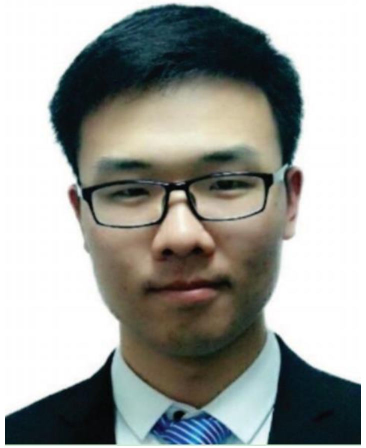
Y^2Seq2Seq: Cross-Modal Representation Learning for 3D Shape and Text by Joint Reconstruction and Prediction of View and Word Sequences





Zhizhong Han[1,2] **Mingyang Shang**[1] **Xiyang Wang**[1] **Yu-Shen Liu**[1] **Matthias Zwicker**[2]

[1]School of Software, Tsinghua University, Beijing, China

[2]Department of Computer Science, University of Maryland, College Park, USA

Content

- Background
- Motivation
- Current solution
- The key idea of Y^2Seq2Seq
- Problem statement
- Technical details
- Results
- Contributions

Background

- With the development of 3D modeling and scanning techniques, more and more 3D shapes become available on the Internet with detailed physical properties, such as texture, color, and material.

- With large 3D datasets, however, shape class labels are becoming too coarse of a tool to help people efficiently find what they want, and visually browsing through shape classes is cumbersome.

Motivation

- To alleviate this issue, an intuitive approach is to allow users to describe the desired 3D object using a text description.

- Jointly understanding 3D shape and text by learning a cross-modal representation, however, is still a challenge because it requires an efficient 3D shape representation that can capture highly detailed 3D shape structures.

Current solution

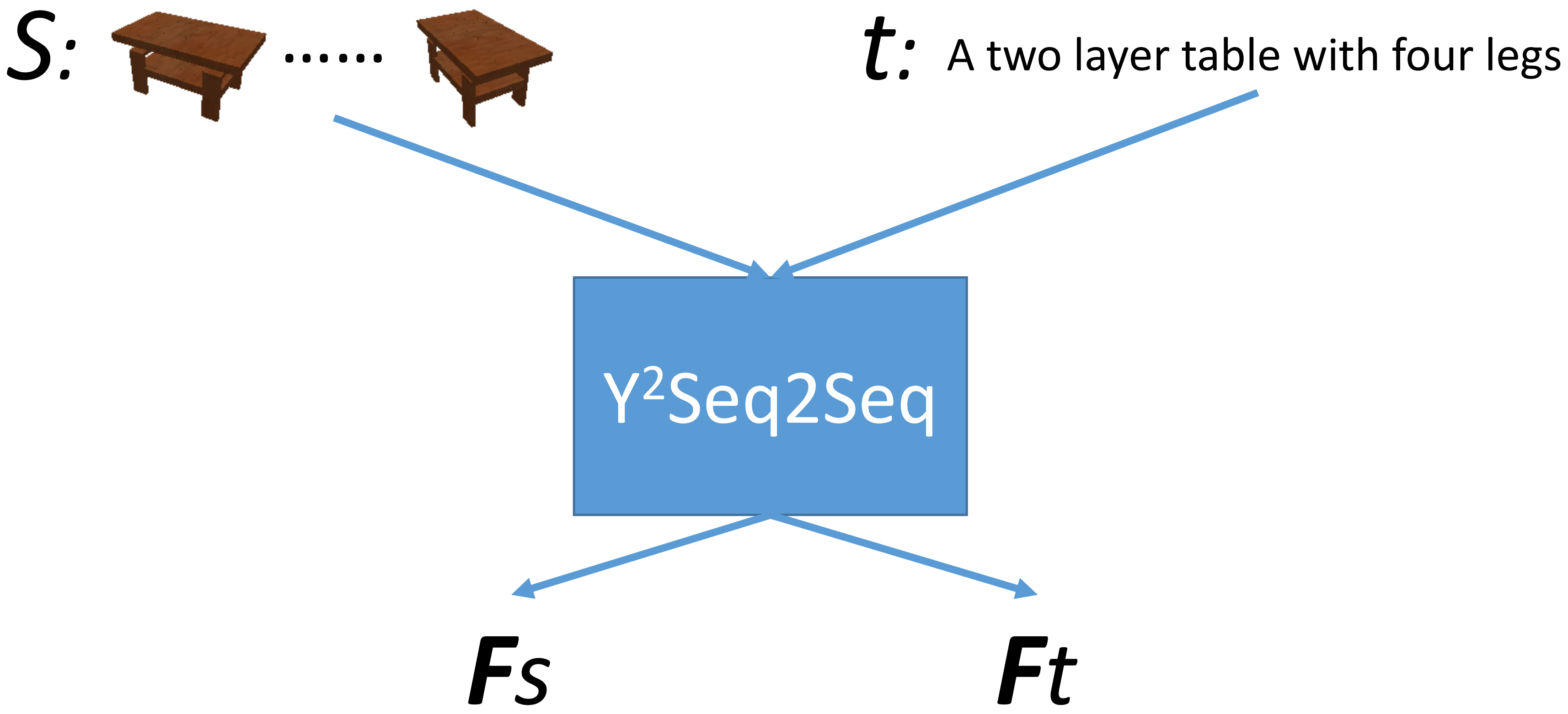
- A 3D-Text cross-modal dataset was recently released, where a combined multimodal association model was also proposed to capture the many-to-many relations between 3D voxels and text descriptions.
- However, this strategy is limited to learning from low resolution voxel representations due to the computational cost caused by the cubic complexity of 3D voxels.
- This leads to low discriminability of learned cross-modal representations due to a lack of detailed geometry information.

