Appendix

Due to space limitation of the main paper, we provide supplementary results and details in the appendix, including: 1) Supplementary related work, 2) Supplementary implementation detail, 3) Supplementary results and analysis, 4) Limitation analysis. We also provide the visualization results of OS-3DETIC together with the baseline and ground truth in the same folder as a supplementary video.

A Supplementary Related Work

A.1 open-set Detection

open-set object detection targets to detect the novel classes that are never provided labels during the training Bansal et al. (2018); Gu et al. (2021); Rahman et al. (2020a;b); Zhou et al. (2022); Zareian et al. (2021); Radford et al. (2021). The classic open-set object detection method directly replaces the classifier with language embedding layer Bansal et al. (2018). To advance the embedding layer, more popular approaches aim at leveraging image-text pairs to extract the rich semantics from text thus broadening the detector Radford et al. (2021); Gu et al. (2021); Zareian et al. (2021). Existing open-set 3D detection Cen et al. (2021; 2022); Wong et al. (2020) is a similar but different setting, compared with our open-set 3D detection. The "Detection" in Open-set 3D Object Detection is to identify the unknown objects from known ones. It does not classify each unknown object into specific categories. However, the "Detection" in our setting (open-set 3D detection) is to localize and classify each object with a specific bounding box and category. The most similar work to us is Detic Zhou et al. (2022), which utilizes ImageNet21K to broaden the classifier of the 2D detector. Yet, it is infeasible to directly use the same method to broaden the classifier of the point-cloud detector, due to the large gap between the image and point-cloud. Different from Detic which transfers knowledge from ImageNet with the image-level class to 2D detection within the same modality, we propose to transfer the knowledge from ImageNet to a totally different modality, point-cloud, with customized pseudo-label strategy and de-biased cross-modal contrastive learning.

A.2 Point-Cloud Detection

Early works on point-cloud object detection discretize and project points onto Bird's Eye View (BEV) or front-view images, and process 2D Lidar feature using standard 2D CNN networks, such as PIXOR Yang et al. (2018), MV3D Chen et al. (2017), SqueezeSeg Wu et al. (2018; 2019); Xu et al. (2020). A more natural way is to directly process each point using PointNet-alike backbones such as PointRCNN Shi et al. (2019) and PointFusion Xu et al. (2018), which is, however, limited by its high computation cost Xu et al. (2021b). Recent popular method is the voxel representation Zhou & Tuzel (2018), which can not only be processed efficiently using 3D sparse convolution Yan et al. (2018); Shi et al. (2020b;a), but also preserve approximately similar information to raw point-cloud with small voxel size. Recently, vision transformer dominates the field of image field Dosovitskiy et al. (2020); Wu et al. (2020); Liu et al. (2021c), and point-cloud transformer is also gradually developed Misra et al. (2021); Zhao et al. (2021). Our method is based on 3DETR Misra et al. (2021).

A.3 Zero-Shot Learning in Point-cloud

Previous zero-shot (open-set) learning works in the point-cloud field mainly study classification. Image2Point Xu et al. (2021a) directly inflates the 2D model pre-trained on large-scale image dataset, and shows a significant improvement for point-cloud classification. PointCLIP Zhang et al. (2021) leverages CLIP pre-trained embedding to broaden vocabulary of the point-cloud classifier. In Cheraghian et al. (2019b;a; 2021), PointNet is pre-trained on seen classes and classifies unseen objects by calculating the similarity to the seen class. Recently, although zero-shot semantic segmentation in point-cloud is studied Liu et al. (2021a); Michele et al. (2021).

