Parts2Words: Learning Joint Embedding of Point Clouds and Texts by Bidirectional Matching between Parts and Words

Chuan Tang[1] Xi Yang[1,4*] Bojian Wu[2] Zhizhong Han[3] Yi Chang[1,4*]

[1]School of Artificial Intelligence, Jilin University, China [2]Zhejiang University
[3]The Department of Computer Science, Wayne State University, USA
[4]Engineering Research Center of Knowledge-Driven Human-Machine Intelligence, MoE, China

Abstract

Shape-Text matching is an important task of high-level shape understanding. Current methods mainly represent a 3D shape as multiple 2D rendered views, which obviously can not be understood well due to the structural ambiguity caused by self-occlusion in the limited number of views. To resolve this issue, we directly represent 3D shapes as point clouds, and propose to learn joint embedding of point clouds and texts by bidirectional matching between parts from shapes and words from texts. Specifically, we first segment the point clouds into parts, and then leverage optimal transport method to match parts and words in an optimized feature space, where each part is represented by aggregating features of all points within it and each word is abstracted by its contextual information. We optimize the feature space in order to enlarge the similarities between the paired training samples, while simultaneously maximizing the margin between the unpaired ones. Experiments demonstrate that our method achieves a significant improvement in accuracy over the SOTAs on multi-modal retrieval tasks under the Text2Shape dataset. Codes are available at here.

1. Introduction

Interaction scenarios, such as metaverse, and computer-aided design (CAD), create a larger number of 3D shapes and text descriptions. To enable a more intelligent process of interaction, it is important to bridge the gap between 3D data and linguistic data. Recently, 3D shapes with rich geometric details have been available in large-scale 3D deep learning benchmark datasets [5, 34]. Beyond 3D shapes themselves, text descriptions can also provide additional information. However, it is still hard to jointly understand 3D shapes and texts, since representing different modalities in a common semantic space is still very challenging.

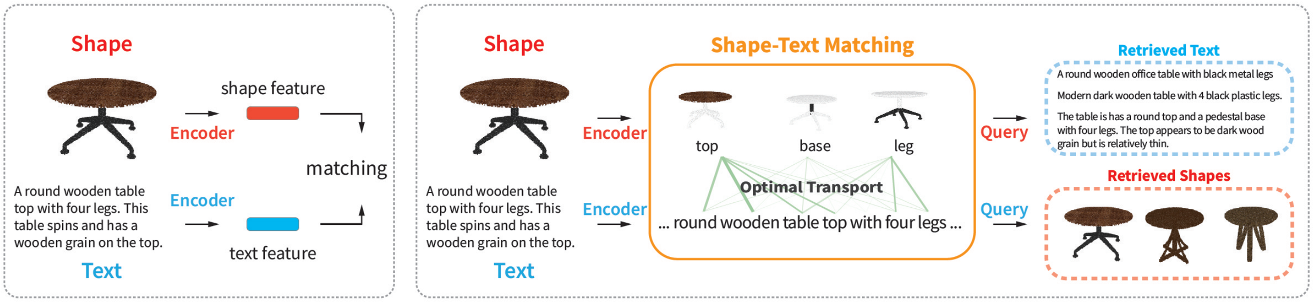
The existing methods aim at learning a joint embedded space to connect various 3D representations with texts, such as voxel grids [6] and multi-view rendered images [11, 12]. However, due to the low resolution and self-occlusions, it is hard for those methods mentioned above to improve the ability of joint understanding of shapes and texts. On the other hand, previous shape-text matching methods [6, 12, 37] usually take the global features of the entire 3D shape for text matching, making it challenging to capture the local geometries, and thus are not suitable for matching detailed geometric descriptions.

Regional-based matching approaches are commonly employed in the image-text matching task [21–23, 27], whereby visual-text alignment is established at the semantic level to enhance the performance of retrieval. These models compute the local similarities between regions and words and then aggregate the local information to obtain the global metrics between the heterogeneous pairs. However, these two-stage methods based on the pre-trained segmentation networks split the connection between matching embeddings and segmentation prior information.

In this paper, we introduce an optimal transport based shape-text matching method to achieve fine-grained alignment and retrieval of 3D shapes and texts, as shown in Figure 1. To mitigate the influence of low-resolution or self-occlusions, we directly represent the shape as point clouds and learn a part-level segmentation prior. Afterward, we leverage optimal transport to build the regional cross-modal correspondences and achieve more precise retrieval results. Our main contributions are summarized as follows:

- We propose a novel end-to-end network framework to learn the joint embedding of point clouds and texts, which enables the bidirectional matching between parts from point clouds and words from texts.
- We leverage optimal transport theory to obtain the best matches between parts and words and incorporate Earth Mover's Distance (EMD) to describe the matching score.
- To the best of our knowledge, our proposed network

[*] Corresponding authors.


Shape
shape feature
Encoder
matching
A round wooden table top with four legs. This table spins and has a wooden grain on the top.
Encoder
text feature
Text
Shape
Encoder
Shape-Text Matching
top
base
leg
Optimal Transport
... round wooden table top with four legs ...
A round wooden table top with four legs. This table spins and has a wooden grain on the top.
Encoder
Text
Query
Query
Retrieved Text
A round wooden office table with black metal legs
Modern dark wooden table with 4 black plastic legs.
The table is has a round top and a pedestal base with four legs. The top appears to be dark wood grain but is relatively thin.
Retrieved Shapes

(a) global-based shape-text matching

(b) our proposed method

Figure 1. Comparison between the global-based matching method and our proposed method. The proposed end-to-end framework aims to learn the joint embedding of point clouds and text by matching parts to words. It can either retrieve shapes using text or vice versa. Our novelty lies in the way of jointly learning embeddings of point clouds and texts.

achieves SOTA results in joint 3D shape/text understanding tasks in terms of various evaluation metrics.

2. Related Work

2.1. Joint embedding of 3D shapes and text

In recent pioneering work, Chen *et al.* [6] introduce a novel 3D-Text cross-modal dataset by annotating each 3D shape from ShapeNet [5] with natural language descriptions. In order to understand the inherent connections between text and 3D shapes, they employ CNN+RNN and 3D-CNN to extract features from text and 3D voxelized shapes respectively. They use a full multi-modal loss to learn the joint embedding and calculate the similarity between both modalities. Han *et al.* [12] propose Y^2Seq2Seq, which is a view-based method, to learn cross-modal representations by joint reconstruction and prediction of view and word sequences. Although this method can extract texture information from multiple rendered views by CNN and acquire global shape representation by RNN, it ignores local information aggregation such as part-level features of 3D shapes, which proves to be useful for 3D-Text tasks. To take a step further, ShapeCaptioner [11] detects shape parts on 2D rendered images, but it is still struggling to fully understand 3D shapes due to the inaccurate boundaries and self-occlusion. TriCoLo [37] learn the joint embedding space from three modalities by contrastive learning.