The key idea of Y^2Seq2Seq

- We resolve this issue by proposing to learn cross-modal representations of 3D shape and text from
  - *View sequences*, where each 3D shape is represented by a view sequence.
  - *Word sequences*, where each sentence is represented by a word sequence.
- Our deep learning model captures the correlation between 3D shape and text by simultaneously
  - *Reconstructing* each modality itself
  - And, *predicting* one modality from the other.

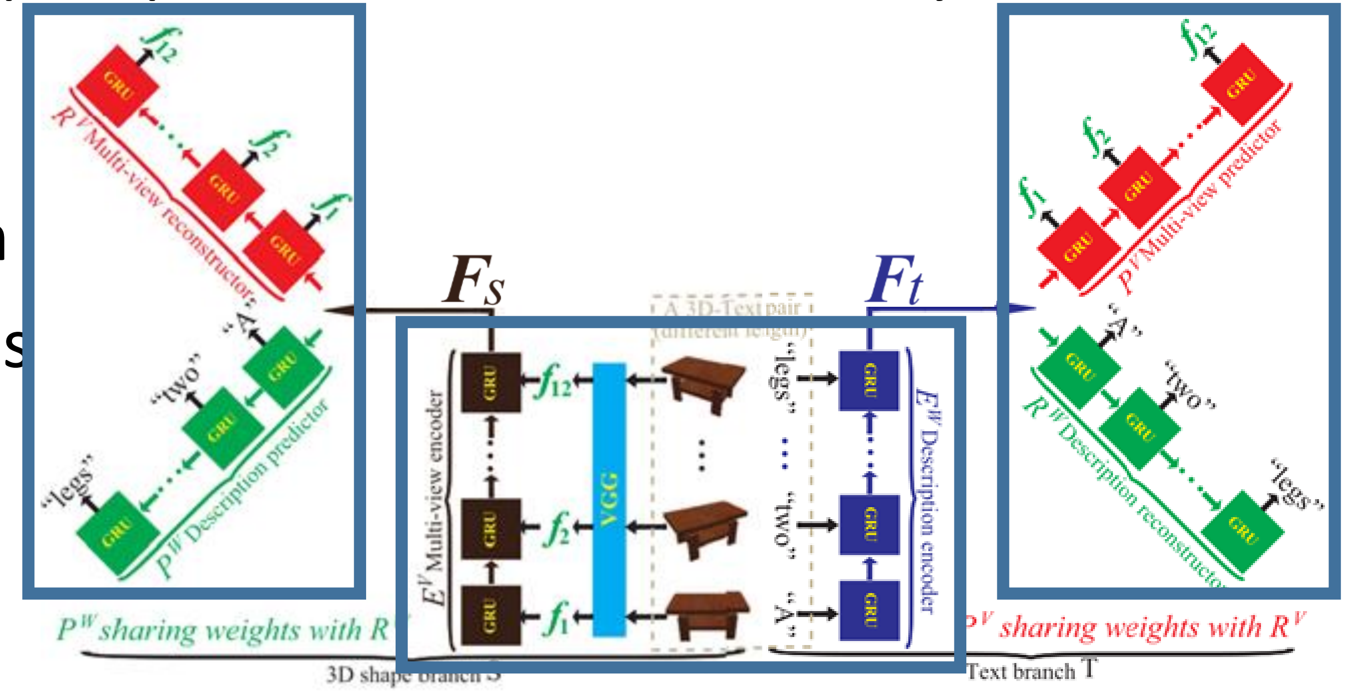
Problem statement

- Jointly earn the feature of a 3D shape *s* and the feature of a sentence *t* describing *s*.



Technical details

- Y²Seq2Seq is formed by two branches:
  - A 3D shape branch S
  - A text branch T
- Each branch is a "Y" like seq2seq model which is formed by:
  - One RNN encoder
  - Two RNN decoders
- The encoder output in each branch is the learned features of 3D shape and text.

