Structure-Aware 3D VR Sketch to 3D Shape Retrieval

Ling Luo[1,2]   Yulia Gryaditskaya[1,2,3]   Tao Xiang[1,2]   Yi-Zhe Song[1,2]

[1]SketchX, CVSSP, University of Surrey, United Kingdom
[2]iFlyTek-Surrey Joint Research Centre on Artificial Intelligence
[3]Surrey Institute for People Centred AI and CVSSP, University of Surrey, United Kingdom

arXiv:2209.09043v1 [cs.CV] 19 Sep 2022

Abstract

We study the practical task of fine-grained 3D-VR-sketch-based 3D shape retrieval. This task is of particular interest as 2D sketches were shown to be effective queries for 2D images. However, due to the domain gap, it remains hard to achieve strong performance in 3D shape retrieval from 2D sketches. Recent work demonstrated the advantage of 3D VR sketching on this task. In our work, we focus on the challenge caused by inherent inaccuracies in 3D VR sketches. We observe that retrieval results obtained with a triplet loss with a fixed margin value, commonly used for retrieval tasks, contain many irrelevant shapes and often just one or few with a similar structure to the query. To mitigate this problem, we for the first time draw a connection between adaptive margin values and shape similarities. In particular, we propose to use a triplet loss with an adaptive margin value driven by a 'fitting gap', which is the similarity of two shapes under structure-preserving deformations. We also conduct a user study which confirms that this fitting gap is indeed a suitable criterion to evaluate the structural similarity of shapes. Furthermore, we introduce a dataset of 202 VR sketches for 202 3D shapes drawn from memory rather than from observation. The code and data are available at `https://github.com/Rowl1ng/Structure-Aware-VR-Sketch-Shape-Retrieval`.

1. Introduction

Progress in VR/AR technologies poses requirements on the availability of 3D content, thereby contributing to the active development of research on 3D shapes [3, 8, 39, 2]. This is further accompanied by the growth of large 3D shapes collections. In our work, we study *3D VR sketching*, 3D sketching by means of a VR headset and VR controllers[1], in a minimal sketching interface as an intuitive tool

for 3D shapes retrieval and exploration.

Sketch was proven to be an efficient media for 2D images retrieval [41, 29]. However, despite recent attempts to address the domain gap between 3D shapes and 2D sketches [24, 20], instance-level retrieval of 3D shapes from 2D sketches remains challenging at large. On the contrary, recent works [22, 23] demonstrate the advantage of 3D VR sketches for 3D shapes retrieval. Despite encouraging results, these models return many irrelevant shapes among the top retrieval results (Fig. 1). This is largely due to the abstraction and distortion inherent in VR sketches. In this work, we strive for a system that accounts for such inaccuracies during training and returns among the top retrieval results shape that are structurally similar to the query sketch.

Traditionally fine-grained retrieval is evaluated by ranking the retrieval results for each query according to a used distance metric and counting the percentage of instances that have ground-truth among the top-k ranked results. As freehand sketches frequently contain distortions, top-k accuracy does not allow to take into account the cases where the top retrieval results perfectly match the sketch but do not contain the ground-truth itself, as we show in Fig. 2. To measure structural similarities between 3D shapes, we adopt a *fitting gap* [32, 33] that evaluates how similar a given 3D shape is to another 3D shape under a structure-preserving deformation. In particular, we propose an *adaptive margin scheme* for a triplet loss, commonly used for retrieval tasks [35, 28, 36, 11], to explicitly exploit this fitting gap. We further leverage the recent progress in deep structure-preserving deformation methods [39, 14] to compute the fitting gap on the fly during training. We demonstrate that our approach outperforms losses proposed in the context of deformation-aware retrieval [32, 33] in the task of the structure-aware 3D shape retrieval from a 3D VR sketch. Furthermore, for the first time, we leverage the fitting gap as an evaluation criterion for retrieval. To this end, we conduct a user study that shows that this criterion by far outperforms