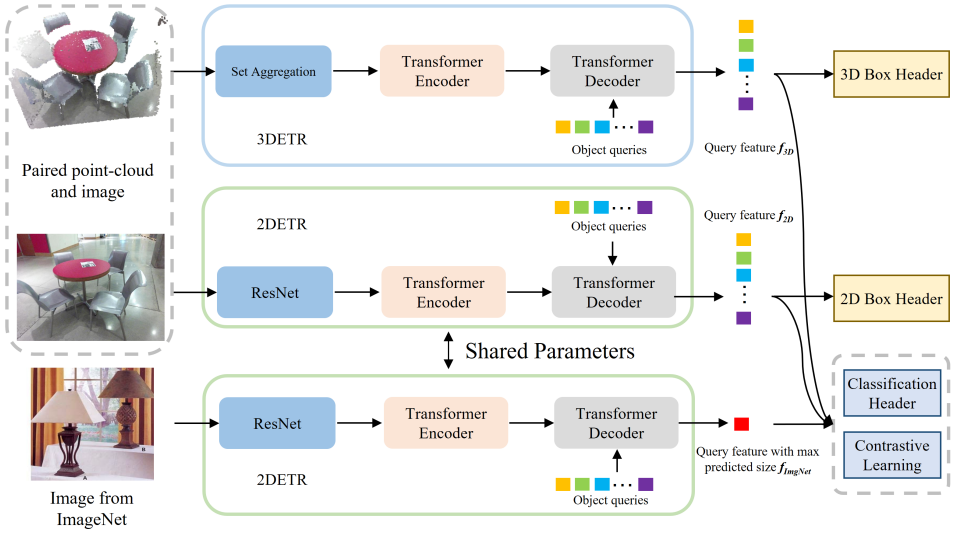


Figure 6: Network architecture of OS-3DETIC

Table 4: Supplementary baselines (AP_{25}) on unseen classes of SUN RGB-D.

Method	toilet	bed	chair	bathtab	sofa	dresser	scanner	fridge	lamp	desk	mean
Detic-ImageNet-box	4.18	0.08	2.69	0.12	0.08	1.08	0.06	1.20	0.08	6.97	1.65
Detic-ModelNet-box	4.52	3.37	3.86	0.10	1.09	4.56	0.13	0.25	0.27	2.45	2.06

B Supplementary Implementation Detail

B.1 Network Architecture

Fig. 6 illustrates the network architecture of OS-3DETIC. The overall network consists of three branches: the point-cloud branch, the paired image branch, and the ImageNet branch. 3DETR Misra et al. (2021) is used as the backbone of the point-cloud branch, which composes of Set Aggregation, Transformer Encoder, and Transformer Decoder. DETR Carion et al. (2020) is used as the backbone of paired image and ImageNet branch, note that the parameters are shared between these two branches. Besides, there are four heads with four different losses concatenate after 3DETR and DETR. 3D and 2D box head are used to regress bounding boxes, while the classification head that is shared among three branches are used to predict category label. Besides, the contrastive learning head aggregates features from three branches and align the distribution among them.

B.2 Data Preprocessing for ScanNet

ScanNet Dai et al. (2017) is a richly annotated dataset of 3D reconstructed meshes of indoor scenes. It contains ∼1.2K training examples collected from hundreds of different scenes and annotated with semantic and instance segmentation labels. Since each reconstructed scene in ScanNet is paired with multiple RGB images, while OS-3DETIC only consumes a single paired image at a time, thus we use the raw data of ScanNet instead of processed meshes. And finally, there are ∼4.6K and ∼4.5K samples in the train and test set, respectively. Furthermore, to generate oriented bounding for ScanNet, we first calculate the center and size of each instance from the instance map and then use principal component analysis (PCA) to compute the heading angle for each bounding box. Other than the above differences, we follow the data preprocessing as VoteNetQi et al. (2019a) does.

Table 5: Supplementary results (AP_{25}) on resampled unseen classes of SUN RGB-D. (Set 1)

Method	bed	bathtub	dresser	fridge	desk	stand	counter	bookshelf	microwave	stool	mean
3DETR Misra et al. (2021)	0.20	0.02	1.24	0.17	2.68	0.18	0.29	0.19	0.00	0.08	0.51
Ours	**1.35**	**47.25**	**9.51**	**10.17**	**13.39**	**25.96**	**3.65**	**17.28**	**0.71**	**4.58**	**13.39**
Improvement	**+1.15**	**+47.23**	**+8.27**	**+10.00**	**+10.71**	**+25.78**	**+3.36**	**+17.09**	**+0.71**	**+4.5**	**+12.88**

Table 6: Supplementary results (AP_{25}) on resampled unseen classes of SUN RGB-D. (Set 2)