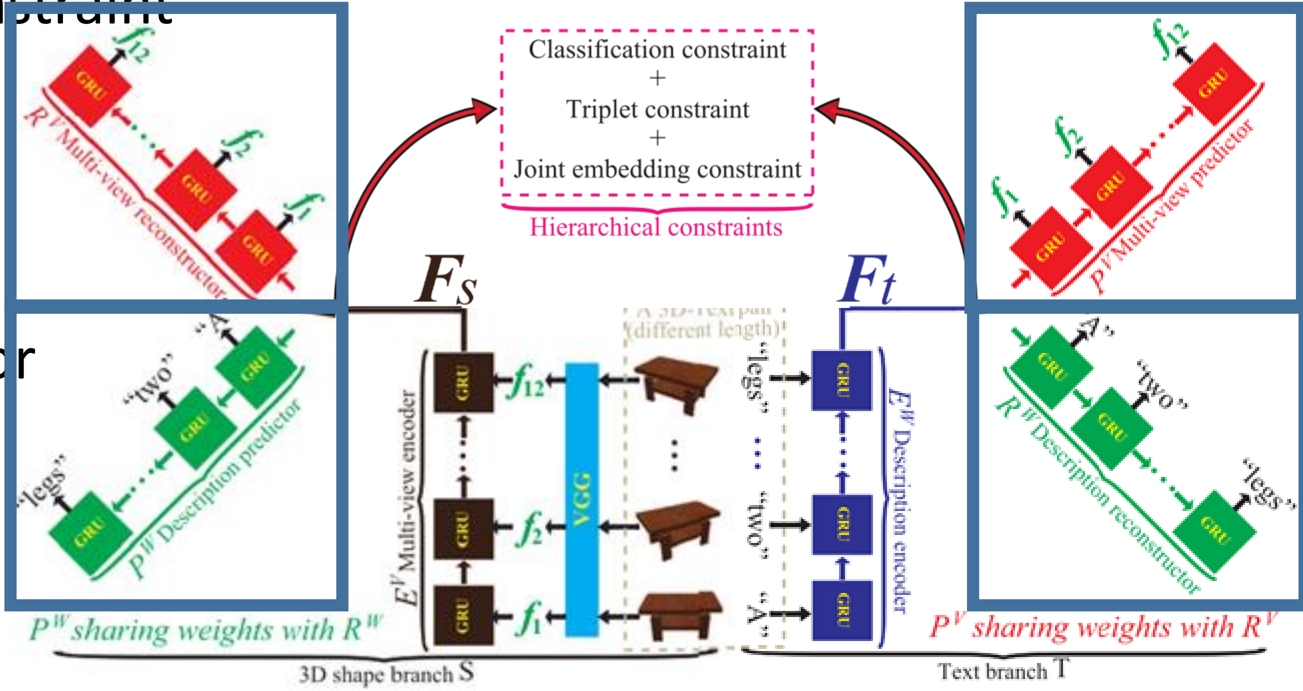


Technical details

- Hierarchical constraints are further proposed to increase the discriminability of learned features.
  - Class level: Classification constraint
  - Instance pair level: Triplet constraint
  - Instance level: Joint embedding constraint
- Two “Y” like Seq2Seq are coupled
  - 3D reconstructor and 3D predictor

  are sharing parameters.
  - Text reconstructor and text predictor

  are sharing parameters.



Technical details

- 3D shape branch S
  - 3D to 3D reconstruction

$$L_{V2V} = \frac{1}{N} \sum_{i \in [1,N]} \|\boldsymbol{f}_i' - \boldsymbol{f}_i\|_2^2$$