---

[1]Note that in general, a 3D sketch can also be created using Kinect or some 2D digital devices. As different sketching setups result in different quality of sketches and their distortions, we refer to sketches created with VR headsets/controllers as 3D VR sketches.
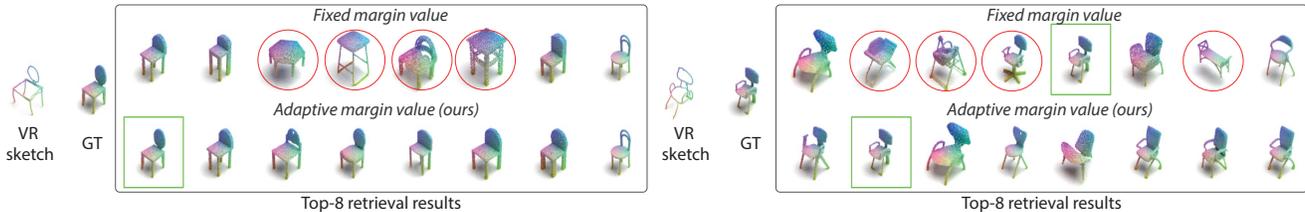
Figure 1: We propose an adaptive margin scheme within a triplet loss ubiquitously used for retrieval tasks. As shown in this figure, this strategy allows to obtain more 'relevant' retrieval results that closer match the input sketch and the ground-truth 3D shape structure. The green square highlights the ground-truth 3D shape among the retrieval results, while the red circles mark erroneous retrieval results (dissimilar shapes).
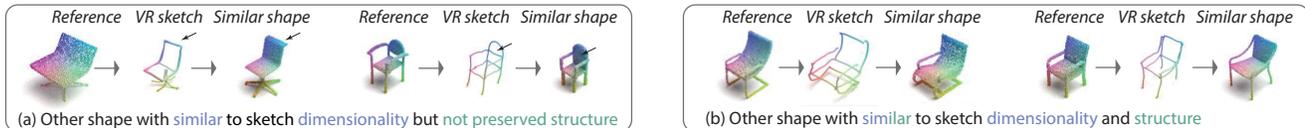

Figure 2: The VR sketches shown here were collected from the 'reference' 3D shapes [23]. The inherent distortions and sparsity of VR sketches pose challenges for the fine-grained 3D shape retrieval task. There may be a similar 3D shape that matches the overall sketch dimensions similar or better than the ground-truth but may have a worse (a) or better match (b) in structure and details than the ground-truth. Therefore, we argue for a practical retrieval system that should be able to not only take the overall shape dimensions into account (a), but match the shapes with a similar structure (b). Likewise, the evaluation requires criteria that allow evaluating the structural similarity of the retrieved shapes (b).

the Chamfer distance in measuring structural similarities, as judged by a human observer. Last but not least, we also show that this fitting gap-based criterion provides additional insights into retrieval performance.

In summary, our contributions include: (i) a novel formulation of a triplet loss with an adaptive margin value driven by the fitting gap between a ground-truth 3D shape and a given gallery shape, (ii) a perceptual study of 3D shapes distance measures on their effectiveness on evaluating structural similarities, (iii) adoption of the fitting gap as a distance metric to evaluate how well the retrieved results preserve structure and provides more fine-grained insights into retrieval performance, and (iv) a new test set of VR sketches drawn from memory.

2. Related Work

Sketch Based Retrieval Our work is inspired by the success of fine-grained and category-level 2D-sketch-based image retrieval (SBIR) methods [34, 27, 41, 21, 4, 29, 42]. Similar to majority of these works we leverage the triplet loss, commonly used in the context of fine-grained and category-level retrieval [35, 28, 36, 11].

