Text2Shape

Presented by Akilesh B

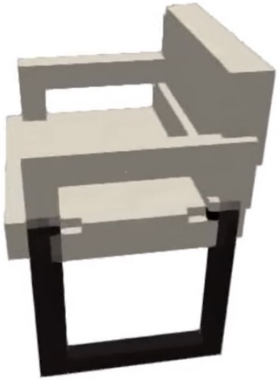
Introduction

- Goal is to connect 3D shapes with natural language descriptions.

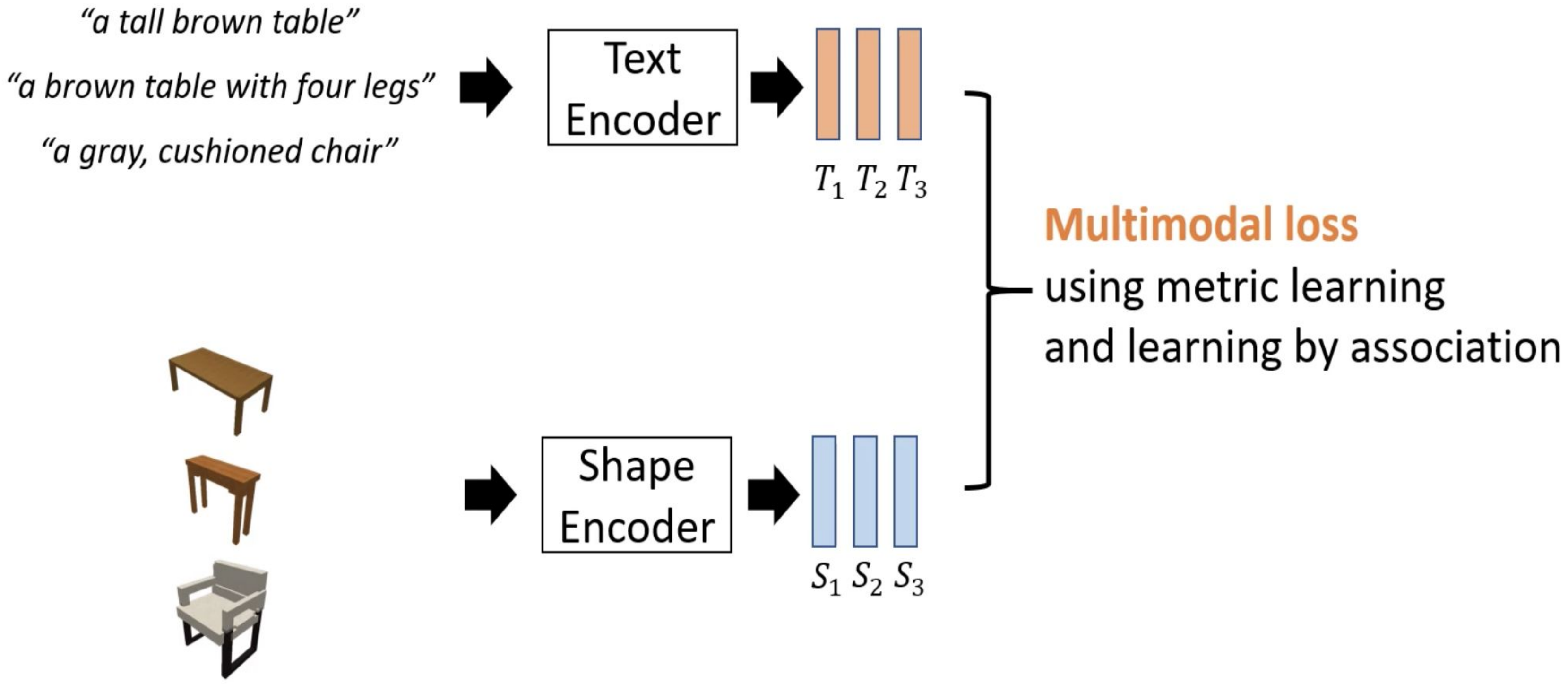
“a brown table with four legs”

“a tall brown table”

“a gray, cushioned chair”