In feature space

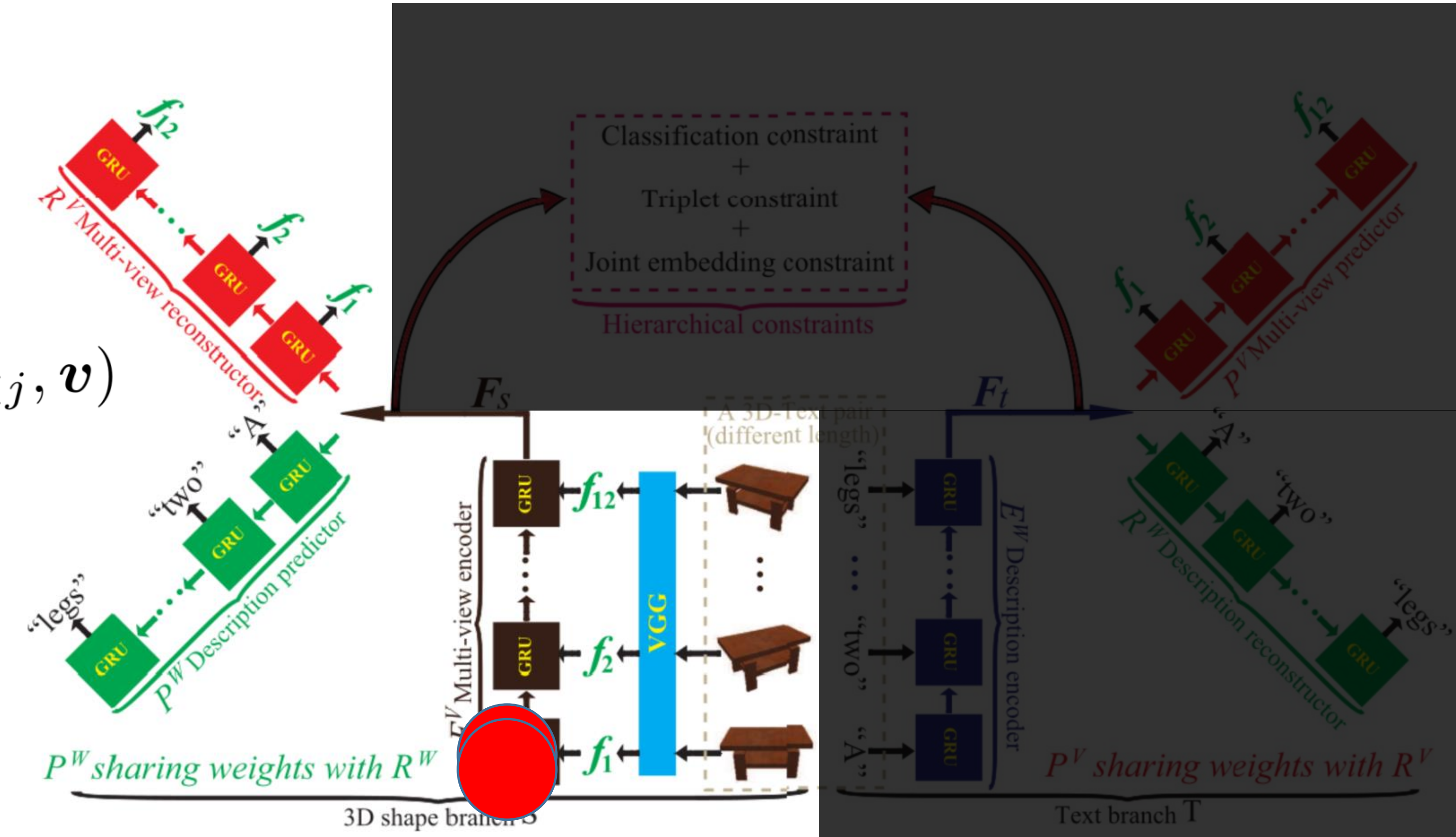
  - 3D to text prediction

$$L_{V2W} = -\sum_{j \in [1,M]} \log p(w_j | w_{<j}, \boldsymbol{v})$$

  - Total losses

$$L_S = \alpha L_{V2V} + \beta L_{V2W}$$

α and β control the balance between the two losses.



Technical details

- Text branch T
  - Text to text reconstruction

$$L_{W2W} = -\sum_{j\in[1,M]} \log p(w_j | w_{<j}, \boldsymbol{w})$$

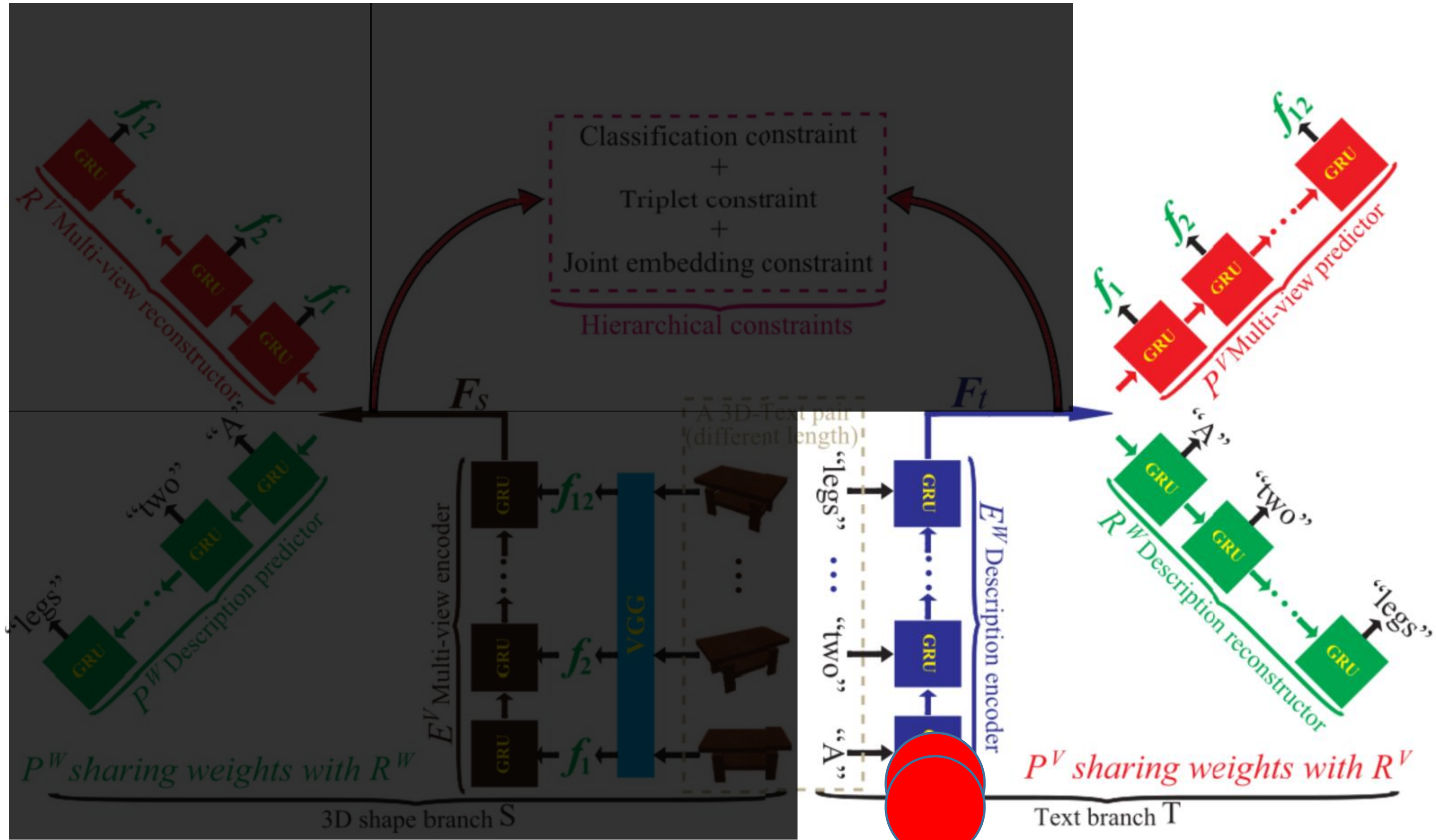
  - Text to 3D prediction

$$L_{W2V} = \frac{1}{N}\sum_{i\in[1,N]} \|\boldsymbol{f}_i'' - \boldsymbol{f}_i\|_2^2$$

  - Total losses

$$L_{\mathrm{T}} = \gamma L_{W2W} + \delta L_{W2V}$$

ϒ and δ control the balance between the two losses.



