



# DepthCLIP

Renrui Zhang<sup>\*1</sup>, Ziyao Zeng<sup>\*2</sup>, Ziyu Guo<sup>1</sup>

<sup>1</sup>Peking University, <sup>2</sup>ShanghaiTech University

Accepted by ACM Multimedia 2022 (Brave New Idea)

<sup>\*</sup>indicates equal contributions

# Outline

- 1. Task
- 2. Related Works
- 3. Our DepthCLIP
- 4. Results & Analysis & Limitations & Future direction
- 5. Conclusion

# Task

- Zero-shot Training-free Monocular Depth Estimation
  - Monocular Depth Estimation:
    - Infer pixel-wise depth from monocular images
    - Important in industrial field
      - Self-driving cars need to infer depth to conduct 3D object detection from monocular images
      - Since Lidar is expensive and stereo cameras are hard to adjust.
  - Zero-shot Training-free Transfer
    - When transfer pre-trained model to a new dataset, we require no extra data nor extra training
      - To achieve efficient and effective transfer
    - Important in industrial field
      - Self-driving cars need to infer depth when entering a complete new environment, and they might have no data nor time to finetune its model

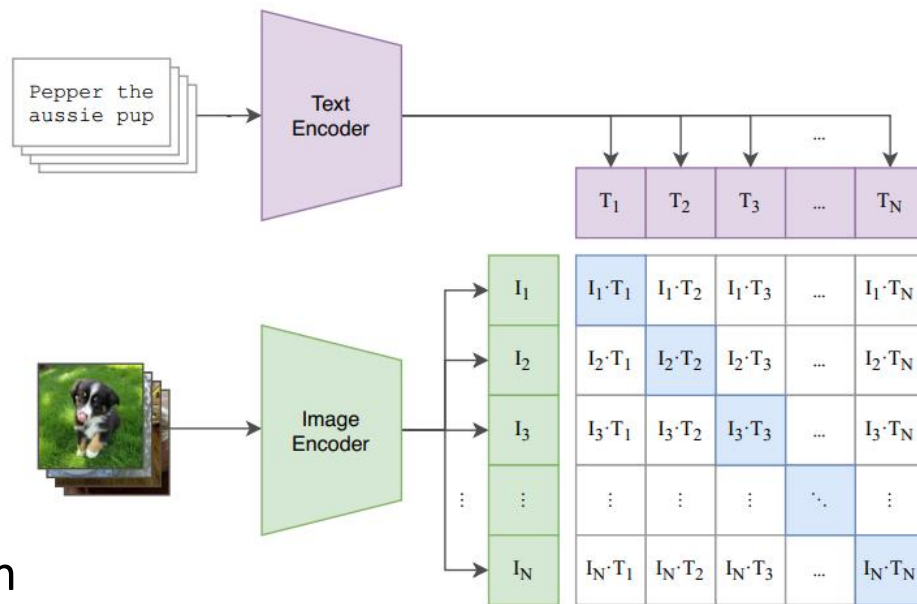
# Related Work

- Monocular Depth Estimation:
  - Fully Supervised Method:
    - ASTRansformer, DORN, RPSF, ....
    - Require pixel-wise depth annotation, costly
  - Unsupervised Method:
    - Learn from ego-motion of unlabeled monocular video.
    - Trains itself with widely available binocular stereo images
      - Generate paired stereo images of given monocular images, exploit epipolar geometry to solve depth
    - Require special modality of data
- Our DepthCLIP
  - Differs from all works in this filed
  - Pretrained with an image classification pre-text task