Method	chair	bathtub	sofa	lamp	desk	table	counter	pillow	sink	stool	mean
3DETR Misra et al. (2021)	1.12	0.02	0.22	0.00	0.37	0.19	0.46	0.00	0.18	0.04	0.26
Ours	**4.38**	**45.70**	**3.23**	**2.72**	**7.62**	**10.64**	**10.39**	**4.16**	**27.45**	**2.16**	**11.85**
Improvement	**+3.26**	**+45.68**	**+3.01**	**+2.72**	**+7.25**	**+10.45**	**+9.93**	**+4.16**	**+27.27**	**+2.12**	**+11.59**

Table 7: Supplementary results (AP_{25}) on resampled unseen classes of SUN RGB-D. (Set 3)

Method	toilet	bathtub	sofa	fridge	lamp	table	counter	bin	microwave	stool	mean
3DETR Misra et al. (2021)	2.72	0.05	0.29	0.29	0.02	1.37	0.39	0.42	0.00	0.07	0.56
Ours	**50.97**	**44.30**	**4.73**	**13.60**	**0.01**	**10.27**	**4.34**	**16.11**	**0.72**	**2.54**	**14.76**
Improvement	**+48.25**	**+44.25**	**+4.44**	**+13.31**	**-0.01**	**+8.90**	**+3.95**	**+15.69**	**+0.72**	**+2.47**	**+14.20**

B.3 Perform Open-Set Detection on Fully-Supervised Setting

As shown in Tables 1 and 2, the performance of state-of-the-art detectors is relatively lower than the number that reported in their original paper, which is because we modify their fully-supervised setting to open-set. Specifically, compared with the fully-supervised version, the only difference is that we slightly modify the classification head that can predict both seen class and unseen class, e.g., 10 seen classes and 10 unseen classes, then the output channel size of the classification head is 20. Note that no ground truth is provided for unseen classes during training. During inference, we evaluate the results of unseen classes.

C Supplementary Results and Analysis

C.1 (Newly Added) Results of Other Baselines

As a supplement to the main results in Tables 1 and 2, in this section, we report the results of another two baselines: `Detic-ImageNet-box` and `Detic-ModelNet-box`. These two baselines are the extension of `Detic-ImageNet` and `Detic-ModelNet`, respectively, based on which the suffix `-box` denotes that the ground truth bounding boxes of both seen and unseen categories are used during training. Intuitively, these two baselines handle localization by ground truth annotation, and address classification via using ImageNet or ModelNet as Detic Zhou et al. (2022) does. Results are presented in Table 4, which demonstrates that using bounding box supervision further outperforms `Detic-ImageNet` and `Detic-ModelNet` by 0.74% and 0.54%, respectively. However, both of them are inferior to OS-3DETIC by a large gap, even though they use the grounding truth supervision of unseen classes. The results of `Detic-ImageNet` and `Detic-ModelNet` further indicate that due to the large gap between image and point-cloud, directly extends the idea of Detic Zhou et al. (2022) to open-set 3D detection does not perform well.

C.2 Results on Other Randomly Resampled Unseen Classes

In order to verify the robustness against the sampling of unseen classes, we further shuffle and randomly resample multiple sets of unseen classes. As shown in Tables 5 to 7, for different settings of unseen classes, we compare OS-3DETIC with the baseline method (3DETR that trained on seen class). The results show that OS-3DETIC outperforms the baseline and achieves 13.3% mAP_{25} that averages over all sets. Besides, even the worst-performing set (Set 2) overtakes the baseline by a large margin of 11.59%, which demonstrates OS-3DETIC is robust to the sampling of unseen classes. By the way, we observe that there are some classes that perform well across all sets of

Table 8: Supplementary results (AP_{25}) of 3DETR in SUN-RGBD.

Class	toilet	bed	chair	bathtub	sofa	dresser	scanner	fridge	lamp	desk	table	stand	cabinet	counter	bin	bookshelf	pillow	microwave	sink	stool	mean
3DETR	89.13	82.61	66.08	76.84	57.29	26.6	13.1	24.02	25.05	28.34	49.49	60.1	17.24	27.92	45.14	29.51	20.72	9.42	32.03	13.38	39.70

Table 9: Supplementary results (AP_{25}) of 3DETR in ScanNet.