Technical details

- Hierarchical constraints
  - Class level

$$L_{\mathrm{C}1} = -\log p(c' = c|\boldsymbol{F}_s) - \log p(c' = c|\boldsymbol{F}_t)$$

  - Instance pair level

$$L_{\mathrm{C}2} = [\|\boldsymbol{F}_{s+} - \boldsymbol{F}_{t+}\|_2^2 + \|\boldsymbol{F}_{s+} - \boldsymbol{F}_{t-}\|_2^2 + \mu]_+ + [\|\boldsymbol{F}_{t+} - \boldsymbol{F}_{s+}\|_2^2 + \|\boldsymbol{F}_{t+} - \boldsymbol{F}_{s-}\|_2^2 + \mu]_+,$$

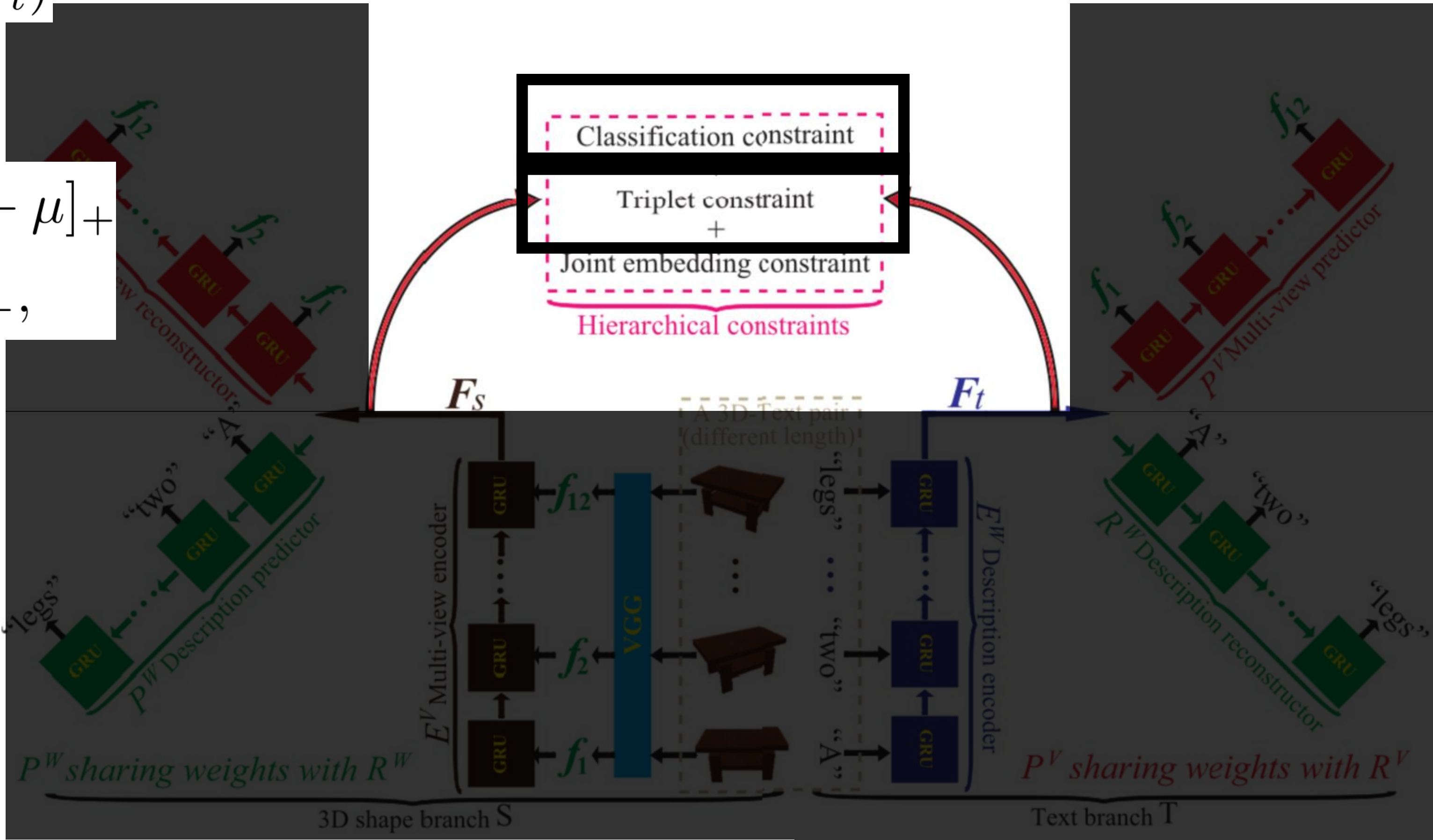
  - Instance level

$$L_{\mathrm{C}3} = \|\boldsymbol{F}_s - \boldsymbol{F}_t\|_2^2$$

  - Total losses

$$L_{\mathrm{C}} = \phi L_{\mathrm{C}1} + \varphi L_{\mathrm{C}2} + \psi L_{\mathrm{C}3}$$