# Related Work: CLIP

## Contrastive Language-Image Pretraining

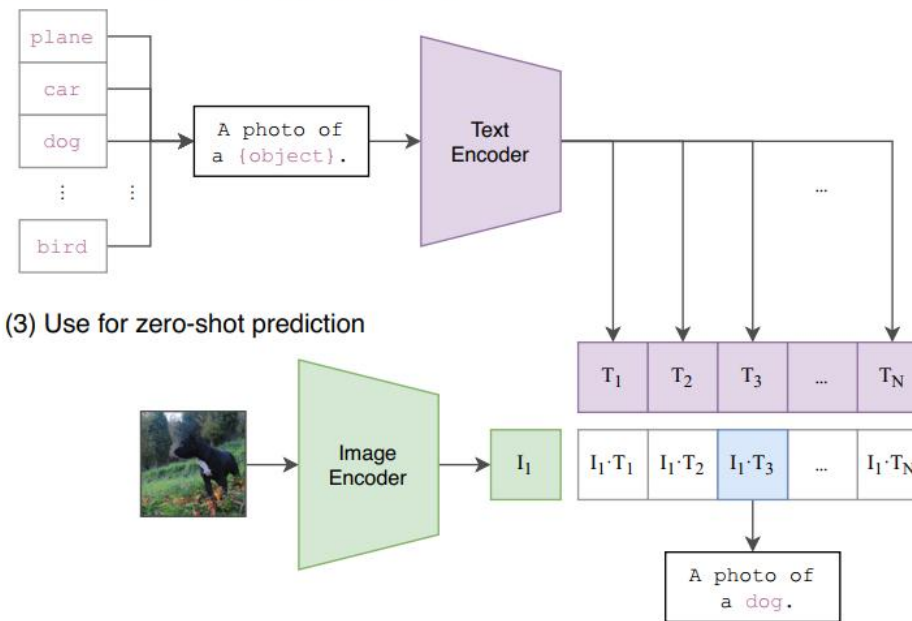
(1) Contrastive pre-training



Group images and text with similar semantic information.

i.e. Understand image using text semantic

(2) Create dataset classifier from label text



(3) Use for zero-shot prediction

### • Train

- In each batch, encode images and corresponding text into feature space
- Maximize cosine similarity between corresponding pair (dog image with “dog”)
- Minimize the rest (dog image with “cat” or “cake”)

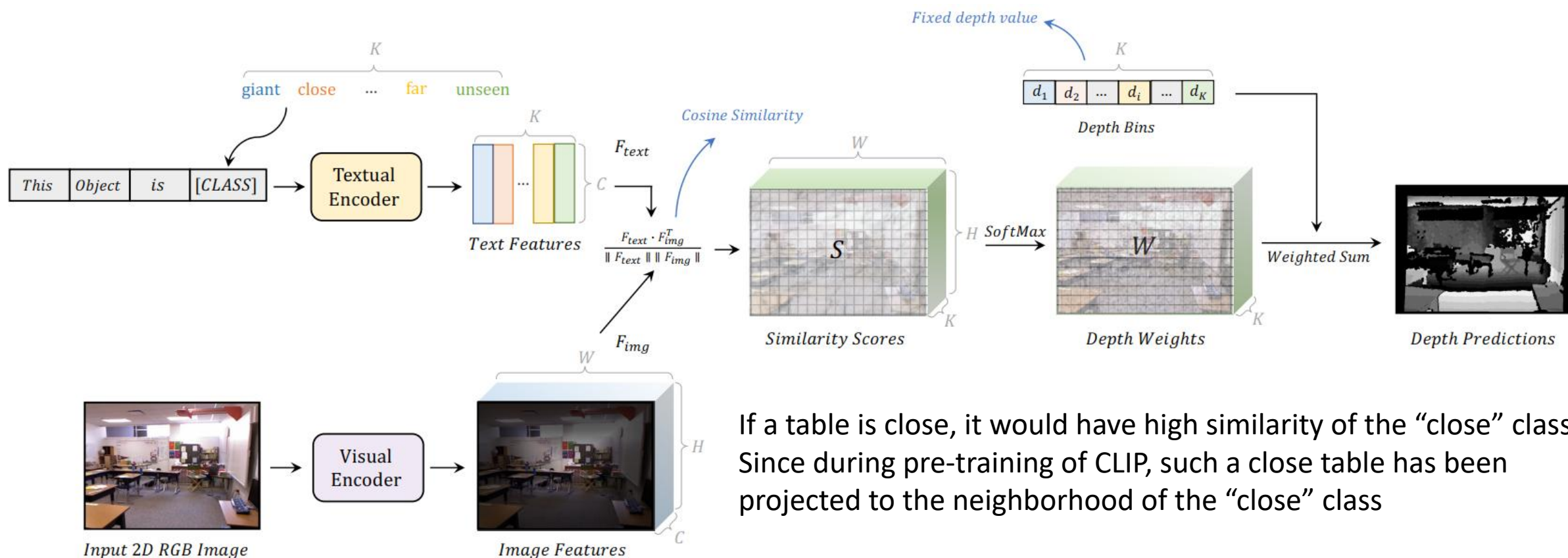
### • Test

- Form prompts, like “A photo of [class]”, substitute all classes in test dataset
- In each batch, calculate similarity between the test image with all prompts
- Choose the class who has max prompt-image similarity, as prediction

The dog image would be projected to the neighborhood of “dog”, thus has high similarity with “dog”.

# DepthCLIP Pipeline

CLIP has the ability to attach each image with corresponding text with similar semantic meaning.



# Quantified Results

- Exceeds the mathematical lower bound significantly, surpasses some existing unsupervised methods, and even draws near some fully-supervised methods.

