NYU Depth v2 [22]. The table is divided by different supervisions and pre-training datasets. Lower bound is obtained by randomly making predication for each pixel within depth range 0-10m.

Can Language Understand Depth? MM ’22, October 10–14, 2022, Lisboa, Portugal
4.8 Depth Predication Visualization

As shown in Figure 5, we visualize depth predictions of our DepthCLIP, compared with ground truth and input RGB images. It could be witnessed that our prediction is a little bit blurred, and focuses more on detailed objects, like the tap on the left column, and the bathtub on the right column. This is discussed in the limitation section below, that the visual encoder of CLIP [19] is pre-trained using a classification pre-text task, in which the model pays more attention to local details that help classification. In this way, backgrounds and local areas would be neglected.

5 LIMITATIONS & FUTURE RESEARCH DIRECTION

One limitation is that our work requires hand-craft depth bin tuning. For various depth distributions of different scenes, DepthCLIP would perform poorly if different scenes share the same depth bin setting. One possible solution is to conduct zero-shot image classification first, which could naturally be done by CLIP [19], then determine class-dependent depth bins for predicated scene class. Learning how to generate class-dependent depth bins for each predicated scene class would be a worthwhile future direction.

Another limitation is its visual encoder of CLIP [19] is trained under a classification-alike pre-training task. As a result, its visual encoder would pay more attention to local details which help identify its scene class. Local regions without noteworthy features for classification would produce non-significant semantic depth reactions, e.g., large areas of wall, ceiling, or floor. One possible solution is to train a vision language model with a regional pre-text task like image segmentation or object detection, to drive more regions’ sensitivity.

6 CONCLUSION

In this paper, we propose DepthCLIP to conduct zero-shot monocular depth estimation with Contrastive Language-Image Pre-Training, which discards the need of training and data with depth labels. It directly applies CLIP [19] that only pre-trained with large-scale image-text pairs to monocular depth estimation task, and achieved satisfying performance. We hope our work could cast a light on the research of bridging semantic vision language knowledge to quantitative vision tasks, and open future research on zero-shot monocular depth estimation.

Acknowledgement. This work is supported by the National Natural Science Foundation of China (No.61971005), the R&D Program of the Shaanxi Province of China (No.2022GY-064).
REFERENCES

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In European conference on computer vision. Springer, 213–229.

[2] Wenjie Chang, Yuexi Zhang, and Zhewei Xiong. 2021. Transformer-based Monocular Depth Estimation with Attention Supervision. (2021).

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020).

[4] David Eigen, Christian Puhrsch, and Rob Fergus. 2014. Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems 27 (2014).

[5] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. 2018. Deep ordinal regression network for monocular depth estimation. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2002–2011.

[6] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyang Fang, Yongjeng Zhang, Hongsheng Li, and Yu Qiao. 2021. Clip-adapter: Better vision-language models with feature adapters. arXiv preprint arXiv:2110.04544 (2021).

[7] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. 2013. Vision meets robotics: The kitti dataset. The International Journal of Robotics Research 32, 11 (2013), 1231–1254.

[8] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. 2017. Unsupervised monocular depth estimation with left-right consistency. In Proceedings of the IEEE conference on computer vision and pattern recognition. 270–279.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. 770–778.

[10] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Song, Zhen Li, and Tom Duerig. 2021. Scaling up visual and visual-language representation learning with noisy text supervision. In International Conference on Machine Learning. PMLR, 4904–4916.

[11] Hualie Jiang, Laiyan Ding, Junjie Hu, and Rui Huang. 2021. PLNet: Plane and Line Priors for Unsupervised Indoor Depth Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 913–922.

[12] Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. 2021. Tip-Adapter: Training-free CLIP-Adapter for Better Vision-Language Modeling. arXiv preprint arXiv:2112.01402 (2021).

[13] Mingliang Zhang, Xinchen Ye, Xin Fan, and Wei Zhong. 2020. Unsupervised depth estimation from monocular videos with hybrid geometric-redefined loss and contextual attention. Neurocomputing 379 (2020), 250–261.

[14] Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. 2021. Tip-Adapter: Training-free CLIP-Adapter for Better Vision-Language Modeling. arXiv preprint arXiv:2111.03930 (2021).

[15] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. 2021. Pointclip: Point cloud understanding by clip. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 8552–8562.

[16] Zhenyu Li, Zehui Chen, Xianming Liu, and Junjun Jiang. 2022. DepthFormer: Exploiting Long-Range Correlation and Local Information for Accurate Monocular Depth Estimation. arXiv preprint arXiv:2203.14211 (2022).

[17] Mingyuan Mao, Renrui Zhang, Honghui Zheng, Teli Ma, Yan Peng, Ernui Ding, Baoshang Zhang, Shumin Han, et al. 2021. Dual-stream network for visual recognition. Advances in Neural Information Processing Systems 34 (2021).

[18] Maren Mel, Muhammad Siddiqui, and Pietro Zunzugh. 2022. End-to-end Learning for Joint Depth and Image Reconstruction from Diffracted Rotation. arXiv preprint arXiv:2204.07076 (2022).

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning. PMLR, 8748–8763.

[20] Yongming Rao, Weni Li, Peng Gao, and Hongsheng Li. 2021. CLIP-Adapter: Better vision-language models for image recognition at scale. In Proceedings of the IEEE conference on computer vision and pattern recognition. 741–750.

[21] Ashutosh Saxena, Min Sun, and Andrew N. Ng. 2008. Make3D: Depth Perception from a Single Still Image.. In Aaai, Vol. 3. 1571–1576.

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020).

[23] Mingliang Zhang, Xinchen Ye, Xin Fan, and Wei Zhong. 2020. Unsupervised depth estimation from monocular videos with hybrid geometric-redefined loss and contextual attention. Neurocomputing 379 (2020), 250–261.

[24] Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. 2021. Tip-Adapter: Training-free CLIP-Adapter for Better Vision-Language Modeling. arXiv preprint arXiv:2111.03930 (2021).

[25] Wei Zeng, Sicer Karazoglou, and Theo Gevers. 2018. Inferring point clouds from single monocular images by depth intermediation. arXiv preprint arXiv:1812.01402 (2018).

[26] Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. 2021. Tip-Adapter: Training-free CLIP-Adapter for Better Vision-Language Modeling. arXiv preprint arXiv:2111.03930 (2021).

[27] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. 2021. Pointclip: Point cloud understanding by clip. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 8552–8562.

[28] Renrui Zhang, Han Qiu, Tai Wang, Xuanzhuo Xu, Ziyu Guo, Yu Qiao, Peng Gao, and Hongsheng Li. 2021. MonoODET: Depth-aware Transformer for Monocular 3D Object Detection. arXiv preprint arXiv:2203.13310 (2022).

[29] Renrui Zhang, Longtian Qiu, Wei Zhang, and Ziyao Zeng. 2021. VT-CLIP: Enhancing Vision-Language Models with Visual-guided Texts. arXiv preprint arXiv:2112.02399 (2021).

[30] Chong Zhou, Chen Change Loy, and Bo Dai. 2021. DenseCLIP: Extract Free Dense Labels from CLIP. arXiv preprint arXiv:2112.01071 (2021).

[31] Kaiyong Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2021. Learning to prompt for vision-language models. arXiv preprint arXiv:2109.01134 (2021).

[32] Xinyu Zhou, Rohit Girishkar, Armand Joulin, Phillip Krähenbühl, and Ishan Misra. 2022. Detecting Twenty-thousand Classes using Image-level Supervision. arXiv preprint arXiv:2201.02605 (2022).