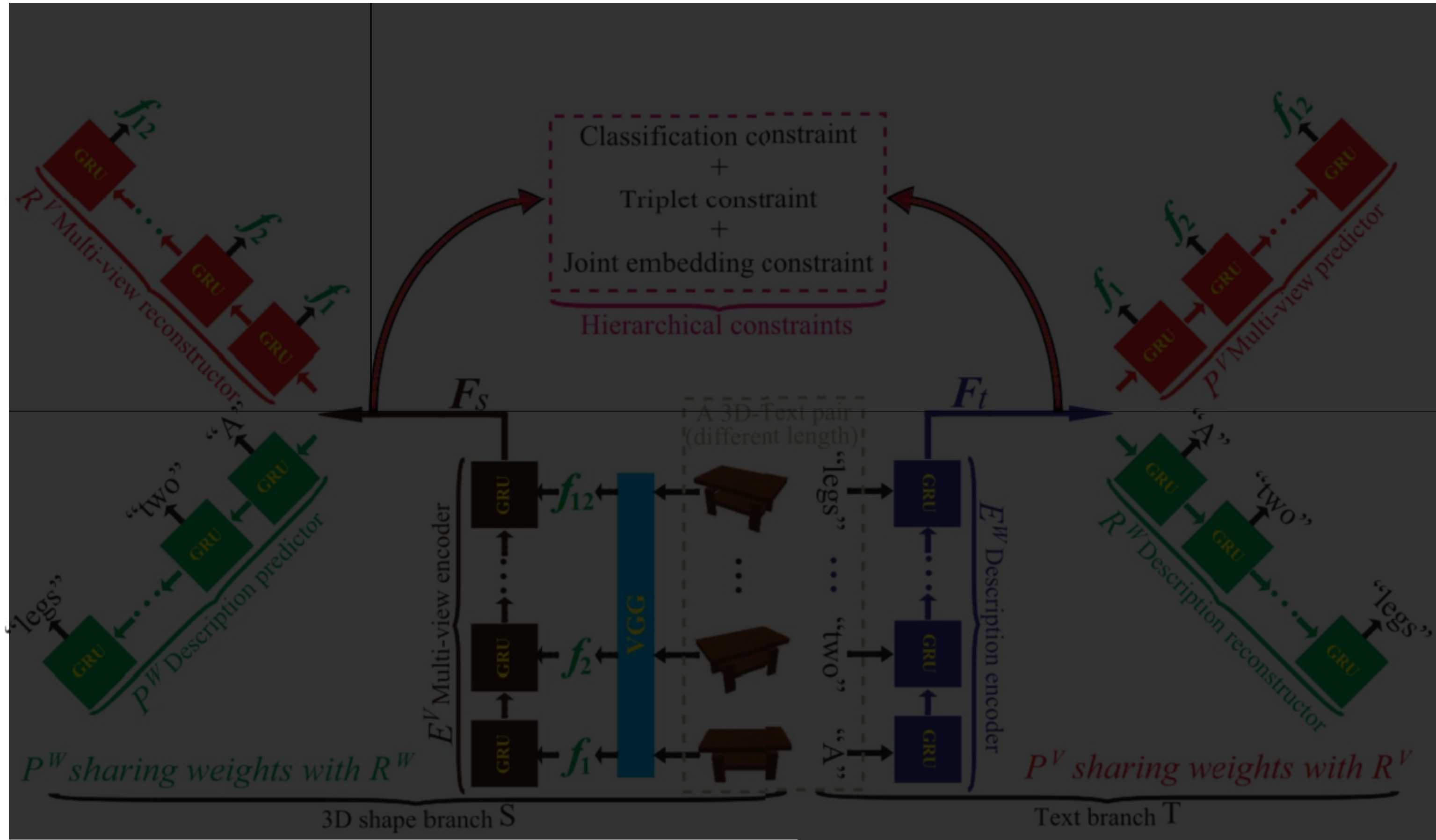
Φ, φ and ψ control the balance between the two losses.



Technical details

- Objective function

$$\min L_{\mathrm{S}} + L_{\mathrm{T}} + L_{\mathrm{C}}$$



Results

- Experimental evaluation in
  - Cross-modal retrieval, from 3D to text and from text to 3D.
  - 3D shape captioning.
- 3D-Text cross-modal dataset
  - Primitive subset, 7560 shapes and 191850 descriptions (Primitive class).
  - ShapeNet subset, 15038 shapes and 75344 descriptions (Chair class and Table class).

Results

- Effect of coupled “Y” under primitive subset

Table 2: Effect of coupled “Y” like Seq2Seq under primitive subset.

	Metrics	Rec	Pre	R+P	C-S	C-T	C-Y
S2T	RR@1	1.47	1.33	1.33	69.87	73.73	**80.13**
	RR@5	1.73	2.67	4.27	70.40	81.60	**82.53**
	NDCG@5	1.44	1.37	1.05	69.87	67.92	**80.16**
T2S	RR@1	2.27	2.06	1.85	46.92	72.57	**92.45**
	RR@5	4.70	7.58	3.34	63.37	87.73	**95.99**
	NDCG@5	1.76	3.01	0.97	41.79	70.21	**88.52**

Results

- Effect of coupled “Y” under ShapeNet subset

Table 3: Effect of coupled “Y” like Seq2Seq under ShapeNet subset.

	Metrics	Rec	Pre	R+P	C-S	C-T	C-Y
S2T	RR@1	0.07	0.07	0.13	1.61	1.74	**1.88**
	RR@5	0.34	0.34	0.34	6.03	6.17	**7.51**
	NDCG@5	0.07	0.07	0.08	1.44	1.42	**1.65**
T2S	RR@1	0.13	0.11	0.07	0.42	0.77	**1.04**
	RR@5	0.34	0.32	0.35	1.20	3.26	**4.25**
	NDCG@5	0.24	0.21	0.20	0.82	1.98	**2.62**

Results

- Effect of hierarchical constraints under primitive subset

Table 4: Effect of hierarchical constraints under primitive subset.