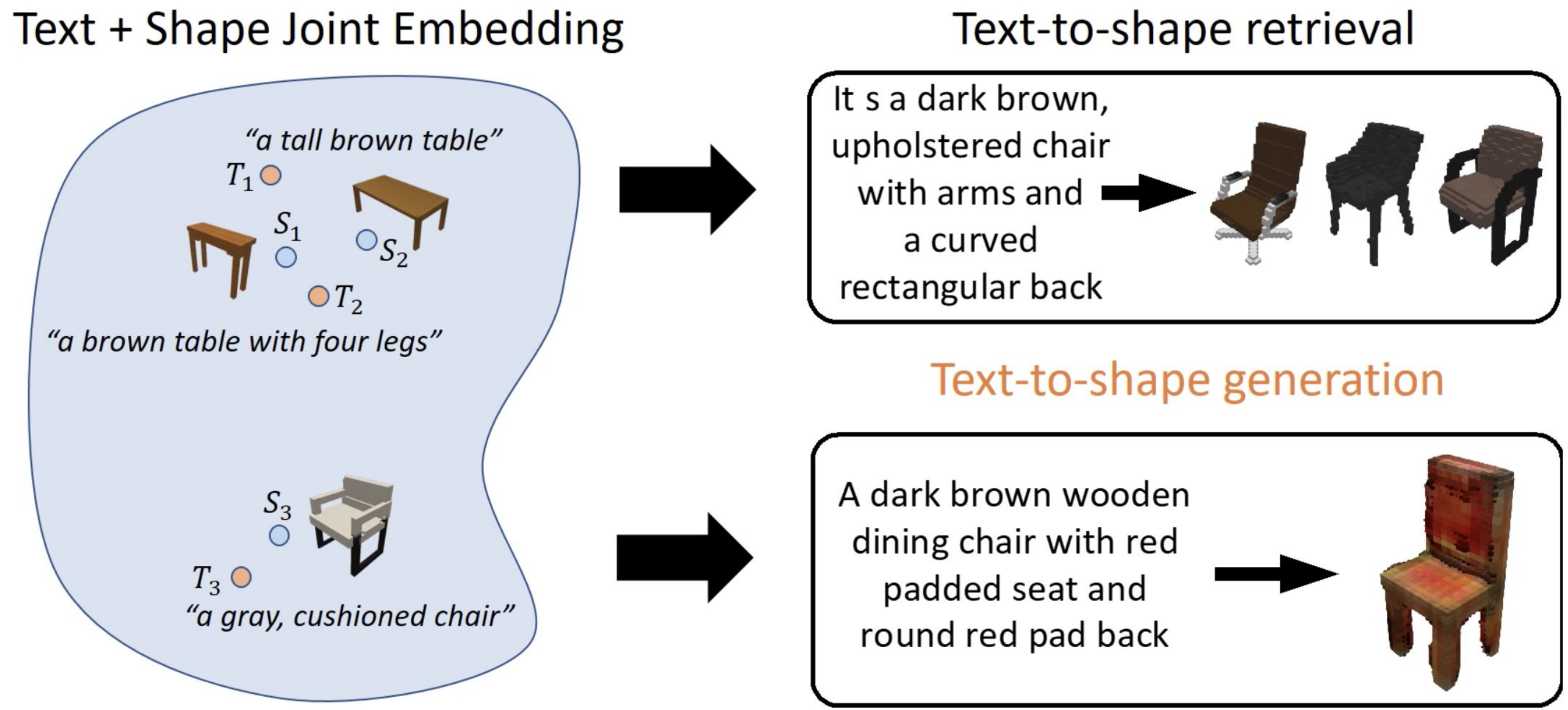


tl;dr



"a tall brown table"
"a brown table with four legs"
"a gray, cushioned chair"
Text Encoder
T_1 T_2 T_3
Shape Encoder
S_1 S_2 S_3
Multimodal loss
using metric learning
and learning by association

tl;dr



Text + Shape Joint Embedding
"a tall brown table"
T_1
S_1
S_2
T_2
"a brown table with four legs"
S_3
T_3
"a gray, cushioned chair"
Text-to-shape retrieval
It s a dark brown, upholstered chair with arms and a curved rectangular back
Text-to-shape generation
A dark brown wooden dining chair with red padded seat and round red pad back

Dataset

- Take shapes from ShapeNet (use only chairs and tables).
- Compute colored solid voxelization.
- Collect natural language descriptions from people.
- On average 5 descriptions for each chair/table.