However, 2D-sketch-based 3D shape retrieval (SBSR) was mostly demonstrated in the context of *category-level* retrieval: the pioneering hand-crafted feature based methods [6, 40, 18, 19] and the more recent deep learning based methods [37, 34, 45, 5, 13, 25]. Qi *et al*. [24] are the first to target fine-grained retrieval of 3D shapes from 2D sketches. They show that existing methods for Fine-Grained (FG)

SBIR and category-level SBSR perform poorly on the task of FG-SBSR. Similarly to this work, we are interested in fine-grained retrieval of a 3D shape (retrieval of a particular shape instance), but consider a 3D VR sketch as a query. In the supplemental, we show that our method outperforms [24] results by 7.44 points in top-1 retrieval accuracy on a similar training and test set sizes on a chair category.

VR-Sketch-Based Shape Retrieval We are not the first to consider 3D-VR-SBSR, but as with 2D sketches, most of the existing methods for 3D-shape retrieval target category-level retrieval [16, 17, 15, 38, 22]. The earlier works [16, 17, 15, 38] relied on the dataset of 300 human-drawn 3D sketches of 30 classes, each with 10 sketches obtained with a Kinect-based virtual 3D drawing system (not available online anymore). Luo *et al*. [22] proposed a heuristics-based algorithm to generate synthetic 3D sketches from the reference 3D shapes, that they leveraged to train deep models. Giunchi *et al*. [9, 10] considered fine-grained retrieval of 3D shapes when a VR sketch is created on top of some reference 3D shape. In their case, the VR sketch comprises a set of dense ribbon-like strokes. Luo *et al*. [23] studied FG-3D-VR-SBSR from sparse quick sketches and introduced the first dataset comprising 1,497 3D VR sketch and 3D shape pairs of chairs. They performed benchmarking on common 3D shape retrieval methods and demonstrated encouraging results. In our work, we exploit their dataset and propose an adaptive margin scheme for a triplet loss, taking into account structural similarities between shapes.

Triplet loss with adaptive margin Triplet loss is commonly used for retrieval tasks [35, 28, 36, 11]. It requires defining for each training set sample the sets of positive ('similar') and negative ('dissimilar') samples. It results in the latent space where 'similar' instances are encoded closer than 'dissimilar' ones. It performs well in classification tasks, but is less suitable for the continuous ratings [12], which is the case for our problem, as we want all top-k retrieval results to be structurally similar to the query sketch. Variants of adaptive margin schemes were shown to be efficient for pose estimation [43], image retrieval [44, 12] and facial expression recognition [31].

Deformation-aware retrieval Our problem of retrieving a 3D shape that matches the structure of a 3D query sketch bears similarities with the problem of deformation-aware retrieval [32, 33]. These works address the problem of retrieving a clean and complete 3D model from a noisy and/or partial 3D scan. They argue for the scenario where the goal is to retrieve the shape that after the deformation is the closest to the query. In this context, they define the *fitting gap* as the difference between the deformed gallery model and the query. In our case, the *fitting gap* is measured between gallery shapes and a ground-truth 3D shape of a query sketch. Our work is closest to [33], where they exploit the fitting gap to ensure that the gallery shapes are embedded closer to the query shapes they can deform into. They leverage the regression loss, first proposed in [32]. Instead, we propose a strategy for an adaptive margin value setting driven by a fitting gap. We show its superiority over the regression loss. Further, Uy et al. [33] jointly train for the deformation and retrieval in an alternating fashion, keeping one module fixed when optimizing the other and vice versa. Their deformation module and embedding space are optimized for deforming the retrieved shapes to the target shapes. In our work, the deformation module *is trained first* and then used to guide the training to obtain the embedding space where shapes/sketches with similar structure are close to each other. Our perceptual study indicates the validity of this measure for 3D shapes structural similarity, outperforming existing ones. We aim at 3D VR sketch understanding and retrieval of a gallery shape that matches the query sketch best, prioritizing structural similarities.

3. Method

Following [23], we represent 3D shapes and 3D sketches as point clouds, and train the model via a Siamese training with a triplet loss [35, 28, 36, 11]. As a 3D sketch and shape encoder we exploit PointNet++[26], where the same set of weights is used to encode both modalities.

