X-Trans2Cap: Cross-Modal Knowledge Transfer using Transformer for 3D Dense Captioning
Supplementary Material

A. Overview

In this supplementary material, we illustrate the implementation details, the efficiency of the model and the results of subjective evaluation in Section B, Section C and Section D, respectively. After that, we provide more experiments of Scan Dense Captioning (DC) on Nr3D dataset in Section E. Then we discuss the effectiveness of each attribute in instance representation in Section F.

B. Implementation Details

In our experiment, we adopt the PointNet++ to generate 3D object features ($O_m^{f_{3d}}$) in Oracle DC, and applies proposals' features from VoteNet in the Scan DC. Furthermore, in the test with Oracle DC, we use ground truth category as O_m^{cls} while adopting the predicted results from detector in Scan DC task. We train the network for 30 epochs by using Adam optimizer with a batch size of 32. The probability of random mask in CMF module is set as 0.2 when achieving the best, and it does not greatly change the result. It should be noted that both teacher and student networks are trained from scratch. The learning rate is initialized as 0.0005 with the decay as 0.1 for every 10 epochs. Experiments are conducted on RTX2080Ti GPUs.

C. Running Time Evaluation

We investigate the running time of our model in this section. Table 1 shows the number of parameters and inference time of per scan in Oracle DC setting. X-Trans2Cap (3D) can speed up more than **20**× compared with its baseline and X-Trans2Cap using extra 2D modality.

D. Subjective Evaluation

We conduct a subjective evaluation with three volunteers on randomly selected 100 descriptions generated by Scan2Cap and X-Trans2Cap with *Oracle DC* setting on ScanRefer datasets. The subjective evaluation results are shown in Table 2. In practice, each volunteer is asked to manually check whether the descriptions correctly reflect two aspects of the object: object color attributes and spatial relations in local environment. As observed from Table 2, X-Trans2Cap can generate more faithful captions regarding the attributes and spatial relationships.

Table 1. The complexity analysis between X-Trans2Cap using both 3D and 2D inputs and only 2D input. Here underline correspond to the time and parameters for the 2D feature extractor.

Method	2D	#Param (M)	Inference (s)
TransCap	✗	19.9	0.4
TransCap	✓	60.0+19.9	8.1+0.4
X-Trans2Cap	✗	**19.9**	**0.4**
X-Trans2Cap	✓	60.0+38.8	8.1+0.9

Table 2. Subjective evaluation in Oracle DC setting. We measure the accuracy of two aspects (object colors and spatial relations) in the generated captions.

Design	Extra 2D	Attribute	Relation
Scan2Cap	✗	61.82	66.86
X-Trans2Cap	✗	**68.73** (+6.91)	**75.54** (+8.68)
Scan2Cap	✓	64.21	69.00
X-Trans2Cap	✓	**70.12** (+5.91)	**78.97** (+9.97)

E. Scan Dense Captioning on Nr3D

In Table 3, we compare the results of Scan DC on Nr3D, including the results without and with extra 2D input in the inference phase. All methods exploit the same network, *i.e.*, VoteNet, to generate proposals. *3D-2D Proj.* projects proposals back to 2D images and captions in a 2D manner. However, it achieves the lowest captioning scores, which reveals that it cannot directly handle the 3D dense captioning task. Though Scan2Cap achieves better results than *3D-2D Proj.*, it also cannot generate faithful captioning results. Not surprisingly, X-Trans2Cap obtains the highest score in all metrics. Specifically, it not only gains a +2.9 improvement in CIDEr@0.25 score upon baseline TransCap, but also achieves +5.5 boost over Scan2Cap. Finally, the experiment also confirms that our X-Trans2Cap can improve 3D visual detection as well.

Table 3. Comparison of 3D dense captioning obtained by X-Trans2Cap and previous methods, taking 3D Scans as the input on Nr3D dataset. We average the scores of the above captioning metrics, which are with the IoU percentage between the predicted bounding box and the ground truth over 0.25 and 0.5, respectively. The 'Extra 2D' means that whether using the extra 2D modality as above. *3D-2D Proj.* represents the method in Scan2Cap, *i.e.*, 3D proposal projected to 2D images. 'Proposals' shows the methods exploited to obtain 2D or 3D proposals.