- 15,038 ShapeNet chairs and tables
- 75,344 descriptions
- 16.3 words per description on average

Data sample

A sofa chair with four legs having cushion on its seat

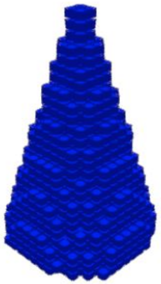
Stuffed chair covered with light and dark blue striped fabric. It has grey feet and arm rests.

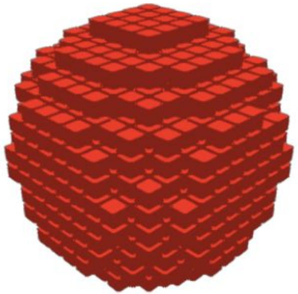
Armchair

A cushionded chair with four legs, curved arms and tones of blue stipping, pillow top seat

cushion sofa like chair, blueish and ivory stripes

Procedurally generated data





1) The blue cone is large tall.
2) A large high navy cone.
3) A large blue tall conical shape.

1) A large red sphere.
2) The ball is large crimson.
3) The large spherical shape is scarlet.

USP

- Reed et. al [4] utilizes pre-training on large image datasets.
- It relies on fine-grained category-level labels for each image (e.g. bird species).
- Learning by association to establish implicit cross-modal links between similar descriptions and shape instances.