In addition, other work is attempting to establish connections between 3D shapes and natural language in other ways. Liu *et al.* [32] design a new approach for high-fidelity text-guided 3D shape generation. Text4Point [18] implements implicit alignment between 3D and text modalities using 2D images. ShapeGlot [1] explores how fine-grained differences between the shapes of common objects are expressed in language, grounded on 2D and/or 3D object representations. They build a dataset of human utterances to develop neural language understanding (listening) and production (speaking) models. ChangeIt3D [2] addresses the task of language-assisted 3D shape edits and deformations, which involves modifying or deforming a 3D shape with the assistance of natural language descriptions. VLGrammar [14] employs compound probabilistic context-free grammars to induce grammars for both image and language within a joint learning framework. PartGlot [26] learns the semantic part segmentation of 3D shape geometry, exclusively relying on part referential language.

2.2. Point-based 3D deep learning

Point clouds have been important representations of 3D shapes due to their simplicity and compactness. PointNet [35] and PointNet++ [36] are the pioneer works to understand this kind of irregular data. After that, lots of studies [29, 42] are proposed to improve the interpretability of network for point clouds in different tasks, such as 3D segmentation [30, 31, 40], 3D classification [30, 31, 40], 3D reconstruction [9, 10, 13, 19] and 3D completion [15, 16, 41].

2.3. Image-text matching

The image-text matching task allows the image or text to mutually find the most relevant instance from the multi-modal database. Most existing methods can be roughly categorized into two types: global matching methods and regional matching methods.

Global matching methods. m-RNN [33] aims to extract the global representation from both images and texts and then calculate the similarity score. VSE [25] learns to map images and text to the same embedding space by optimizing a pairwise ranking loss. VSE++ [8] tries to improve the performance by exploiting the hard negative mining strategy during training.

Regional image-text matching. These methods extract image region representation from existing detectors and then take latent visual-semantic correspondence at the level of image regions and words into consideration. DeFrag [23]

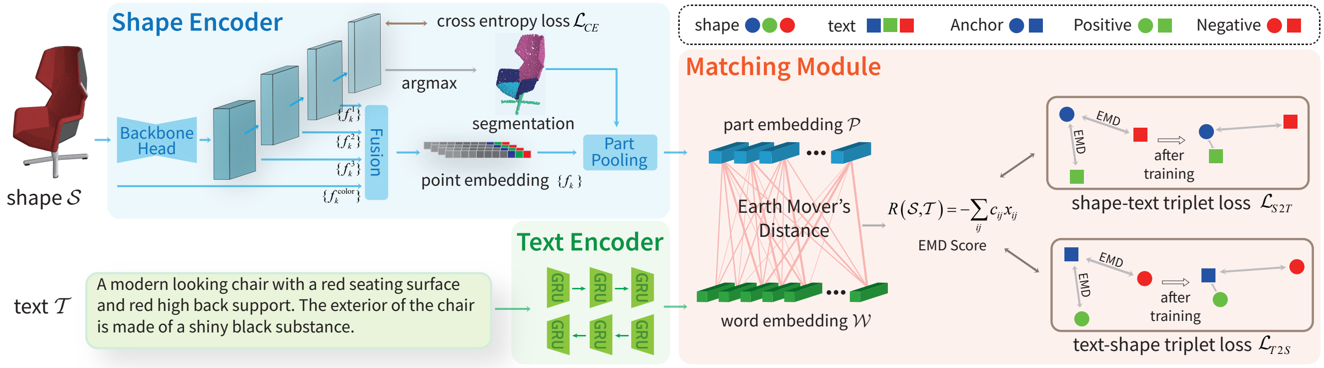

Shape Encoder
cross entropy loss $\mathcal{L}_{CE}$
argmax
segmentation
shape $\mathcal{S}$
Backbone Head
$\{f_k^1\}$
$\{f_k^2\}$
$\{f_k^3\}$
$\{f_k^{color}\}$
Fusion
point embedding $\{f_k\}$
Part Pooling
text $\mathcal{T}$
A modern looking chair with a red seating surface and red high back support. The exterior of the chair is made of a shiny black substance.
Text Encoder
GRU
shape ●●● text ■■■ Anchor ●■ Positive ●■ Negative ●■
Matching Module
part embedding $\mathcal{P}$
Earth Mover's Distance
word embedding $\mathcal{W}$
$R(\mathcal{S},\mathcal{T}) = -\sum_{ij} c_{ij} x_{ij}$
EMD Score
EMD
after training
shape-text triplet loss $\mathcal{L}_{S2T}$
text-shape triplet loss $\mathcal{L}_{T2S}$

Figure 2. Overview. The proposed network includes three modules: shape encoder, text encoder, and matching module. The shape encoder learns the part embedding from the input 3D shape, and the text encoder learns the word embedding from the corresponding text description. Then we utilize Earth Mover's Distance to measure the discrepancy between parts and words in order to obtain the best matches. To achieve the goal, we also leverage triplet loss to enhance the similarity between the paired training samples, and push the unmatched ones far apart.

and DVSA [22] propose visual semantic matching by inferring their inter-modal alignment, these methods first detect object regions and then acquire the region-word correspondence, finally aggregate the similarity of all possible pairs of image regions and words in the sentence to infer the global image-text similarity. Inspired by Up-Down [3], SCAN [27] takes a step towards attending to important image regions and words with each other as context for inferring the image-text similarity. Recently, some works SMAN [21], CASC [43], R-SCAN [28], RDAN [17], DP-RNN [7], PFAN [39] attempt to improve SCAN and try to achieve better performance.

3. Our Method

Overview As shown in Figure 2, our proposed network includes three modules: a shape encoder, a text encoder, and a matching module. To encode a 3D shape $\mathcal{S}$, we first uniformly sample points and then use a point-based backbone network that aims to predict point labels and concatenate hierarchical hidden features to acquire the representation of each point. Then, we aggregate these representations in the same part to extract its embedding $\mathcal{P} \in \{p_i | i \in [1, n]\}$ of the input shape $\mathcal{S}$ with n parts. For the text encoder, we use the Bi-directional Gate Recurrent Unit (GRU) to learn context-sensitive embedding $\mathcal{W} = \{w_j | j \in [1, m]\}$ of each word in the text $\mathcal{T}$, where the text $\mathcal{T}$ has m words.

To achieve the matching between $\mathcal{P}$ and $\mathcal{W}$, we elaborate an optimal transport-based matching module to generate the similarity score between $\mathcal{S}$ and $\mathcal{T}$. Finally, we use both segmentation loss and matching loss to train our model.