	Metrics	No	$+L_{C1}$	$+L_{C1}+L_{C2}$	$+L_{C}$
S2T	RR@1	80.13	83.07	88.53	**94.13**
	RR@5	82.53	85.73	88.80	**94.13**
	NDCG@5	80.16	82.43	88.33	**94.10**
T2S	RR@1	92.45	93.20	95.99	**96.66**
	RR@5	95.99	97.50	97.53	**97.57**
	NDCG@5	88.52	89.36	95.52	**95.87**

Results

- Effect of hierarchical constraints under ShapeNet subset

Table 5: Effect of hierarchical constraints under ShapeNet subset.

	Metrics	No	$+L_{C1}$	$+L_{C1}+L_{C2}$	$+L_{C}$
S2T	RR@1	1.88	2.82	3.42	**6.77**
	RR@5	7.51	10.19	10.59	**19.30**
	NDCG@5	1.65	2.40	2.59	**5.30**
T2S	RR@1	1.04	1.83	1.92	**2.93**
	RR@5	4.25	6.36	6.89	**9.23**
	NDCG@5	2.62	4.07	4.40	**6.05**

Results

- Effect of voxel resolution under ShapeNet subset

Table 6: Effect of voxel resolution under ShapeNet subset.

	Metrics	32^3	64^3	128^3
S2T	RR@1	6.77	7.31	**7.64**
	RR@5	19.30	19.97	**20.64**
	NDCG@5	5.30	5.43	**5.48**
T2S	RR@1	**2.93**	2.37	2.70
	RR@5	9.23	8.81	**9.82**
	NDCG@5	6.05	5.61	**6.27**

Results

- Cross-modal retrieval under primitive subset

Table 7: The comparison in cross-modal retrieval under primitive subset.

	Methods	RR@1	RR@5	NDCG@5
S2T	ML	24.67	29.87	24.38
	DS	80.50	85.87	80.36
	MiViSE	17.87	24.13	16.44
	SLR	1.20	2.80	1.15
	LBAT	5.20	6.13	5.25
	LBAM	89.20	90.53	89.48
	FTST	92.00	92.40	91.98
	FMM	93.47	93.47	93.47
	Our	**94.13**	**94.13**	**94.10**
T2S	ML	25.93	57.24	25.00
	DS	81.77	90.70	81.29
	MiViSE	8.21	15.42	6.84
	SLR	4.08	9.49	2.31
	LBAT	5.06	15.29	5.92
	LBAM	91.13	98.27	91.90
	FTST	94.24	97.55	95.20
	FMM	95.07	**99.08**	95.51
	Our	**96.66**	97.57	**95.87**

Results

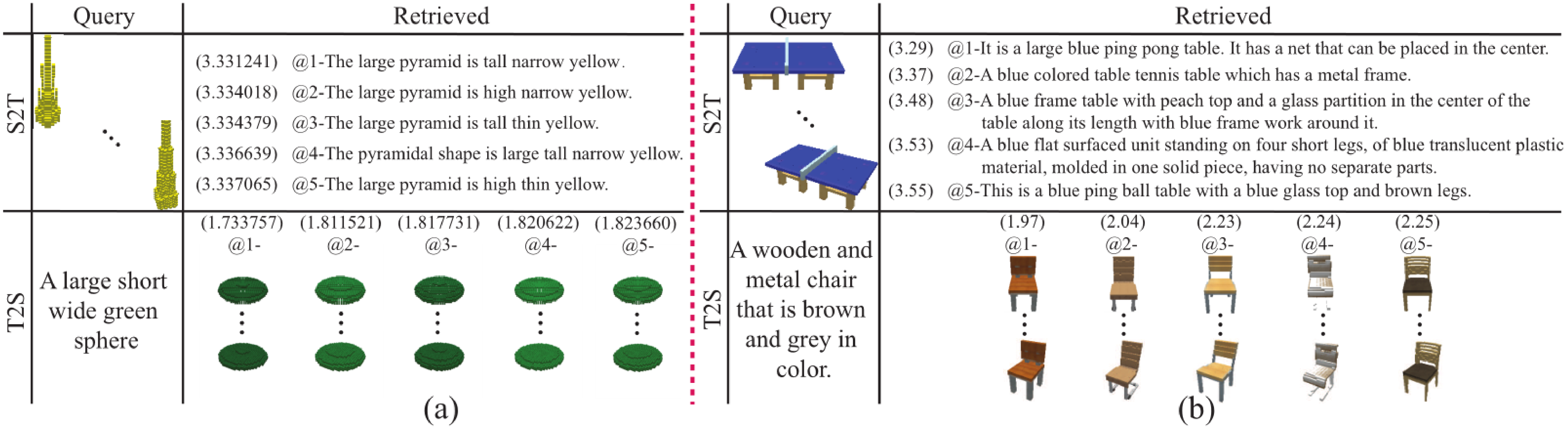
- Cross-modal retrieval under ShapeNet subset

Table 8: The comparison in cross-modal retrieval under ShapeNet subset.

	Methods	RR@1	RR@5	NDCG@5
S2T	ML	0.13	0.47	0.11
	DS	0.13	0.60	0.13
	MiViSE	0.20	0.40	0.10
	SLR	0.27	0.40	0.11
	LBAT	0.20	0.80	0.12
	LBAM	0.07	0.34	0.07
	FTST	0.94	3.69	0.85
	FMM	0.83	3.37	0.73
	Our	**6.77**	**19.30**	**5.30**
T2S	ML	0.13	0.61	0.36
	DS	0.12	0.65	0.38
	MiViSE	0.11	0.31	0.20
	SLR	0.11	0.38	0.24
	LBAT	0.04	0.20	0.12
	LBAM	0.08	0.34	0.21
	FTST	0.22	1.63	0.87
	FMM	0.40	2.37	1.35
	Our	**2.93**	**9.23**	**6.05**

Results

- Cross-modal retrieval visualization



Results

- 3D shape captioning under primitive subset

Table 9: The comparison in 3D shape captioning under primitive subset.