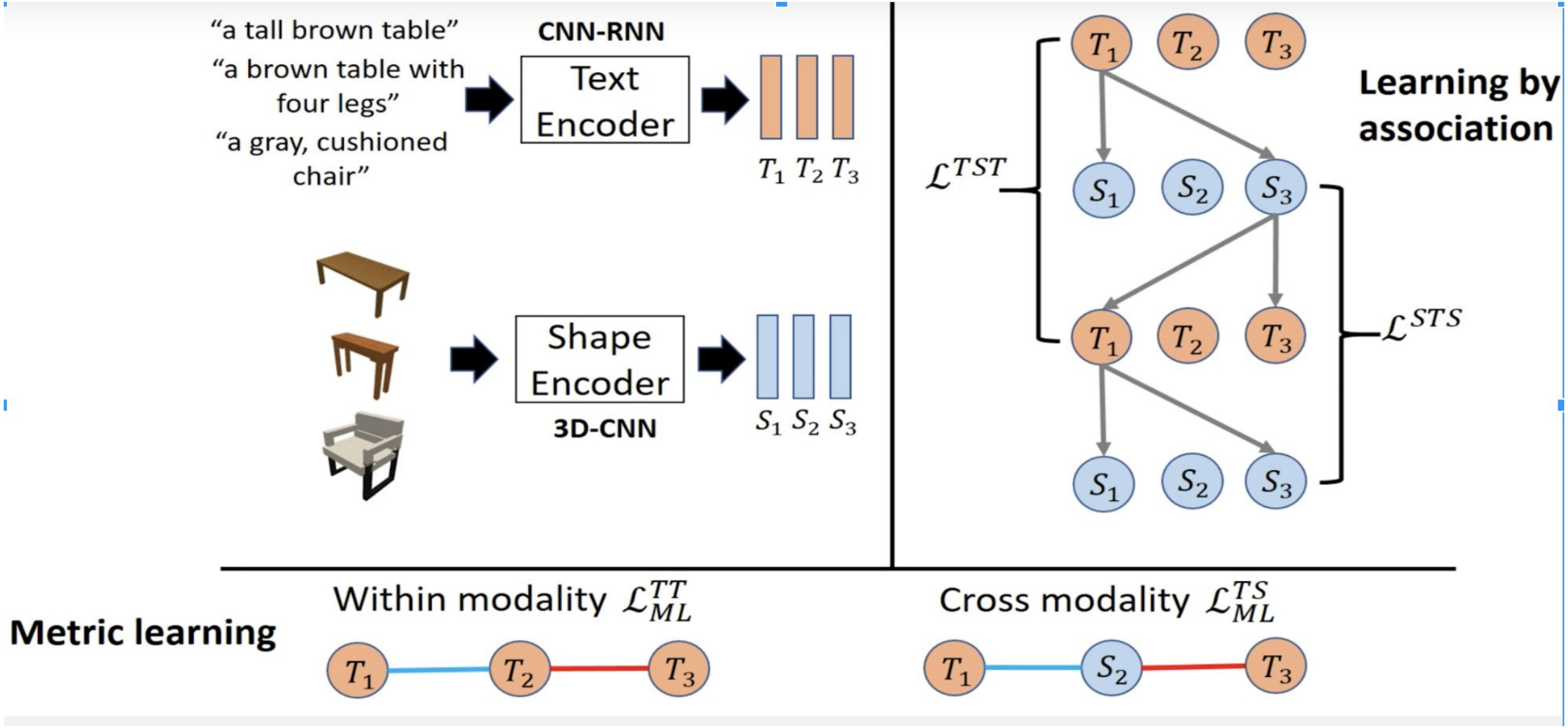
Goal

The learnt joint embedding should:

1. cluster similar text together and similar shapes together.
2. keep text descriptions close to their associated shape instance.
3. separate text from shapes that are not similar.

1) Is achieved by generalizing the learning by association approach.
2) and 3) we jointly optimize an association learning objective with metric learning.

Method



"a tall brown table"
"a brown table with four legs"
"a gray, cushioned chair"
CNN-RNN
Text Encoder
T_1 T_2 T_3
Shape Encoder
3D-CNN
S_1 S_2 S_3
Learning by association
$\mathcal{L}^{TST}$
$\mathcal{L}^{STS}$
Metric learning
Within modality $\mathcal{L}_{ML}^{TT}$
Cross modality $\mathcal{L}_{ML}^{TS}$

Method

- CNN + GRU encoder for text (produce text embeddings T).
- 3D-CNN for shape (produce shape embeddings S).
- Define text-shape similarity matrix $M_{ij} = T_i \cdot S_j$ (n shapes, m descriptions)
- $P_{ij}^{TS} = e^{Mij} / \sum_{j'} e^{Mij'}$
- Similarly, compute probability of associating shape i to description j by replacing M with M^T.
- Round-trip probability = $P_{ij}^{TST} = (P^{TS} P^{ST})_{ij}$

Association Learning

- For a given description i, goal is to have P_{ij}^{TST} be uniform over the descriptions j which are similar to description i.
- Roundtrip loss L_R^{TST} as the cross-entropy between the distribution P^{TST} and the target uniform distribution.
- To associate text descriptions with all possible matching shapes:
  - $P_j^{visit} = \sum_i P_{ij}^{TS} / m$
  - L_H^{TST} cross entropy between the P_j^{visit} and the uniform distribution over the shapes.

- $L^{TST} = L_R^{TST} + \lambda L_H^{TST}$
- In addition to TST round-trip, they impose a STS round-trip.
- L^{STS} takes similar form as above.

Multimodal metric learning

- Given: triplet (x_i, x_j, x_k) of text description embeddings.
- (x_i, x_j) belong to the same instance class (positive pair).
- (x_i, x_k) belong to different instance classes (negative pair).
- Constraint: $F(x_i;\theta) . F(x_j;\theta) > F(x_i;\theta) . F(x_k;\theta) + \alpha$
- F maps to the metric space and α is the margin.

$$\mathcal{L}_{ML}^{TT} = \frac{1}{2|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \left[\log\left(V_i + V_j\right) - m_{i,j}\right]_+^2$$