We denote the feature embedding of a sketch S as $s \in \mathbb{R}^{512}$ (source) and a shape T as $t \in \mathbb{R}^{512}$ (target). For a given batch of N sketch-shape pairs, in [23] the triplet loss

is defined as follows:

$$L_T = 1/|N_S^n| \sum_{t_S^n \in N_S^n} [d(s, t_S^p)^2 - d(s, t_S^n)^2 + m]_+, \quad (1)$$

where t_S^p is an embedding of the shape T_S^p matching the considered sketch S. $N_S^n = \cup T_S^n$ is a set of negative shape examples, T_S^n, which are all shapes in a batch excluding T_S^p. m is a margin, $[\cdot]_+$ denotes the clipping of negative values (a hinge loss), and d is a Euclidean distance in the embedding space. Feature vectors are L_2 normalized. The triplet loss ensures that the source sketch and its matching shape are closer in the feature space than the source sketch and all other 3D shapes.

We propose an adaptive margin that determines how far to push the embeddings of a sketch s and its matching shape from the embeddings of shapes t_S^n in the negative set of shapes based on how similar the 3D shapes T_S^p (sketch matching shape) and T_S^n (some shape from the set of negative shapes) are. In particular, let $\delta(\cdot, \cdot)$ be some distance between two 3D shapes. Then, for each shape, we compute distances to each other shape. Let $M_S(T_S^p, N_S^n)$ be the maximum distance from the shape T_S^p to shapes in N_S^n. We compute the margin value for each pair of shapes T_S^p and T_S^n as follows:

$$m \coloneqq m(T_S^p, T_S^n) = \alpha + (\beta - \alpha)\frac{\delta(T_S^p, T_S^n)}{M_S(T_S^p, N_S^n)}, \quad (2)$$

where α and β are method hyper-parameters. This mapping has nice properties: in the unlikely case that all the shapes in the batch are similar, this loss will try to push them far apart, still satisfying fine-grained retrieval requirements. In the more likely case and on condition of a sufficiently large batch size, $M_S(T_S^p, N_S^n)$ will be similar between batches, ensuring consistent behavior of the mapping driven only by $\delta(T_S^p, T_S^n)$.

3.1. Choice of the distance δ in 3D space

One obvious choice of the distance between two 3D shapes is the Chamfer distance [1], which is by far the most commonly used distance for training and evaluating similarities between two 3D shapes [7]. In this work, we consider additional criteria to compare 3D shapes which prioritize structural similarities.

First, we define an *asymmetric fitting gap* as the distance between the sketch matching shape T_S^p and the deformed target/gallery shape T_S^n to the T_S^p:

$$\delta_{CD}^A(T_S^p, T_S^n) \coloneqq CD(\mathcal{D}(T_S^n; T_S^p), T_S^p), \quad (3)$$

where $\mathcal{D}(T_S^n; T_S^p)$ is a deformation operator that deforms T_S^n to T_S^p, and CD denotes the Chamfer distance. We refer to this definition of $\delta_{CD}^A(\cdot, \cdot)$ as the *asymmetric fitting gap* since we only deform the target shapes to the source
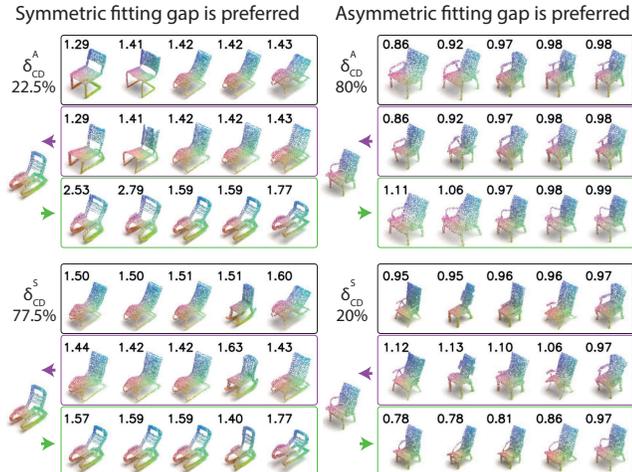
Figure 3: Examples of shapes for which asymmetric or symmetric fitting gaps are preferred (Sec. 4.1). The shapes in the black rectangles show the top-5 ranked 3D shapes according to the asymmetric δ_{CD}^A or symmetric δ_{CD}^S fitting gaps. The shapes in the purple and green rectangles are the deformation results and their respective fitting gap values. The arrows indicate the deformation direction. Thus, the shapes in the purple rectangles are the gallery shapes deformed towards the query shape.