3.1. The Shape Encoder

We use 3D point cloud data as visual input and use a semantic segmentation network as the shape encoder. Our shape encoder extracts the embedding of parts with different semantic types on each input shape by aggregating the features of corresponding points in the segmented parts. We feed a shape $\mathcal{S}$ to the segmentation backbone (using PointNet [35] or PointNet++ [36]) to acquire the semantic prediction for part assignment $\{a_i | i \in [1, n]\}$ and extract the point-wise features $f_k, k \in [1, l]$, where l is the number of points in $\mathcal{S}$. We then aggregate these point-wise features into embedding of parts, which can be instantly sent into the matching module. The outputs $f_k^1, f_k^2, f_k^3, k \in [1, l]$ of the last three layers of the backbone are extracted to form the point features f_k. Besides, we also explicitly utilize the color representation $f_k^{\text{color}}, k \in [1, l]$ of the input shape for better performance, our combined feature fusion module makes full use of the color information and semantic information of 3D shape at the same time. Thus, the point-wise feature representation f_k is computed by Equation (1):

$$f_k = \text{fc}(\text{fc}(f_k^1) + \text{fc}(f_k^2) + \text{fc}(f_k^3) \oplus \text{mlp}(f_k^{\text{color}})), \quad (1)$$

where mlp stacks multiple fully connection layers fc to perform nonlinear mapping on color feature f_k^{color}. Then, pooling is applied to the point-wise features within the same semantic part, in order to extract the embedding of each part p_i, as shown in Equation (2).

$$p_i = \underset{k \in a_i}{pooling}(f_k), i = 1, \ldots, n \quad (2)$$

3.2. The Text Encoder

The text encoder aims to extract local features at the word level, as shown in Equation (3), we use Bi-directional GRU to extract the context-sensitive word embedding $\mathcal{W}$. Each text description $\mathcal{T}$ is firstly represented by the embedding of every single word e_j in the text through a word embedding layer. Then, we encode the context of w_j in the bi-directional GRU. For the forward $\overrightarrow{GRU}$, the hidden state $\overrightarrow{h_j}$ can be calculated from the embedding e_j of the current word at each time-step, and the hidden state h_{j-1} from the previous time-step. Similarly, for the backward $\overleftarrow{GRU}$, the current hidden state $\overleftarrow{h_j}$ is calculated from the embedding e_j of the current word and the hidden state h_{j+1} from the next. Finally, the context-sensitive word embedding is obtained by the averages of the hidden states in the two directions.

$$
\begin{aligned}
\overrightarrow{h_j} &= \overrightarrow{GRU}\left(e_j, h_{j-1}\right), j \in [1, m] \\
\overleftarrow{h_j} &= \overleftarrow{GRU}\left(e_j, h_{j+1}\right), j \in [1, m] \\
w_j &= \frac{\overrightarrow{h_j} + \overleftarrow{h_j}}{2}, j \in [1, m]
\end{aligned}
\tag{3}
$$

3.3. The Matching Module

We introduce the optimal transport method to evaluate the similarity of 3D shape and text using their embedding $\mathcal{P}$ and $\mathcal{W}$. Similar to the transport of goods between producers (part p_i) and consumers (word w_j), p_i contains u_i stock of goods and w_j has v_j capacity. Then, the transport cost between p_i and w_j is defined as c_{ij}, and the matching flow is defined as x_{ij}. We formulate our matching problem in Equation (4):

$$
\begin{aligned}
\min_{x_{ij}} \quad & \textstyle\sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} x_{ij} \\
\text{s.t.} \quad & x_{ij} \geq 0, i = 1, \ldots, n, j = 1, \ldots, m \\
& \textstyle\sum_{j=1}^{n} x_{ij} = u_i, \quad i = 1, \ldots n \\
& \textstyle\sum_{i=1}^{m} x_{ij} = v_j, \quad j = 1, \ldots m \\
& \\
& c_{ij} = 1 - \frac{{p_i}^T w_j}{\|p_i\| \|w_j\|}
\end{aligned}
\tag{4}
$$

where the cost c_{ij} between p_i and w_j is measured by cosine distance. We adopt a simple discrete uniform distribution for the EMD node weight settings of u_i and v_j, which constrains the upper limit of the summation of matching flow.

Then, we use the Sinkhorn algorithm to compute the optimal matching flow $\hat{x}_{ij}$ which is the solution of the earth mover's distance function in Equation (4). After optimization, we calculate the similarity R_{EMD} between shape $\mathcal{S}$ and text $\mathcal{T}$ as shown by Equation (5)

$$
R_{EMD}(\mathcal{S}, \mathcal{T}) = -\sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} \hat{x}_{ij}
\tag{5}
$$

3.4. Objective Function

As shown in Equation (6), we train the proposed network using the multi-task learning strategy, and learn the part segmentation and matching simultaneously. The weight β is employed to balance the two tasks.

Table 1. Retrieval results on Text2shape compared to the SOTAs.

Method	S2T			T2S		
	RR@1	RR@5	NDCG@5	RR@1	RR@5	NDCG@5
Text2Shape [6]	0.83	3.37	0.73	0.40	2.37	1.35
Y^{2}Seq2Seq [12]	6.77	19.30	5.30	2.93	9.23	6.05
TriCoLo [37]	16.33	45.52	12.73	10.25	29.07	19.85
Global-Max (Our)	12.60	32.96	9.48	8.53	24.09	16.46
Global-Avg (Our)	7.63	24.07	6.23	8.60	24.82	16.83
LocalBaseline (Our)	12.81	34.71	9.82	7.87	23.55	15.84
LocalBaseline+ (Our)	17.77	44.58	13.91	11.94	31.62	21.92
Parts2Words (Our)	**19.38**	**47.17**	**15.30**	**12.72**	**32.98**	**23.13**

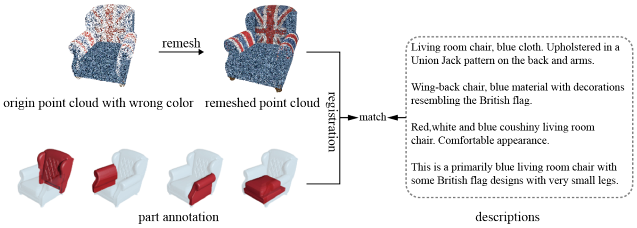

remesh
origin point cloud with wrong color
remeshed point cloud
registration
match
Living room chair, blue cloth. Upholstered in a Union Jack pattern on the back and arms.
Wing-back chair, blue material with decorations resembling the British flag.
Red,white and blue coushiny living room chair. Comfortable appearance.
This is a primarily blue living room chair with some British flag designs with very small legs.
part annotation
descriptions

Figure 3. Data pre-process. We rectify the incorrect color input provided by ShapeNet [5] and add the part segmentation annotation from PartNet [34] by registration. Finally, the new segmentation dataset will be combined with the Text2Shape [6].