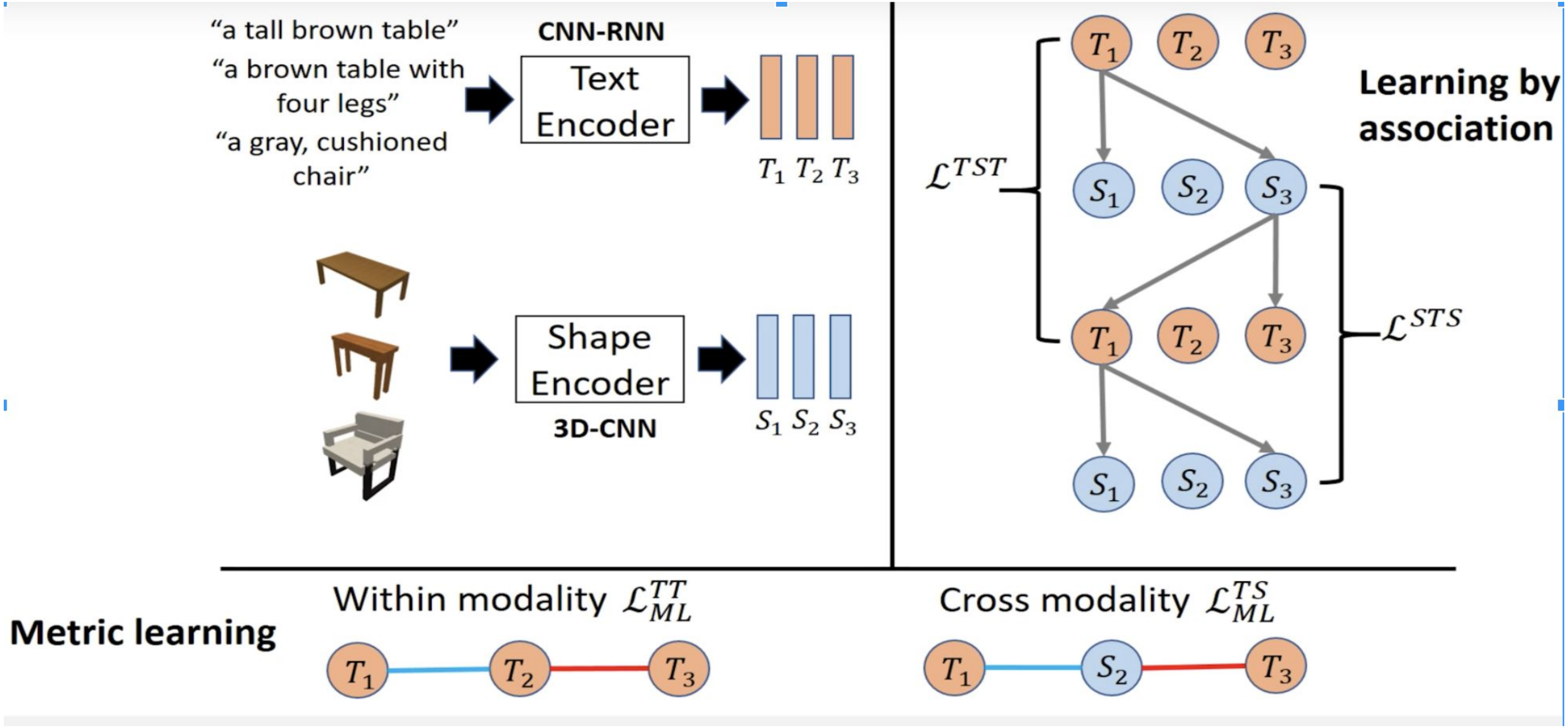
- m_{ij} denotes the similarity between x_i and x_j
- $V_l = \sum_{k \epsilon Nl} \exp\{\alpha + m_{l,k}\}$
- N_l is a negative set: set of indices that belong to an instance class other than the class l is in.
- P is a positive set : both indices i and j belong to the same instance class.

- Extend this for cross-modal similarities.
- $F(x_i;\theta) \,.\, F(y_j;\theta) > F(x_i;\theta) \,.\, F(y_k;\theta) + \alpha$
- x represents text embeddings and y represents shape embeddings.
- text-to-shape loss L_{ML}^{TS} can be derived similarly.