Method	Supervision	Pre-training	Transfer	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$	rel $\downarrow$	$\log_{10} \downarrow$	rmse $\downarrow$
Make3D[23]	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
ASTransformer[2]	depth	-	-	0.902	0.985	0.997	0.103	0.044	0.374
DepthFormer[14]	depth	-	-	0.921	0.989	<b>0.998</b>	0.096	0.041	0.339
RPSF[20]	depth	-	-	<b>0.952</b>	<b>0.989</b>	0.997	<b>0.072</b>	<b>0.029</b>	<b>0.267</b>
Lower Bound	-	-	-	0.140	0.297	0.471	1.327	0.323	2.934
vid2depth[18]	unsupervised	KITTI monocular video[7]	0-shot	0.268	0.507	0.695	0.572	-	1.637
Zhang et al.[29]	unsupervised	KITTI monocular video[7]	0-shot	0.350	0.617	0.799	0.513	0.529	1.457
<b>Ours-DepthCLIP</b>	language	CLIP[21]	0-shot	<b>0.394</b>	<b>0.683</b>	<b>0.851</b>	<b>0.388</b>	<b>0.156</b>	<b>1.167</b>

**Table 1: Results of Monocular Depth Estimation on NYU Depth v2[24]. 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.**

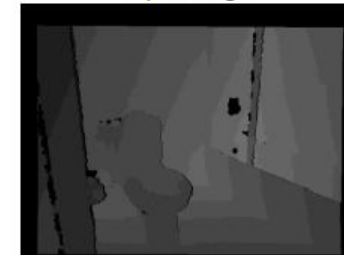


# Results Visualization

- Could tell contours, predict resonable depth
  - The background and details are blurred.
  - CLIP is pre-trained under a classification pre-text task
  - Details and background that are unimportant for classification would be neglected during feature extraction.
  - In the future, we could explore pre-trained model with regional pre-text tasks like segmentation



Input Image



Ground Truth



Our Predication



Input Image



Ground Truth

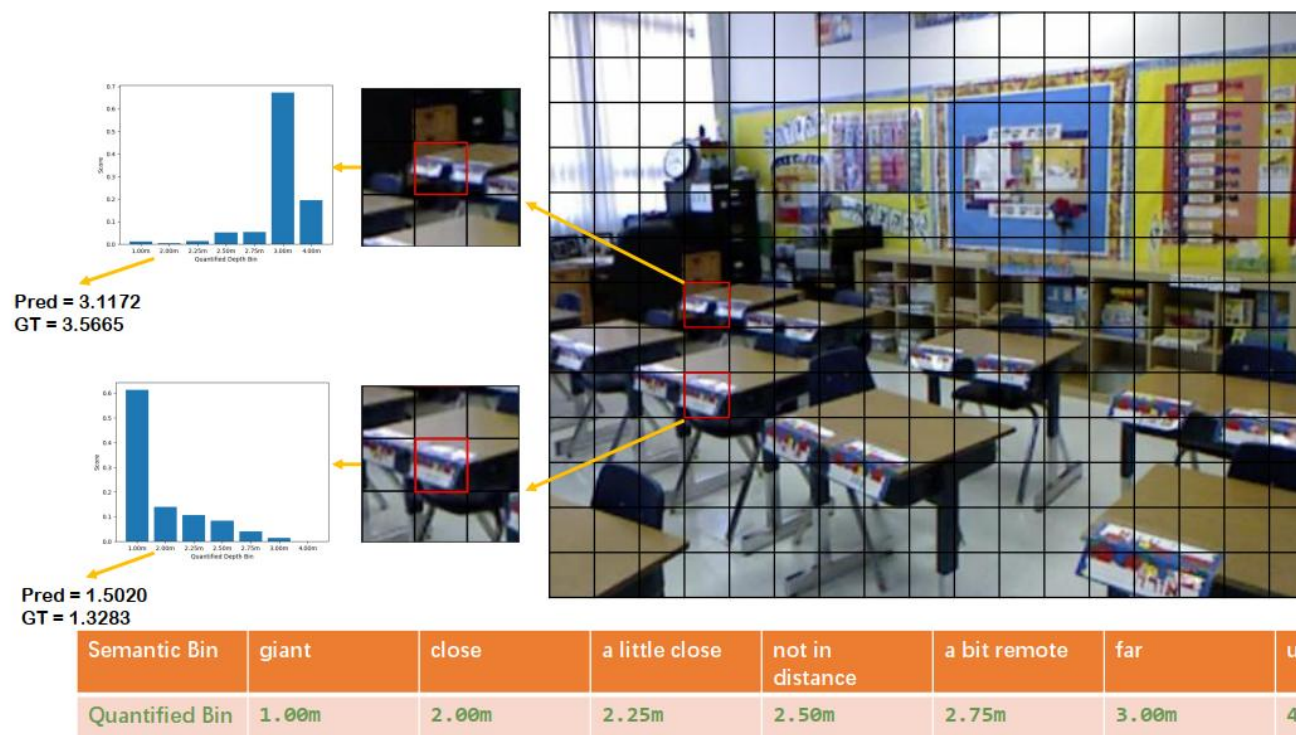


Our Predication



# Semantic Bin Response

- Different patches have different distance semantic responses
  - Depth CLIP could distinguish between a close patch and a remote patch, and make proper distance response.