Class	toilet	bed	chair	sofa	dresser	table	cabinet	bookshelf	pillow	sink	bathtub	fridge	desk	stand	counter	door	curtain	box	lamp	bag	mean
3DETR	91.14	64.84	67.82	68.16	35.33	53.8	38.3	43.38	44.57	62.91	77.67	43.86	52.89	52.05	30.78	44.05	37.97	12.37	17.38	9.78	47.45
3DETR-color	91.02	62.49	66.62	68.00	39.37	53.19	36.79	42.02	46.05	65.98	77.11	48.20	50.41	55.54	32.63	46.80	38.03	13.76	19.54	11.62	48.26

unseen classes, such as "toilet" and "bathtub", while there are also some cases that perform poorly, such as "microwave" and "lamp". This phenomenon indicates that the inherent difficulty of each class directly relates to the final results.

C.3 (NEWLY ADDED) RESULTS OF THE CLOSE-SET FULLY-SUPERVISED DETECTOR

Here, we report the fully supervised results of both seen and unseen categories in Tables 8 and 9, which can be regarded as our upper bound. Even though OS-3DEIC outperforms a wide range of baselines, there is still a large gap, compared to the fully supervised setting, indicating the huge space for further improvement.

C.4 (NEWLY ADDED) RESULTS ON ENABLING COLOR INFORMATION

Paired image is used as the intermediary to bridge RGB image and xyz point-cloud, yet, during inference, paired image is not required. In this section, we would like to investigate what if RGB color is available during inference. Specifically, RGB color is used as initial feature of each point which is further processed by 3DETR Misra et al. (2021). Results are presented in Tables 9 and 10, where 3DETR-color denotes 3DETR with the input of colored point-cloud, and OS-3DETIC-color represents OS-3DETIC with the input of colored point-cloud. As we can see, RGB colors marginally help 3DETR and OS-3DETIC. Intuitively, however, RGB colors are very helpful for human recognition of objects. The reason is that 3DETR is not designed for processing color information, the official version of 3DETR only uses xyz geometry, and OS-3DETIC is mainly based on 3DETR. Nonetheless, finding a way to fully exploit RGB colors is an interesting problem. We will investigate this further in the future.

C.5 (NEWLY ADDED) ABLATION ON DISTANCE-AWARE TEMPERATURE

To further uncover how the proposed Distance-Aware Temperature works, we perform this ablation study that sets γ (γ is the base of exponential function) range from 0.5 to 1.5, and the results are shown in Table 11. As we can see, when $\gamma < 1$, which means L_{DECC} focusing on distant objects, the mAP_{25} drops. And when $\gamma = 1$, the distance-aware temperature degenerates to class-based contrastive learning. And when $\gamma > 1$, which means L_{DECC} focusing on close-by objects, the mAP_{25} is improved first and then drops. As we discussed in Section 3.3, intuitively, distance-aware temperature leverages the prior that the correlation between close-by objects is stronger than that of distant objects.

C.6 (NEWLY ADDED) ABLATION ON DECOUPLING PSEUDO-LABEL AND DE-BIASED CROSS-MODAL CONTRASTIVE LEARNING

As we discussed in Section 3.2, "*There is significant synergy between pseudo-label strategy and our proposed de-biased cross-modal contrastive learning*", the pseudo-label is beneficial for true positive sampling of de-biased contrastive learning, and de-biased contrastive learning in turn helps to generate better pseudo-labels. Therefore, the proposed OS-3DETIC works in an iterative scheme that the improvement achieved by the de-biased cross-modal contrastive learning will be merged into the updated pseudo-label.

Table 10: Supplementary results (AP_{25}) of OS-3DETIC-color in ScanNet.

Class	toilet	bed	chair	sofa	dresser	table	cabinet	bookshelf	pillow	sink	mean
OS-3DETIC-color	50.21	2.74	7.33	18.93	2.91	14.44	2.65	4.75	4.35	22.71	13.10

Table 11: Ablation on Distance-Aware Temperature.