ground-truth (sketch matching) shape. The intuition behind the asymmetric fitting gap follows the findings by Uy et al. [32] that while one shape can be easily deformed into another, it might not be true in the opposite direction. The direction of the deformation is driven by the observation that the structure of the retrieval results is more likely to be what the user had in mind if the retrieval results can be deformed to the target shapes.

Then, we define a *symmetric fitting gap* $\delta_{CD}^S(T_S^p, T_S^n)$ as:

$$1/2\left(CD(\mathcal{D}(T_S^n; T_S^p), T_S^p) + CD(T_S^n, \mathcal{D}(T_S^p; T_S^n))\right). \quad (4)$$

The rational behind the symmetric fitting gap is that if the deformation can be accurately performed in both directions, then the shapes are more likely to have the similar structure. However, the risk is that the results for some shapes may be biased due to the efficiency of the deformation of the source into the target shapes.

As the deformation operator, we leverage a recent neural deformation operator [39] that allows efficient computation of the margin value during training, and was shown to work well on diverse classes of shapes. We train the deformation operator on the 3D shapes from the chair class of the ShapeNet dataset. We visualize the deformation results in supplemental materials.

4. Perceptual study

In this section, we study the appropriateness of the fitting gap as an evaluation criterion of the structural similarities between 3D shapes, as judged by a human observer. First, we compare the asymmetric and symmetric fitting gaps. Second, we evaluate the asymmetric fitting gap against the Chamfer distance and F-score.

4.1. Symmetric vs. asymmetric fitting gap

Data preparation We select a subset of 24 diverse shapes from the ShapeNet chair category, on which the top-5 retrieved shapes for two versions of the fitting gap differ visually strongly from each other (some examples are shown in Fig. 3). Then, for each shape, we perform a nearest neighbor retrieval among the 5,274 shapes. We represent shapes as point clouds by uniformly sampling 1,024 points.

Task We showed to 39 participants the top 5 ranked retrieval results according to two versions of the fitting gap. We asked the participants to select the retrieval results that *have the most similar structure to the reference*, as judged from left to right.

Results For all shapes and all participants, the asymmetric fitting gap was chosen 49.50% of the time, while the symmetric fitting gap was 50.50% of the time. While on average there is no clear preference for one over the other, we observe that the best measure depends on a shape, as shown in Fig. 3: on different shapes, one measure is chosen over another with a clear gap. Which fitting gap is preferred is driven by how easy one or another deformation direction is. Thus, the example on the left shows the shape for which computing deformation in both direction results in more reliable ranking. Meanwhile, for the example on the right the ranked results are dominated by the deformation direction from the query to the gallery shapes.

4.2. Fitting gap vs. Chamfer distance

In this experiment, we select a subset of 25 diverse shapes from the ShapeNet chair category. Then, for each shape, we perform a nearest neighbor retrieval among 5,274 shapes according to one of the 3 measures described below. The task in this study is the same as the one described in the previous section.

We evaluate the Chamfer distance against the two variants of the asymmetric fitting gap. We choose the asymmetric fitting gap because it requires less computation and performs similarly to the symmetric fitting gap on average. As the first variant of the fitting-gap-based measures, we consider an asymmetric fitting gap, where the distance after the deformation is computed via Chamfer distance (Eq. (3)). We also evaluate the asymmetric fitting gap, based on the F-score with the threshold values set to 0.01:

$$\delta_F^A(T_S^p, T_S^n) := \text{F-score}(\mathcal{D}(T_S^n; T_S^p), T_S^p) \quad (5)$$