# Depth Distribution Gap

- Different scenes have different depth distribution
- The same depth class should be projected to different depth bins in different scenes.
- In major experiments, we project the same depth class to the same depth bin.



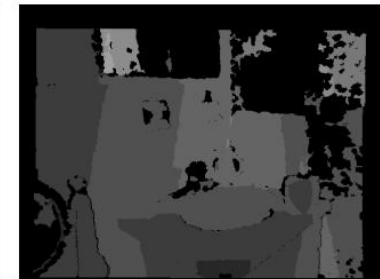
RGB Image of Classroom



RGB Image of Bathroom



Depth Map of Classroom



Depth Map of Bathroom

# Class-dependent Depth Bin Ablation

- DepthCLIP is sensitive to depth bin
- Set different bins for different scenes could improve performance
- In the future, we could predict scene of the input image first, then use a learnable class-dependent depth bin to achieve a better performance.

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]

Class: Bathroom	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$	rel $\downarrow$	$\log_{10} \downarrow$	rmse $\downarrow$
Original bin	0.333	0.631	0.814	0.549	0.175	0.922
Class-dependent 1	0.248	0.490	0.699	0.754	0.219	1.237
Class-dependent 2	0.236	0.460	0.675	0.801	0.229	1.308
<b>Class-dependent 3</b>	<b>0.425</b>	<b>0.723</b>	<b>0.893</b>	<b>0.373</b>	<b>0.141</b>	<b>0.745</b>
Class-dependent 4	0.129	0.302	0.535	1.072	0.287	1.682
Best partition's gain	+0.092	+0.092	+0.079	-0.176	-0.034	-0.177

Class: Classroom	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$	rel $\downarrow$	$\log_{10} \downarrow$	rmse $\downarrow$
Original bin	0.308	0.533	0.742	<b>0.372</b>	0.193	1.826
Class-dependent 1	<b>0.312</b>	0.565	0.820	0.383	0.179	1.694
Class-dependent 2	0.310	0.583	0.830	0.397	0.175	1.636
Class-dependent 3	0.231	0.452	0.600	0.407	0.246	2.138
<b>Class-dependent 4</b>	0.276	<b>0.637</b>	<b>0.844</b>	0.461	<b>0.173</b>	<b>1.544</b>
Best partition's gain	-0.032	+0.104	+0.102	+0.088	-0.020	-0.282

Class: All	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$	rel $\downarrow$	$\log_{10} \downarrow$	rmse $\downarrow$
<b>Original bin</b>	<b>0.394</b>	<b>0.683</b>	<b>0.851</b>	0.388	<b>0.156</b>	<b>1.167</b>
Class-dependent 1	0.373	0.653	0.828	0.467	0.166	1.228
Class-dependent 2	0.366	0.641	0.819	0.496	0.170	1.248
Class-dependent 3	0.333	0.621	0.818	<b>0.353</b>	0.176	1.290
Class-dependent 4	0.288	0.548	0.752	0.663	0.201	1.439
Best partition's gain	-	-	-	-	-	-

# Prompts Ablation

- Robust to prompt design
- Different prompts could catch the same distance relationship, since only relative distance matters

Prompt number	Prompt design details (in semantic token words)					
Original prompt	['giant', 'extremely close', 'close', 'not in distance', 'a little remote', 'far', 'unseen']					
Prompt 1	['extremely close', 'close', 'middle', 'a little far', 'far', 'quite far', 'unseen']					
Prompt 2	['extremely close', 'very close', 'close', 'a little close', 'a little far', 'far', 'unseen']					
Prompt 3	['giant', 'close', 'a little close', 'not in distance', 'a bit remote', 'far', 'unseen']					

Prompt number	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$	rel $\downarrow$	$\log_{10} \downarrow$	rmse $\downarrow$
<b>Original prompt</b>	<b>0.394</b>	<b>0.683</b>	<b>0.851</b>	0.388	<b>0.156</b>	<b>1.167</b>
Prompt 1	0.341	0.623	0.816	0.379	0.175	1.274
Prompt 2	0.377	0.667	0.845	0.385	0.161	1.196
Prompt 3	0.380	0.670	0.846	<b>0.375</b>	0.160	1.196

# Conclusion

- Without any further training, DepthCLIP could surpass some existing unsupervised methods and even approach some fully-supervised networks.
  - We are the first to conduct zero-shot training-free adaptation from the semantic language knowledge possessed by a pre-trained model (CLIP), to a downstream task that needs quantified knowledge (monocular depth estimation).
  - Hope our work could cast a light on the research of bridging semantic vision-language knowledge to the quantified task.