γ	0.5	0.7	0.9	1.0	1.1	1.3	1.5
mAP_{25}	11.35	12.27	12.75	12.92	13.03	12.87	12.54

Table 12: Ablation on decoupling Pseudo-Label and DECC. (mAP_{25})

Iteration	0	1	2	3	4
Pseudo-Label	1.16	5.86	9.10	10.45	10.86
Pseudo-Label + DECC	1.16	6.55	10.58	12.35	13.03
Improvement	**+0.00**	**+0.69**	**+1.48**	**+1.90**	**+2.17**

In order to decouple the improvement of pseudo-label and de-biased cross-modal contrastive learning, in this experiment, we disable the de-biased cross-modal contrastive learning altogether. The results are given in Table 12. As we can see, with the updating of pseudo-label, mAP_{25} of both `Pseudo-Label` and `Pseudo-Label + DECC` are improved, while the improving of `Pseudo-Label + DECC` is better than that of `Pseudo-Label`, further demonstrating the effectiveness of de-biased cross-modal contrastive learning.

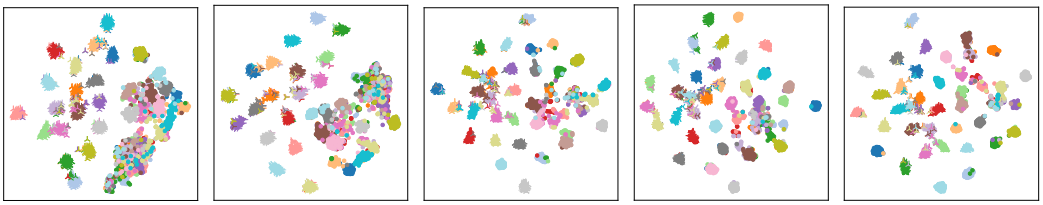
C.7 Visualization of Learned Representations

Inspired by Chuang et al. (2020), we investigate the learned representation via comparison among T-SNEs. Specifically, we compare OS-3DETIC with four different settings, they are baseline setting, baseline and pseudo-label with position-based contrastive learning, baseline and pseudo-label with class-based contrastive learning, and unbiased contrastive learning with ground truth. Please note that the unbiased contrastive learning is the upper bound of OS-3DETIC. Besides, we visualize the feature both before and after the linear layer of contrastive loss. The results are shown in Figs. 7 and 8. The "Dot" and "Triangle" markers denote the feature of point-cloud and images from ImageNet, respectively. Fig. 7 (a) can be mainly divided into three parts: ImageNet features (left), seen classes features of point-clouds (right-top) and unseen classes features of point-clouds (right-down). The comparison among these three parts indicates that, in the baseline setting, ImageNet features are clearly clustered, and that of seen classes could barely distinguish from each other, while those of unseen classes are almost indistinguishable. Moreover, comparing the first four sub-figures of Fig. 7, we find that the "Pseudo-Label", "Class-based contrastive learning" and "Distance-aware temperature" help the clustering of the representations of unseen classes progressively.

Above observations can be also found in the feature distribution after the linear layer, we evaluate the gain of de-biased contrastive learning by observing the ImageNet and point-cloud feature distributions. Specifically, if these two feature distributions are close to each other, then the contrastive strategy works. As shown in Fig. 8 (e), in the unbiased setting, the two feature distributions overlap with each other, and in the biased setting (Fig. 8 (b)), there is a significant difference between these two distributions. Our de-biased setting (Fig. 8 (d)) performs better than the biased setting, while there are still some hard examples that could be improved.

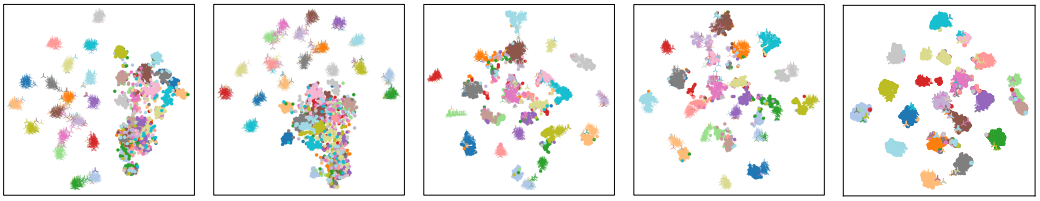
D Limitation Analysis

The proposed OS-3DETIC transfers open-set knowledge from image-level annotated ImageNet to 3D detector. The major limitation is that we assume the calibration matrix between Camera and Depth sensor is available. Even though calibrating the camera and depth sensor is a mature technique, this assumption hinders the using of the massive unpaired image and point-cloud data. Besides, loosely calibrated camera and depth sensor may also lead to mAP loss. It is interesting to investigate knowledge transfer through unpaired image and point-cloud. As we analyzed in Sec-