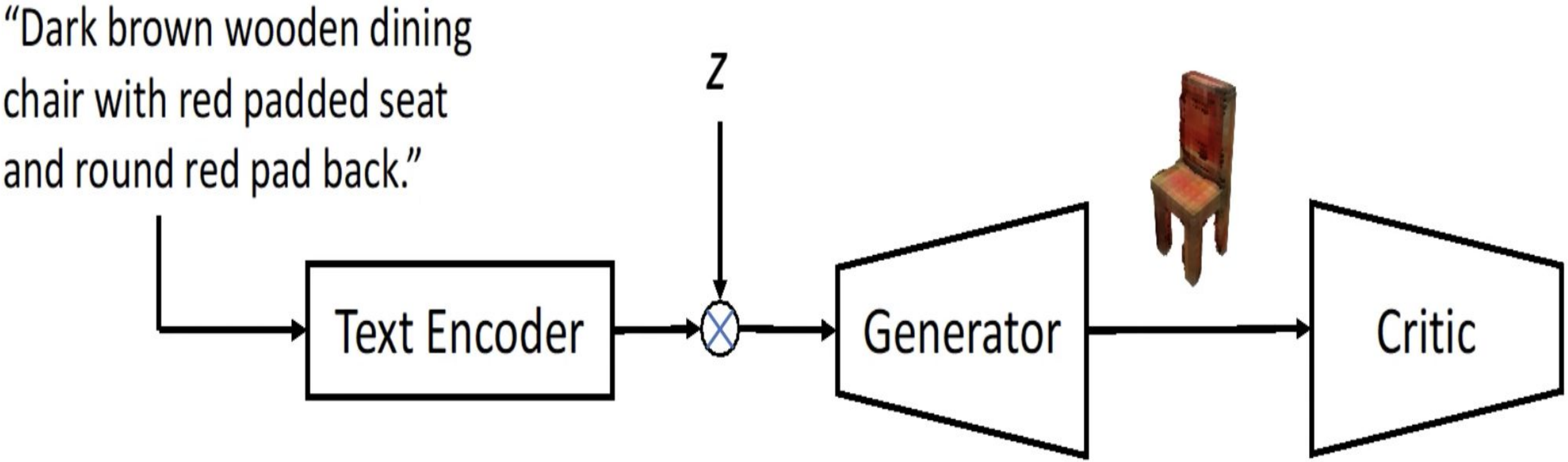
Full multimodal loss

- Combine the association losses with the metric learning losses to form the final loss function used to train the text and shape encoders.
- $L_{total} = L^{TST} + L^{STS} + \gamma\,(L_{ML}^{TT} + L_{ML}^{TS})$.

Complete picture



Generation



Novel conditional Wasserstein GAN

Objective function

$$\mathcal{L}_{\text{CWGAN}} = \mathbb{E}_{t\sim p_T}[D(t, G(t))] + \mathbb{E}_{(\tilde{t},\tilde{s})\sim p_{\text{mis}}}[D(\tilde{t}, \tilde{s})] - 2\mathbb{E}_{(\hat{t},\hat{s})\sim p_{\text{mat}}}[D(\hat{t}, \hat{s})] + \lambda_{GP}\mathcal{L}_{GP}$$

$$\mathcal{L}_{GP} = \mathbb{E}_{(\bar{t},\bar{s})\sim p_{GP}}[(\|\nabla_{\bar{t}}D(\bar{t}, \bar{s})\|_2 - 1)^2 + (\|\nabla_{\check{s}}D(\bar{t}, \bar{s})\|_2 - 1)^2]$$

p_{mat} : matching text-shape pairs.

p_{mis} : mismatching text-shape pairs.

t: text embeddings concatenated with randomly sampled noise vectors.

p_{GP} : randomly choose samples from p_{mat} or p_{mis} with 0.5 each.

Metrics

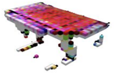
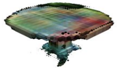
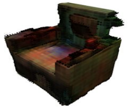
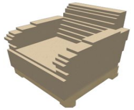
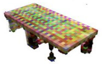
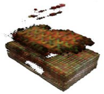
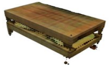
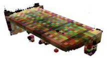
- Occupancy: IoU => Mean intersection-over-union (IoU) between generated voxels and ground truth shape voxels.
- Realism: inception score => train a chair/table shape classifier and compute the inception score.
- Color: Earth Mover's Distance => downsample voxel colors in HSV space and compute the Earth Mover's Distance between the ground truth and the generated hue/saturation distributions using L1 as the ground distance.
- Color/occupancy: classification accuracy => Accuracy of whether the generated shape class matches with the ground truth based on a shape classifier.

Quantitative results

Text-to-shape generation evaluation on ShapeNet dataset.

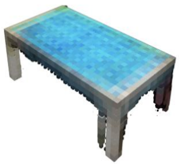
Method	IoU↑	Inception↑	EMD↓	Class Acc.↑
GAN-INT-CLS [10]	9.51	1.95	0.5039	95.57
Ours (CGAN)	6.06	1.95	0.4768	**97.48**
Ours (CWGAN)	**9.64**	**1.96**	**0.4443**	97.37

Qualitative results

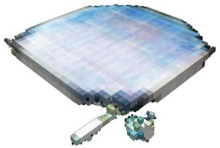
Input Text	GAN-INT-CLS [10]	Ours CGAN	Ours CWGAN	GT
Dark brown wooden dining chair with red padded seat and round red pad back.				
Circular table, I would expect to see couches surrounding this type of table.				
Waiting room chair leather legs and armrests are curved wood.				
A multi-layered end table made of cherry wood. There is a rectangular surface with curved ends, and a square storage surface underneath that is slightly smaller.				
Brown colored dining table. It has four legs made of wood.				