The segmentation network is optimized by the cross-entropy loss L_{CE} only. While, for the matching task, we adopt the paired ranking loss with the semi-hard negative sampling mining strategy [38] to facilitate the network to better converge and avoid getting into a collapsed model, as shown in Equation (7). Specifically, for a positive sampling pair $(\mathcal{S}, \mathcal{T})$, we select the semi-hard negative sampling pair $(\hat{\mathcal{S}}_{semi}, \hat{\mathcal{T}}_{semi})$ which has a smaller similarity score than $(\mathcal{S}, \mathcal{T})$, and calculate the triplet loss for the input shape and text respectively. Similarly, the triplet loss between the sampling pair $(\mathcal{T}, \mathcal{S})$ can also be calculated in the same way.

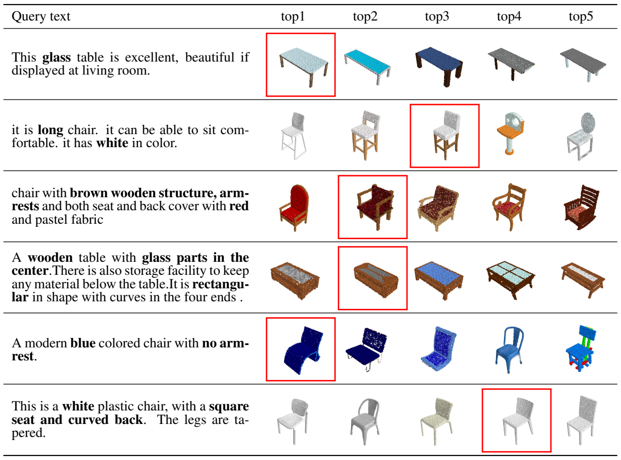
$$
L = L_{CE} + \beta L_{EMD}
\tag{6}
$$

$$
\begin{aligned}
L_{EMD} &= L_{S2T}(\mathcal{S}, \mathcal{T}) + L_{T2S}(\mathcal{T}, \mathcal{S}), \qquad (7) \\
L_{S2T}(\mathcal{S},\mathcal{T}) &= max(\alpha - R_{EMD}(\mathcal{S},\mathcal{T}) + R_{EMD}(\mathcal{S},\hat{\mathcal{T}}_{semi}), 0) \\
L_{T2S}(\mathcal{T},\mathcal{S}) &= max(\alpha - R_{EMD}(\mathcal{T},\mathcal{S}) + R_{EMD}(\mathcal{T},\hat{\mathcal{S}}_{semi}), 0)
\end{aligned}
$$

Here, α is a margin that is enforced between the positive and negative pairs.

4. Experiments

Data preparation We evaluate our proposed network on a 3D-Text cross-modal dataset [6]. However, this dataset does not include 3D point clouds and the segmentation prior. To resolve this issue, we establish our training samples using two additional datasets, ShapeNet [5] and PartNet [34], which share the same 3D models. Moreover, in contrast to the segmentation annotations offered



Query text	top1	top2	top3	top4	top5
This **glass** table is excellent, beautiful if displayed at living room.					
it is **long** chair. it can be able to sit comfortable. it has **white** in color.					
chair with **brown wooden structure, armrests** and both seat and back cover with **red** and pastel fabric					
A **wooden** table with **glass parts in the center**.There is also storage facility to keep any material below the table.It is **rectangular** in shape with curves in the four ends .					
A modern **blue** colored chair with **no armrest**.					
This is a **white** plastic chair, with a **square seat and curved back**. The legs are tapered.					

Figure 4. Retrieval result of the proposed Parts2Words (T2S). For each query sentence, we show the top-5 ranked shapes, the ground truth shapes are marked as red boxes. The words on the corresponding to the details of the retrieved shapes are marked in **bold**.

Query Shape	Retrieval Result
	1. wooden chair with white color sponge on seating area and back support. trapezium shape seating area and eclipse shape back support . 2. brown and white with round back . appear to be wood and a soft material on the seat and back . 3. a fancy dining chair with white padding and brown , wooden frame . the backing be round with wooden framing and a white pad . 4. wooden chair with four leg , gray padded seat and back rest , the back rest be tall and straight . on the top of the back rest there be a semicircular protuberance . the wooden frame have classical style . 5. a wooden chair with white upholstered seat
	1. it be a five - wheeled , black , height - adjustable office chair on coaster wheel . 2. modern irregular shape black chair make of plastic , with an iron structure and with 5 wheel on the bottom 3. this be a black , height - adjustable , office chair without arm . its 5 wheel and single base be make of silver color metal and the seat be make out of a hard , black material . 4. black chair with back support . have single leg with 5 wheel branch from the main support . 5. a black armless computer stool with leg on roller .
	1. brown color , l shape , wooden table . box drawer at both leg side . with rectangular l shape plain top . 2. brown wooden corner desk unit with drawer two set of two 3. brown business desk . have an l shape appearance 4. a brown wooden desk at a 90 degree angle and drawer on both end of the l shape 5. an l shape dark brown colored wooden table
	1. a long brown table that be oblong in shape 2. a modern oval shape wooden table with six design short leg 3. a long brown wooden rectangular table with three full cover leg 4. a brown , rounded wooden table with three leg 5. this be a short brown elongated round table . this table would be use as a coffee table with a small table in the middle of a room .
	1. a wicker round chair with blue plush seat and pillow . the chair be on a circular pedestal 2. round shape brown chair with blue cushion in it 3. a multi color round shape fashion chair with cushion 4. round c shape back chair in a checkered pattern all through the chair . circle cushion seat with a cylinder checkered single pole in the middle to support the cushion . the bottom be circle foundation in same checkered pattern . 5. one short round chair with three royal blue colored cushion and have a round backrest
	1. an oval shape table which be yellow at top with red lining . below that blue color be see . 2. an oval table with a blue shelf under it . the tabletop be yellow with a narrow red trim around it . 3. the object be a large oval table with a yellow top and red rim around the yellow . it also have a blue shelf underneath and black leg . 4. an oval shape pool table with a green felt top , wooden out layer and two moon shape wood leg 5. oval table with 4 leg and material from wood , plastic , brown wood and blue plastic help the table very luxury

Figure 5. Retrieval result of the proposed Parts2Words (S2T). For each query shape, we show the top-5 ranked sentences, the ground truth sentences are marked in red.

by ShapeNet [5], the segmentation annotations present in PartNet [34] demonstrate a level of inherent object classification supervisory capability. We sample point clouds with color from meshes in ShapeNet [5], and assign each point a label using the fine-grained, instance-level, and hierarchical 3D segmentation ground truth of the same 3D shapes in PartNet [34]. We notice that the original point clouds in ShapeNet dataset [5] have the wrong color which is inherited from its inner meshes (The origin point cloud with incorrect color is shown in Figure 3). The incorrect color input will cause problems for our model to understand color information. To mitigate this problem, we first remove the inner mesh from the origin mesh data, then we sample points with the correct color. At the same time, we leverage ICP [4] to map the 3D segmentation ground truth on shapes in PartNet to the sampled point clouds. Finally, we obtained point clouds with color and segmentation ground truth with different granularities. In the 3D-Text dataset that contains chairs and tables, we use 11498 3D shapes as training samples, and the remaining 1434 ones are considered as test data. And each 3D shape has an average of 5 text descriptions.