(a) Baseline (b) Baseline+PL+P (c) Baseline+PL+C (d) Baseline+PL+C+D (e) Unbiased

Figure 7: **T-SNE visualization before linear layer**: "PL, P, C, D" are the abbreviation of "Pseudo-Label", "Position-Based Contrastive Learning", "Class-Based Contrastive Learning" and "Distance Aware Temperature". "Unbiased" represents using the ground truth of unseen classes for contrastive learning. The "Dot" and "Triangle" markers in each figure denote the representations of the point-clouds and the images from ImageNet, respectively. The distinct colors of markers indicate different categories.



(a) Baseline (b) Baseline+PL+P (c) Baseline+PL+C (d) Baseline+PL+C+D (e) Unbiased

Figure 8: **T-SNE visualization after linear layer**: "PL, P, C, D" are the abbreviation of "Pseudo-Label", "Position-Based Contrastive Learning", "Class-Based Contrastive Learning" and "Distance Aware Temperature". "Unbiased" represents using the ground truth of unseen classes for contrastive learning. The "Dot" and "Triangle" markers in each figure denote the representations of the point-clouds and the images from ImageNet, respectively. The distinct colors of markers indicate different categories.

tion 4.3, not all classes achieve the same improvement. The unseen classes can be roughly divided into three groups, they are simple, normal and hard. Take hard categories as example. There are a variety of reasons that lead to these challenges, including but not limited to: too small target, rare samples, clutter background. These challenges also exist in close-set 3D detection.

REFERENCES

Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 384–400, 2018.

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

Jun Cen, Peng Yun, Junhao Cai, Michael Yu Wang, and Ming Liu. Open-set 3d object detection. In *2021 International Conference on 3D Vision (3DV)*, pp. 869–878. IEEE, 2021.

Jun Cen, Peng Yun, Shiwei Zhang, Junhao Cai, Di Luan, Michael Yu Wang, Ming Liu, and Mingqian Tang. Open-world semantic segmentation for lidar point clouds. In *European Conference on Computer Vision (ECCV)*. Springer, 2022.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3D object detection network for autonomous driving. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6526–6534, 2017.

Ali Cheraghian, Shafin Rahman, Dylan Campbell, and Lars Petersson. Mitigating the hubness problem for zero-shot learning of 3d objects. *arXiv preprint arXiv:1907.06371*, 2019a.

Ali Cheraghian, Shafin Rahman, and Lars Petersson. Zero-shot learning of 3d point cloud objects. In *2019 16th International Conference on Machine Vision Applications (MVA)*, pp. 1–6. IEEE, 2019b.

Ali Cheraghian, Shafinn Rahman, Townim F Chowdhury, Dylan Campbell, and Lars Petersson. Zero-shot learning on 3d point cloud objects and beyond. *arXiv preprint arXiv:2104.04980*, 2021.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. *Advances in neural information processing systems*, 33:8765–8775, 2020.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

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

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Zero-shot detection via vision and language knowledge distillation. *arXiv e-prints*, pp. arXiv–2104, 2021.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Ji Hou, Saining Xie, Benjamin Graham, Angela Dai, and Matthias Nießner. Pri3d: Can 3d priors help 2d representation learning? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5693–5702, 2021.

Huajie Jiang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Transferable contrastive network for generalized zero-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9765–9774, 2019.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1097–1105, 2012.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

Bo Liu, Shuang Deng, Qiulei Dong, and Zhanyi Hu. Segmenting 3d hybrid scenes via zero-shot learning. *arXiv preprint arXiv:2107.00430*, 2021a.

Yueh-Cheng Liu, Yu-Kai Huang, Hung-Yueh Chiang, Hung-Ting Su, Zhe-Yu Liu, Chin-Tang Chen, Ching-Yu Tseng, and Winston H Hsu. Learning from 2d: Pixel-to-point knowledge transfer for 3d pretraining. *arXiv e-prints*, pp. arXiv–2104, 2021b.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021c.

Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2929–2938, 2021d.

Björn Michele, Alexandre Boulch, Gilles Puy, Maxime Bucher, and Renaud Marlet. Generative zero-shot learning for semantic segmentation of 3d point clouds. In *2021 International Conference on 3D Vision (3DV)*, pp. 992–1002. IEEE, 2021.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection with vision transformers. *arXiv preprint arXiv:2205.06230*, 2022.

Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2906–2917, 2021.

Jinhyung Park, Chenfeng Xu, Yiyang Zhou, Masayoshi Tomizuka, and Wei Zhan. Detmatch: Two teachers are better than one for joint 2d and 3d semi-supervised object detection. *arXiv preprint arXiv:2203.09510*, 2022.

C. Qi, Or Litany, Kaiming He, and Leonidas J. Guibas. Deep hough voting for 3d object detection in point clouds. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9276–9285, 2019a.

Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9277–9286, 2019b.

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

Shafin Rahman, Salman Khan, and Nick Barnes. Improved visual-semantic alignment for zero-shot object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 11932–11939, 2020a.

Shafin Rahman, Salman H Khan, and Fatih Porikli. Zero-shot object detection: joint recognition and localization of novel concepts. *International Journal of Computer Vision*, 128(12):2979–2999, 2020b.

Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Point-RCNN: 3d object proposal generation and detection from point cloud. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–779, 2019.

Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. PV-RCNN: Point-voxel feature set abstraction for 3d object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10529–10538, 2020a.

Shaoshuai Shi, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2020b.

Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 567–576, 2015.

Kelvin Wong, Shenlong Wang, Mengye Ren, Ming Liang, and Raquel Urtasun. Identifying unknown instances for autonomous driving. In *Conference on Robot Learning*, pp. 384–393. PMLR, 2020.

Bichen Wu, Alvin Wan, Xiangyu Yue, and Kurt Keutzer. SqueezeSeg: Convolutional neural nets with recurrent CRF for real-time road-object segmentation from 3d lidar point cloud. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1887–1893, 2018.

Bichen Wu, Xuanyu Zhou, Sicheng Zhao, Xiangyu Yue, and Kurt Keutzer. Squeezesegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud. In *ICRA*, 2019.

Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision. *arXiv preprint arXiv:2006.03677*, 2020.

Bichen Wu, Ruizhe Cheng, Peizhao Zhang, Peter Vajda, and Joseph E Gonzalez. Data efficient language-supervised zero-shot recognition with optimal transport distillation. *arXiv preprint arXiv:2112.09445*, 2021.

Chenfeng Xu, Bichen Wu, Zining Wang, Wei Zhan, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. Squeezesegv3: Spatially-adaptive convolution for efficient point-cloud segmentation. In *European Conference on Computer Vision (ECCV)*, pp. 1–19. Springer, 2020.

Chenfeng Xu, Shijia Yang, Bohan Zhai, Bichen Wu, Xiangyu Yue, Wei Zhan, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. Image2point: 3d point-cloud understanding with pretrained 2d convnets. *arXiv preprint arXiv:2106.04180*, 2021a.

Chenfeng Xu, Bohan Zhai, Bichen Wu, Tian Li, Wei Zhan, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. You only group once: Efficient point-cloud processing with token representation and relation inference module. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4589–4596. IEEE, 2021b.

Danfei Xu, Dragomir Anguelov, and Ashesh Jain. PointFusion: Deep sensor fusion for 3D bounding box estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018.

Bin Yang, Wenjie Luo, and Raquel Urtasun. PIXOR: Real-time 3d object detection from point clouds. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7652–7660, 2018.

Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14393–14402, 2021.

Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. *arXiv preprint arXiv:2112.02413*, 2021.

Zaiwei Zhang, Bo Sun, Haitao Yang, and Qi-Xing Huang. H3dnet: 3d object detection using hybrid geometric primitives. In *ECCV*, 2020.

Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16259–16268, 2021.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Phillip Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. *arXiv preprint arXiv:2201.02605*, 2022.

Yin Zhou and Oncel Tuzel. VoxelNet: End-to-end learning for point cloud based 3d object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.