White coffee table

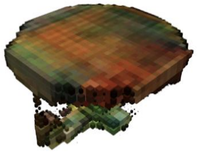
Wooden coffee table



Rectangular glass coffee table



Glass round coffee table



Red round coffee table

Red chair

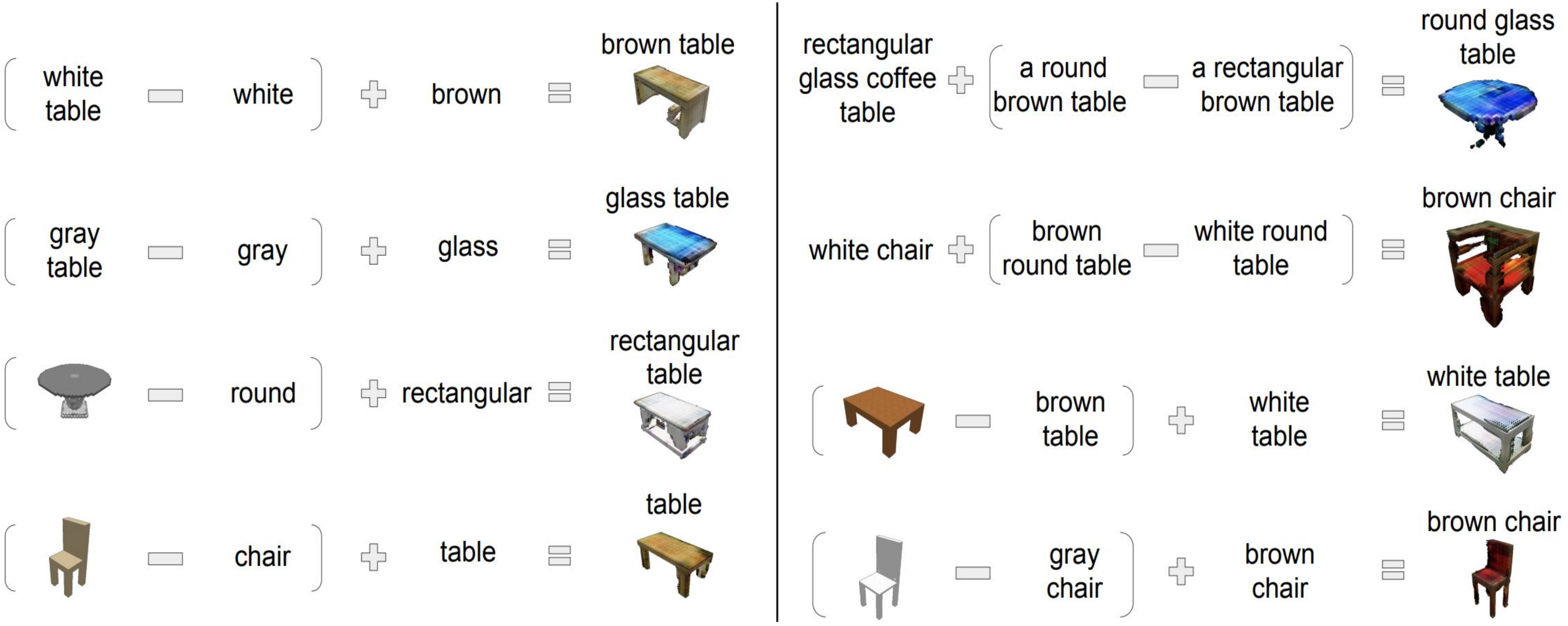
Dining chair

Gray dining chair

Silver leather chair

Gray leather chair

Shape manipulation



Summary : core contributions

- End-to-end instance-level association learning framework for cross-modal associations (text and 3D shapes).
- New problem: text to colored 3D shape generation.
- Conditional Wasserstein GAN formulation.
- Dataset of 3D shape color voxelizations and text descriptions.

References

- Learning by Association: ref1, ref2.
- Metric Learning : ref1.
- Txt2Image (Reed) : ref1.