Model	M	R	C	B-1	B-2	B-3	B-4
SLR-N	0.18	0.44	0.13	0.42	0.31	0.21	0.15
MiV-N	0.35	0.67	0.53	0.66	0.53	0.45	0.39
S2VT	0.47	0.87	0.96	0.88	0.82	0.75	0.70
Our-N	**0.70**	**0.98**	**1.37**	**0.98**	**0.97**	**0.96**	**0.96**
Our	0.54	0.92	1.21	0.92	0.88	0.84	0.80

Results

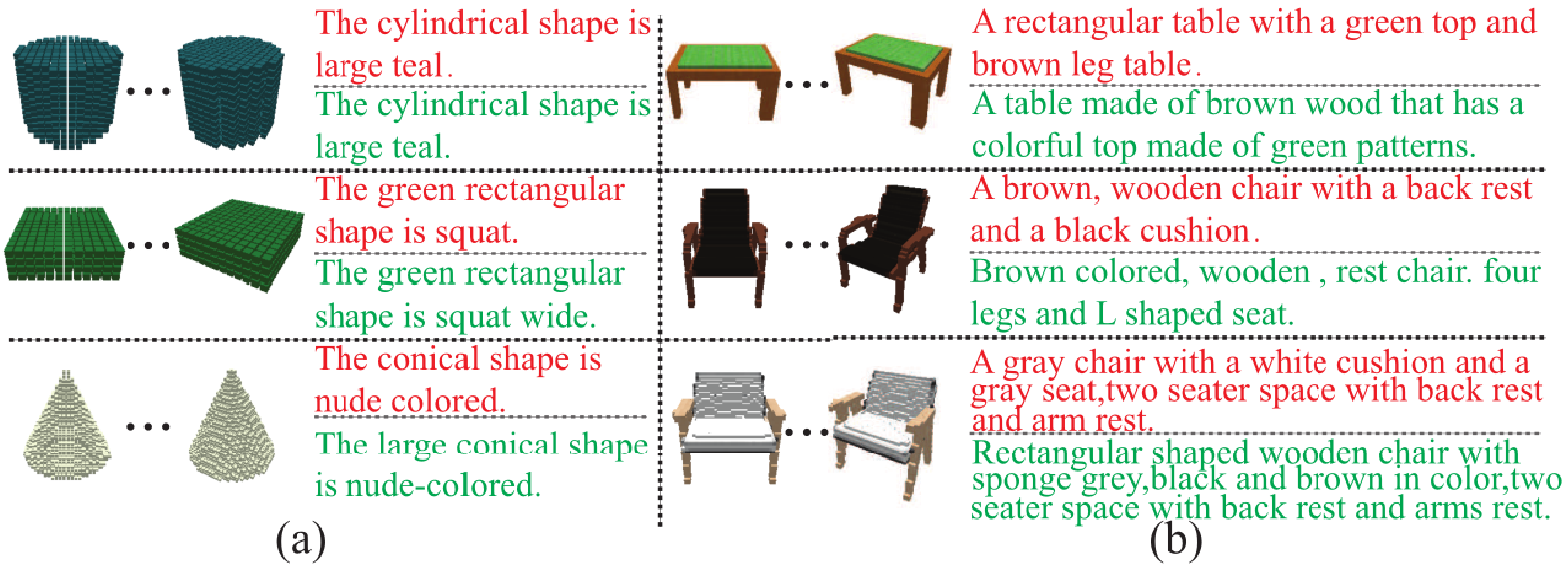
- 3D shape captioning under ShapeNet subset

Table 10: The comparison in 3D shape captioning under ShapeNet subset.

Model	M	R	C	B-1	B-2	B-3	B-4
SLR-N	0.11	0.24	0.05	0.40	0.17	0.08	0.04
MiV-N	0.16	0.36	0.14	0.61	0.35	0.21	0.12
S2VT	0.21	0.45	0.27	0.67	0.43	0.26	0.15
Our1-N	0.22	0.41	0.29	0.57	0.34	0.22	0.17
Our1	**0.29**	**0.56**	**0.71**	**0.80**	**0.65**	**0.53**	**0.46**
Our2-N	0.22	0.41	0.30	0.57	0.34	0.23	0.18
Our2	**0.30**	**0.56**	**0.72**	**0.80**	**0.65**	**0.54**	**0.46**
Our3-N	0.22	0.41	0.31	0.58	0.35	0.24	0.19
Our3	**0.29**	**0.55**	**0.70**	**0.80**	**0.64**	**0.52**	**0.44**

Results

- 3D shape captioning visualization



Contributions

- We propose a deep learning model called Y2Seq2Seq, which enables to learn cross-modal representations of 3D shape and text from view sequences and word sequences.
- Our novel coupled “Y” like Seq2Seq structures have a powerful capability to bridge the semantic meaning of two sequence-represented modalities by joint reconstruction and prediction.
- Our results demonstrate that our novel hierarchical constraints can further increase the discriminability of learned cross-modal representations by employing more detailed discriminative information.

Thank you!