Evaluation metrics We employ recall rate (RR@k, $k = 1, 5$) and NDCG [20] to conduct quantitative evaluation. RR@k is defined as the percentage of correct text/shape in the top k retrieval results.

Parameter Setting We set the number of each point cloud l to 2500, and output the segmentation result and the matching scores between shapes and text. We use the coarse granularity of 17 classes as segmentation ground truth. For the shape encoder module, we tried to use PointNet and PointNet++ respectively as the backbone. In the group pooling module, we use average pooling to aggregate part embedding. Additionally, we ignore the part with less than 1% of total points. In the matching module, we set the dimension of embedding to 1024, which is consistent with [11, 12]. We also use the vocabulary of 3587 unique words and a single-layer Bi-directional GRU as the text encoder. Due to the limited vocabulary of the Text2Shape dataset used in the experiment, the impact of text pre-training methods like BERT on the results is minimal (within 1%). Therefore, we only use a text encoder trained from scratch in the comparative and ablation experiments. For the loss function, we adopt a semi-hard negative mining strategy, and the margin α of the triplet ranking loss is set to 0.2. We divide the training process into two stages. First, we pretrain the model only by semantic segmentation loss with 50 epochs and then train multi-task loss with 20 epochs and we set the balance weight of loss β to 40. Our model uses the Adam [24] algorithm as the optimizer and set the initial learning rate to 0.001. We used RTX 6000 with 24GB for training and set the batch size to 128.

4.1. Comparison results

Table 1 presents the quantitative results on 3D-Text dataset where our method outperforms the latest approaches Text2Shape [6], Y^{2}Seq2Seq [12] and TriCoLo [37] in all

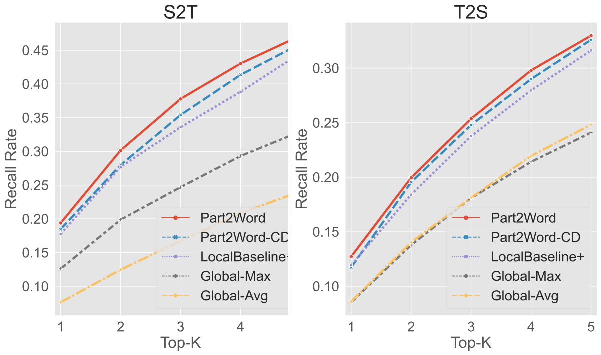


Figure 6. Comparison between global/local-based methods. The result shows the local-based methods (Parts2Words, Parts2Words-CD, LocalBaseline+) outperform the global-based methods (Global-Max, Global-Avg).

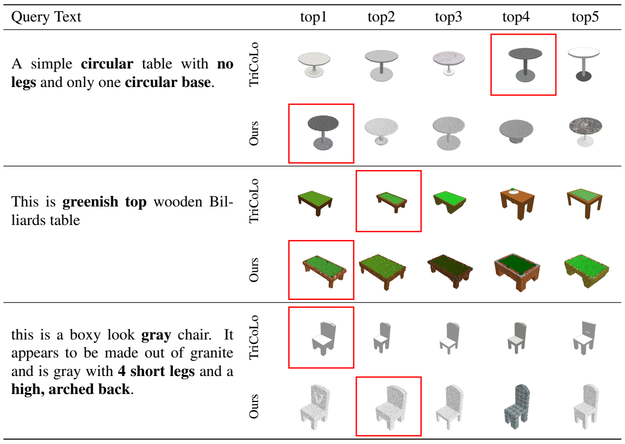

Query Text
top1
top2
top3
top4
top5
A simple circular table with no legs and only one circular base.
This is greenish top wooden Billiards table
this is a boxy look gray chair. It appears to be made out of granite and is gray with 4 short legs and a high, arched back.
TriCoLo
Ours

Figure 7. Retrieval result of the proposed Parts2Words and TriCoLo [37] (T2S). the ground truth shape are marked as red boxes. Words corresponding with the parts in retrieved shapes are marked in **bold**.

Table 2. Component-wise analysis on text2shape, with different backbones on shape encoder and different matching modules.

		S2T			T2S		
Backbones	Matching	RR@1	RR@5	NDCG@5	RR@1	RR@5	NDCG@5
PN++	EMD	16.66	43.11	13.20	11.10	28.89	20.25
PN	CD	18.47	45.98	14.52	11.73	32.60	22.49
PN	EMD	**19.38**	**47.17**	**15.30**	**12.72**	**32.98**	**23.13**

Table 3. Effectiveness of end-to-end training and semi-hard negative mining strategies.

		S2T			T2S		
w e2e	w semi	RR@1	RR@5	NDCG@5	RR@1	RR@5	NDCG@5
	✓	16.31	43.39	12.99	10.68	29.25	20.14
✓		0.00	0.21	0.04	0.08	0.35	0.23
✓	✓	**19.38**	**47.17**	**15.30**	**12.72**	**32.98**	**23.13**

Table 4. Ablation study on the influence of color and segmentation supervision.

	S2T			T2S		
	RR@1	RR@5	NDCG@5	RR@1	RR@5	NDCG@5
+COLOR	7.77	26.94	6.91	5.06	17.21	11.25
+SEG	11.62	29.18	8.53	7.58	21.93	14.91
Parts2Words (COLOR+SEG)	**19.38**	**47.17**	**15.30**	**12.72**	**32.98**	**23.13**

measures. We can observe that our proposed model outperforms other methods on the 3D-Text dataset. Our best result at RR@1 are 19.38 and 12.72 for shape-to-text (S2T) retrieval and text-to-shape (T2S) retrieval, which achieves a 12.64% relative improvement compared to current SOTA methods.