Method	Extra 2D	Proposals	C@0.25	B-4@0.25	M@0.25	R@0.25	C@0.5	B-4@0.5	M@0.5	R@0.5	mAP@0.5
Scan2cap	✗	VoteNet	41.76	24.12	24.98	55.79	23.70	14.88	20.95	47.50	32.17
TransCap	✗	VoteNet	44.32	25.63	25.25	55.69	27.24	17.76	21.60	49.16	34.09
X-Trans2Cap	✗	VoteNet	**47.26**	**27.38**	**25.45**	**56.28**	**30.96**	**18.70**	**22.15**	**49.92**	**34.13**
3D-2D Proj.	✓	VoteNet	8.57	8.49	18.83	44.95	3.93	4.21	16.68	41.24	31.83
Scan2cap	✓	VoteNet	42.24	24.43	25.07	55.88	24.10	15.01	21.01	47.95	32.21
TransCap	✓	VoteNet	45.06	25.79	25.22	55.55	33.45	19.09	22.24	50.00	33.71
X-Trans2Cap	✓	VoteNet	**51.43**	**27.62**	**25.75**	**56.46**	**33.62**	**19.29**	**22.27**	**50.00**	**34.38**

Table 4. Ablation study for applying different instance representation designs. The results are obtained in Oracle DC on the ScanRefer dataset. The upper part shows ablated results for different student input design, and the lower illustrates results using specific input for teacher network.

Model	Teacher Network						Student Network				Metrics			
	O^{f3d}	O^{cls}	O^{b3d}	O^{pe}	O^{f2d}	O^{b2d}	O^{f3d}	O^{cls}	O^{b3d}	O^{pe}	C	B-4	M	R
A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	**87.09**	**44.12**	**30.67**	64.37
B	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	70.41	39.98	28.70	62.09
C	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	82.99	43.39	30.22	**64.57**
D	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	33.52	35.67	26.33	61.78
E	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	86.71	43.92	30.54	64.32
F	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	84.23	43.43	30.24	64.33
G	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	83.85	43.12	30.24	64.52

F. Analysis and Ablation Studies

We further conduct an ablation study on different instance representation designs as shown in Table 4, where the upper part and lower part show the specific designs in teacher and student network, respectively.

Does object class help? From the results of model B in Table 4, it can be found out that there is a dramatic drop in metric of CIDEr, from 87.09 to 70.41, when we discard the object class O^{cls}. Thus, it shows that O^{cls} is an important attribute for instance representation. Note that the ablated model B is still +6 CIDEr higher than that of Scan2Cap.

Does 3D bounding box help? As shown in results of model C in Table 4, removing the 3D bounding box O^{b3d} will not cause a large performance drop, only -4.1 CIDEr from 87.09 to 82.99. This result reflects that X-Trans2Cap utilizes 3D object spatial coordinates to generate captions.

Does positional encoding help? The result of model D demonstrates a tremendous performance decrease in metric of CIDEr when positional encoding O^{pe} is not exploited, where the model can only obtain 33.52 in metric of CIDEr. Since our model only chooses one object as the target object and the remaining ones will be regarded as reference objects, positional encoding helps the model to identify the target one. Without its help, the network can hardly work.

Does 2D input help? The lower part of Table 4 describes the effectiveness of different attributes in the teacher network. There are three conclusions can be obtained: **1)** Discarding 3D features O^{f3d} in teacher network barely hampers the performance (model E). This is because the 3D features also exist in the input of the student network. **2)** Utilizing the pre-trained network to extract 2D features is not necessary (model F). The result of model F shows that even if we only exploit the information of 2D bounding box, there is only an about -2 CIDEr drop for the caption results. **3)** The 2D bounding box information O^{b2d} seems to play a more important role compared with 2D features O^{f2d} (see the model G). Without using O^{b2d}, the model only obtains 83.85 CIDEr, and this result is even 0.4 lower than that of model F (without using O^{f2d}). Such results also emphasize the capability of X-Trans2Cap in real-world applications, *i.e.*, without pre-trained 2D network, only utilizing the 2D bounding box information can still greatly boost the captioning performance.