Additionally, to make our experiment more comparable and convincing, we also conducted four additional experiments by modifying our proposed network to approximate other existing methods. The results are also presented in Figure 1. We designed two global-based matching networks, Global-Max and Global-Avg, to show the performance of point-based global matching, which simply use max pooling (Global-Max) and average pooling (Global-Avg) operation on both features of all points and embeddings of all words separately. Besides, we also presented a local-based matching network called LocalBaseline, by replacing our matching module with the stack cross attention module (SCAN) [27] in image-text matching. We train this network by two approaches: LocalBaseline is a two-stage training approach based on a pre-trained segmentation network, identical to the SCAN which freezes the detector model to produce features from image patches; LocalBaseline+ is an end-to-end training approach, identical to our proposed method. We achieved 1.61 and 0.78 improvements in terms of RR@1 by comparing with LocalBaseline+. The comparison with four modified networks proved that the local-based matching approaches extract more detailed features than the global-based ones, our optimal transport-based matching module learns better joint embeddings than the popular SCAN module, and the proposed end-to-end multi-task learning approach makes a better connection between matching embeddings and segmentation prior information.

As shown in Figure 6, we plot the curve of RR@K of 5 methods, among which Parts2Words, Parts2Words-CD, and LocalBaseline+ are local-based methods, and Global-Avg and Global-Max are global-based methods. Parts2Words-CD was obtained by replacing EMD with CD (Chamfer Distance) on the basis of Parts2Words, and further analysis was conducted on this in subsequent ablation experiments. It can be seen that the recall rate of the local matching methods with segmentation prior is obviously higher than the global matching methods. Meanwhile, we can also observe that the proposed Parts2Words achieves the best results compared to other local matching models.

The examples of T2S and S2T retrieval results are shown in Figure 5 and Figure 4. From Figure 5, we display top-5 retrieved texts, we can see that our model can match an average of 2-3 ground truth texts within top-5 best-matched

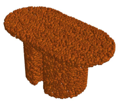
Query Shape	Parts2Words	TriCoLo
	1. a wooden oval brown small table . it have a rectangular hole at the middle below the table - top , seem like it have two leg . 2. a capsule shape wooden table with two half cylinder shape leg . 3. brown color , oval shape , wood material , and physical appearance table 4. this be a light brown wooden table , with a flat capsule shape surface at the top , and two thick leg that be vertically tall , and be semicircle shape with the flat side face the middle create an empty space . 5. brown colored whole wood oval shape coffee table .	1. this oval light wood topped table is on a dark wood base . 2. a wooden oval brown small table . it has a rectangular hole at the middle below the table top seems like it has two legs . 3. an oval shaped table with two legs . it is also wooden and brown . 4. an brown oval table with three section base 5. brown color rectangle shape wood material and physical appearance table
	1. a red chair with a back . 2. a red chair . the seat be curved downward and the back have a gap . 3. red side chair 4. maroon color chair with four leg and rest at back 5. it be a wooden chair . it be red in color .	1. a wooden chair red in color 2. it is a wooden chair . it is red in color . 3. a wooden chair with red colour back and seat with spindle and strong four legs 4. a red wooden kitchen chair with detached back and slightly rounded seat 5. this is wooden chair with four legs and it is in red texture light weight
	1. marble round table with metal leg . table be black color . 2. black , round metal outdoor table . with long curl leg . 3. a center - table with black circular top and four curved leg 4. a stylish black color round table for all - purpose 5. a circular black table .	1. the table is circular with three legs . the table is black and the legs stick out from the top. 2. a black color round shaped wooden table with three legs 3. a black colored round table with four slim shaped legs 4. black round metal outdoor table with long curled legs . 5. black round table three legs wooden material

Figure 8. Retrieval result of the proposed Parts2Words and TriCoLo [37], the ground truth description are marked in red. Each retrieval case under our model can match 2 or 3 ground truth sentences, which is more than TriCoLo.

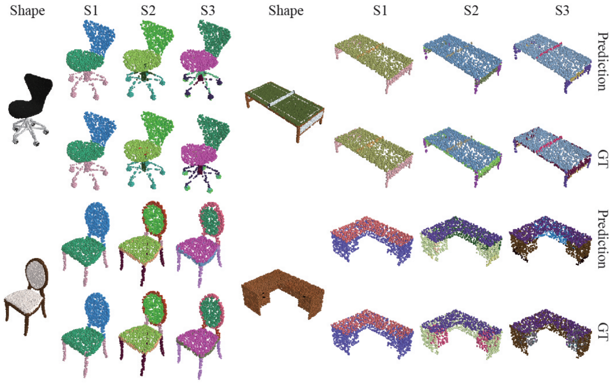


Figure 9. Segmentation results in three different granularity.

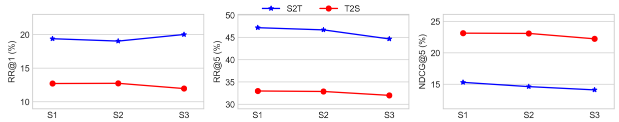


Figure 10. Comparison in terms of Recall@1(RR@1), Recall@5(RR@5) and NDCG@5 using different segmentation granularity. S1, S2, and S3 represent the representing different segmentation granularities from coarse to fine, which have 17 classes, 72 classes, and 90 classes separately. The segmentation with the coarsest granularity **S1** achieves a better performance.

texts for each case. From Figure 4, we can see the top-5 retrieval shapes have a more similar appearance, especially in details such as color and geometry. The results demonstrate that our method could find correspondences in details, such as color and geometric description. In particular, for complex shapes, our model can still achieve superior results.

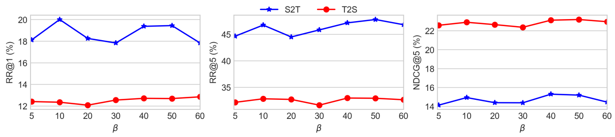


Figure 11. Comparison in terms of Recall@1(RR@1), Recall@5(RR@5) and NDCG@5 using different loss weight β. According to the above result, we choose loss weight β as 40.

We also compare with with TriCoLo [37] by presenting 3 text-to-shape retrieval results and 3 shape-to-text retrieval results, as shown in Figures 7 and 8. For each query text/shape, we display top-5 retrieval shapes/texts ranked by the similarity scores from the two methods. The ground truth shapes/texts are marked in red. And as shown in Figure 7, our model matches the ground truth shapes within top-2 retrieval shapes. For the S2T retrieval result in Figure 8, our model can match 2 or 3 ground truth sentences within top-5 retrieval results, which is more than the ones that TriCoLo can match.

4.2. Ablation Study

Granularity We explore the impact of part embedding extracted under different segmentation granularities on the matching model. The PartNet dataset contains hierarchical segmentation annotation in three granularities from **S1** to **S3**, representing different segmentation granularities from coarse to fine, which have 17 classes, 72 classes, and 90 classes separately. The ground truths and our semantic predictions are shown in Figure 9. As shown in Figure 10, the results show that our method with coarse part segmentation annotation achieves the best performance. We believe that the parts learned through finer segmentation ground truth are hard to align with the simple plain words. Therefore, the coarsest part segmentation annotation **S1** is selected.

Loss Weight We adopt the balance weight β to adjust the joint loss to make our network focus more on the retrieval performance than the segmentation result. We present our retrieval performance when selecting different β to select the most suitable parameter settings, as shown in Figure 11. The retrieval performance improves with the increase of the weight of retrieval loss until β is 40.

Backbone We analyzed the influence of different network backbones on the retrieval results. We use PointNet and PointNet++ respectively as the backbone for the feature extraction network. As shown in Table 2, the comparison results show that the retrieval result gets worse after replacing the backbone with PointNet++.

CD/EMD In Table 2, we replace the EMD for matching similarity calculation by CD (Chamfer Distance), which can be regarded as a local optimal hard matching method. In the Chamfer distance matching flow, each node only corresponds to the node with the most similar individual. The experimental results show that the EMD is better than CD.

Sampling We explore the impact of different negative sample learning strategies based on the triplet ranking loss on retrieval. We compared two strategies: hardest negative mining and semi-hard negative mining. As demonstrated in Table 3, we find that the model using the hardest negative mining results in a collapsed model.

Training We examine the effectiveness of end-to-end training by joint multi-task learning as shown in Table 3. In comparison, we separately train the shape encoder module and the matching module. Our results demonstrate that the end-to-end model outperforms the separate training approach.

Color and Segmentation As shown in Table 4, we separately evaluate the influence of integrating the point cloud color into the feature fusion step and supervising the training with the segmentation loss. The results demonstrate that, after removing these two components (w/o color and w/o seg loss), it is hard for the network to learn to establish a local alignment between color and geometry, which will lead to a significant decrease in accuracy. It indicates that color and part information are crucial for multi-modal retrieval.

4.3. Visualization

We visualized two examples of the best matching pair between the shape and the text, as shown in Figure 12. The normalized similarity matrix is colored by calculating each pairwise distance $D = \{1 - c_{ij}\}$ between parts and words. We can see that our model can accurately find the correspondence (dark red) between parts and words. For the first example, the chair is firstly segmented into 4 parts, then the black seat matches the words ”black” and ”seat” in the text well, and the rest also attends the words ”yellow”, ”black” and ”rest”. Besides, the similarity weight between the part of blue legs and the word ”blue” obtained the highest score.

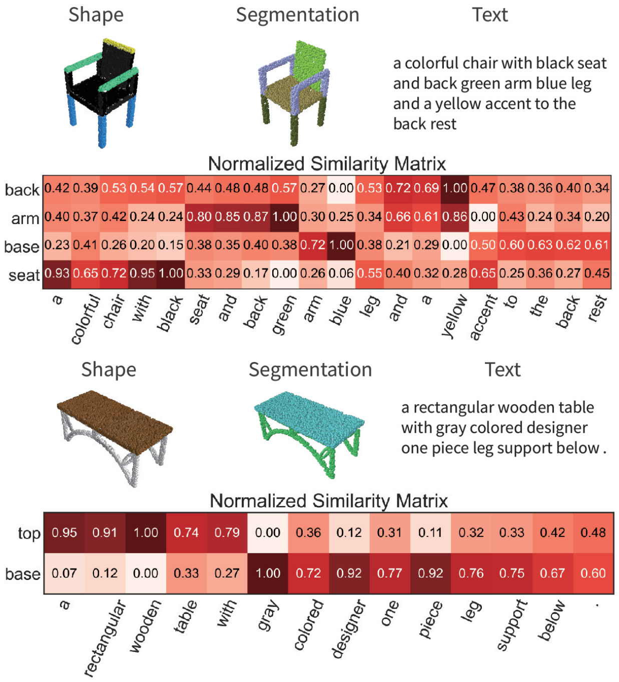


Figure 12. Visualization of the similarity between parts and words. We display a normalized similarity matrix between parts and words. The place with darker color indicates the higher relationship between corresponding parts and words.

5. Conclusion

We introduce a method to learn the joint embedding of 3D point clouds and text. Our method successively increases the ability of a joint understanding of 3D point clouds and text by learning to bidirectionally match 3D parts to words in an optimized space. We obtain the 3D parts by leveraging a 3D segmentation prior, which effectively resolves the self-occlusion issue of parts that is suffered by current multi-view based methods. We also demonstrate that matching 3D parts to words using the optimal transport is an efficient way to merge different modalities including 3D shapes and text in a common space, where the proposed cross-modal earth mover’s distance is also justified to effectively capture the relationship of part-word in this matching procedure. Experimental results show that our method significantly outperforms other state-of-the-art methods.

Acknowledgement. This work is supported by the Young Scientists Fund of the National Natural Science Foundation of China (Grant No.62206106).

References

[1] Panos Achlioptas, Judy Fan, Robert Hawkins, Noah Goodman, and Leonidas J Guibas. Shapeglot: Learning language for shape differentiation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8938–8947, 2019. 2

[2] Panos Achlioptas, Ian Huang, Minhyuk Sung, Sergey Tulyakov, and Leonidas Guibas. ChangeIt3D: Language-assisted 3d shape edits and deformations. *https://changeit3d.github.io/*, 2022. 2

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proc. IEEE Conf. on Computer Vision & Pattern Recognition*, pages 6077–6086, 2018. 3

[4] Paul J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 14(2):239–256, 1992. 5

[5] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 1, 2, 4, 5

[6] Kevin Chen, Christopher B Choy, Manolis Savva, Angel X Chang, Thomas Funkhouser, and Silvio Savarese. Text2shape: Generating shapes from natural language by learning joint embeddings. In *Proc. Asian Conf. on Computer Vision*, pages 100–116, 2018. 1, 2, 4, 5

[7] Tianlang Chen and Jiebo Luo. Expressing objects just like words: Recurrent visual embedding for image-text matching. In *Proc. AAAI Conference on Artificial Intelligence*, pages 10583–10590, 2020. 3

[8] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: improving visual-semantic embeddings with hard negatives. In *Proc. British Machine Vision Conference*, 2018. 2

[9] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan C. Russell, and Mathieu Aubry. A papier-mâché approach to learning 3D surface generation. In *Proc. IEEE Conf. on Computer Vision & Pattern Recognition*, 2018. 2

[10] Zhizhong Han, Chao Chen, Yu-Shen Liu, and Matthias Zwicker. DRWR: A differentiable renderer without rendering for unsupervised 3D structure learning from silhouette images. In *Int. Conf. on Machine Learning*, 2020. 2

[11] Zhizhong Han, Chao Chen, Yu-Shen Liu, and Matthias Zwicker. Shapecaptioner: Generative caption network for 3d shapes by learning a mapping from parts detected in multiple views to sentences. In *Proc. ACM Int. Conf. on Multimedia*, pages 1018–1027, 2020. 1, 2, 5

[12] Zhizhong Han, Mingyang Shang, Xiyang Wang, Yu-Shen Liu, and Matthias Zwicker. Y2seq2seq: Cross-modal representation learning for 3d shape and text by joint reconstruction and prediction of view and word sequences. In *Proc. AAAI Conference on Artificial Intelligence*, volume 33, pages 126–133, 2019. 1, 2, 4, 5

[13] Zhizhong Han, Xiyang Wang, Yu-Shen Liu, and Matthias Zwicker. Multi-angle point cloud-vae:unsupervised feature learning for 3D point clouds from multiple angles by joint self-reconstruction and half-to-half prediction. In *Proc. Int. Conf. on Computer Vision*, 2019. 2

[14] Yining Hong, Qing Li, Song-Chun Zhu, and Siyuan Huang. Vlgrammar: Grounded grammar induction of vision and language. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1645–1654. IEEE, 2021. 2

[15] Tao Hu, Zhizhong Han, Abhinav Shrivastava, and Matthias Zwicker. Render4Completion: Synthesizing multi-view depth maps for 3D shape completion. *arXiv preprint arXiv:1904.08366*, 2019. 2

[16] Tao Hu, Zhizhong Han, and Matthias Zwicker. 3d shape completion with multi-view consistent inference. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 10997–11004. AAAI Press, 2020. 2

[17] Zhibin Hu, Yongsheng Luo, Jiong Lin, Yan Yan, and Jian Chen. Multi-level visual-semantic alignments with relation-wise dual attention network for image and text matching. In *Proc. Int. Joint Conf. on Artificial Intelligence*, pages 789–795, 2019. 3

[18] Rui Huang, Xuran Pan, Henry Zheng, Haojun Jiang, Zhifeng Xie, Shiji Song, and Gao Huang. Joint representation learning for text and 3d point cloud, 2023. 2

[19] Eldar Insafutdinov and Alexey Dosovitskiy. Unsupervised learning of shape and pose with differentiable point clouds. *Advances in Neural Information Processing Systems*, pages 2807–2817, 2018. 2

[20] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Information Systems*, 20(4):422–446, 2002. 5

[21] Z. Ji, H. Wang, J. Han, and Y. Pang. Sman: Stacked multimodal attention network for cross-modal image-text retrieval. *IEEE Trans. Cybernetics*, pages 1–12, 2020. 1, 3

[22] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proc. IEEE Conf. on Computer Vision & Pattern Recognition*, pages 3128–3137, 2015. 1, 3

[23] Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. *Advances in Neural Information Processing Systems*, 27:1889–1897, 2014. 1, 2

[24] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. Int. Conf. on Learning Representations*, 2015. 5

[25] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 2

[26] Juil Koo, Ian Huang, Panos Achlioptas, Leonidas J. Guibas, and Minhyuk Sung. Partglot: Learning shape part segmen-

tation from language reference games. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16484–16493. IEEE, 2022. 2

[27] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proc. Euro. Conf. on Computer Vision*, pages 201–216, 2018. 1, 3, 6

[28] Kuang-Huei Lee, Hamid Palangi, Xi Chen, Houdong Hu, and Jianfeng Gao. Learning visual relation priors for image-text matching and image captioning with neural scene graph generators. *arXiv preprint arXiv:1909.09953*, 2019. 3

[29] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. *Advances in Neural Information Processing Systems*, 31:820–830, 2018. 2

[30] Xinhai Liu, Zhizhong Han, Yu-Shen Liu, and Matthias Zwicker. Point2Sequence: Learning the shape representation of 3D point clouds with an attention-based sequence to sequence network. In *Proc. AAAI Conference on Artificial Intelligence*, pages 8778–8785, 2019. 2

[31] Xinhai Liu, Zhizhong Han, Wen Xin, Yu-Shen Liu, and Matthias Zwicker. L2G auto-encoder: Understanding point clouds by local-to-global reconstruction with hierarchical self-attention. In *Proc. ACM Int. Conf. on Multimedia*, 2019. 2

[32] Zhengzhe Liu, Yi Wang, Xiaojuan Qi, and Chi-Wing Fu. Towards implicit text-guided 3d shape generation. abs/2203.14622, 2022. 2

[33] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). In *Proc. Int. Conf. on Learning Representations*, 2015. 2

[34] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *Proc. IEEE Conf. on Computer Vision & Pattern Recognition*, pages 909–918, 2019. 1, 4, 5

[35] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proc. IEEE Conf. on Computer Vision & Pattern Recognition*, pages 652–660, 2017. 2, 3

[36] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in Neural Information Processing Systems*, 30:5099–5108, 2017. 2, 3

[37] Yue Ruan, Han-Hung Lee, Ke Zhang, and Angel X. Chang. Tricolo: Trimodal contrastive loss for fine-grained text to shape retrieval. abs/2201.07366, 2022. 1, 2, 4, 5, 6, 7

[38] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proc. IEEE Conf. on Computer Vision & Pattern Recognition*, pages 815–823, 2015. 4

[39] Yaxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. Position focused attention network for image-text matching. In *Proc. Int. Joint Conf. on Artificial Intelligence*, pages 3792–3798, 2019. 3

[40] Xin Wen, Zhizhong Han, Geunhyuk Youk, and Yu-Shen Liu. CF-SIS: Semantic-instance segmentation of 3D point clouds by context fusion with self-attention. In *Proc. ACM Int. Conf. on Multimedia*, 2020. 2

[41] Xin Wen, Tianyang Li, Zhizhong Han, and Yu-Shen Liu. Point cloud completion by skip-attention network with hierarchical folding. In *Proc. IEEE Conf. on Computer Vision & Pattern Recognition*, 2020. 2

[42] Wenxuan Wu, Zhongang Qi, and Li Fuxin. Pointconv: Deep convolutional networks on 3d point clouds. In *Proc. IEEE Conf. on Computer Vision & Pattern Recognition*, pages 9621–9630, 2019. 2

[43] Xing Xu, Tan Wang, Yang Yang, Lin Zuo, Fumin Shen, and Heng Tao Shen. Cross-modal attention with semantic consistence for image-text matching. *IEEE Trans. Neural Networks and Learning Systems*, 31(12):5412–5425, 2020. 3