It was observed that the Chamfer distance is sensitive to geometric layout of outliers, and F-score was proposed as an
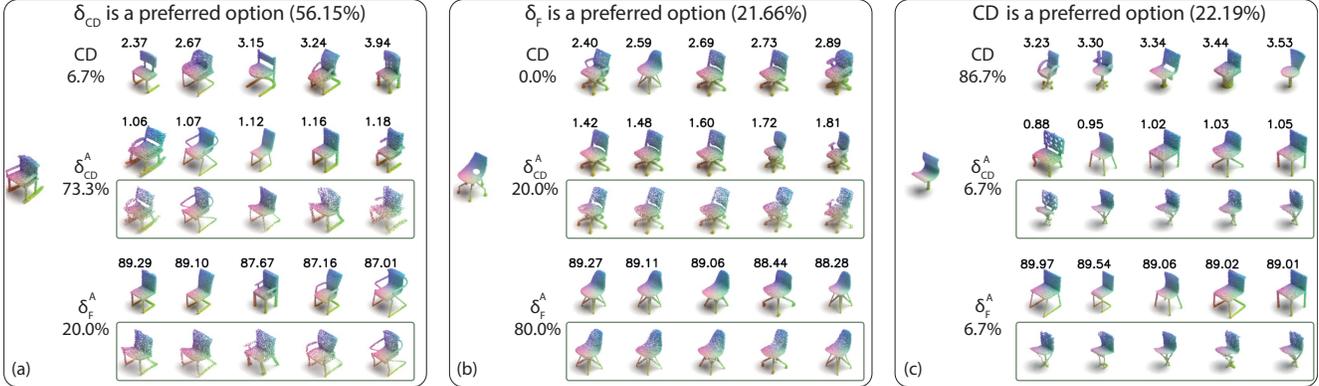
Figure 4: *Human preferences on retrieval results according to different measures, excluding the groundtruth.* The Chamfer Distance (CD), δ_{CD}^A (Eq. (3)) and δ_F^A (Eq. (5)) are selected by participants as the best measure reflecting shapes structure similarity for 16% (4 shapes), 72% (18 shapes) and 12% (3 shapes) of shapes considered in the study, respectively. The values below each measure symbol show the percentage of participants who selected a particular measure for a given shape as the best. The images in green rectangles show the deformations of the top-5 ranked shapes towards the query shape. The numbers above each shape image are the values of the evaluated measure.

additional metric [30]. The ShapeNet 3D shapes are normalized to fit a unit bounding box.

Results We collect responses from 15 participants. Fig. 4 shows representative distributions of human preferences. On average the Chamfer distance CD is preferred 22.19% of the time, δ_{CD}^A (Eq. (3)) is preferred 56.15% of the time, and δ_F^A (Eq. (5)) is preferred 21.66% of the time. We observe a good agreement between the participants: the standard deviations in choosing CD, δ_{CD}^A and δ_F^A comprise 6.85, 7.34 and 7.36 respectively. Yet, there is a larger spread in preferences among the shapes (Fig. 4): the standard deviations in choosing CD, δ_{CD}^A and δ_F^A comprise 22.29, 23.48 and 19.26 respectively. In Fig. 4(c) it can be seen that the Chamfer distance is more likely to be selected when it is too 'easy' to warp any shape to the source groundtruth shape, and the fitting gap is not informative. However, such cases are not very frequent. It also can be seen that the F-score is not very good in capturing the differences between shapes due to small details (Fig. 4(a)). Even though in Fig. 4(b) δ_F^A is preferred by the majority of participants, it can be seen that the fitting gap with the Chamfer distance allows us to retrieve shapes with closer form of the chairs' legs, and containing a hole in the back similar to the ground-truth.

5. Experiments and Analysis

5.1. Datasets

For training and testing we exploit the dataset of 3D VR sketches by Luo *et al.* [23]. This dataset consists of 1,005 unique 3D sketch and shape pairs from the ShapeNet chair category, created by 50 participants. VR sketches in this dataset were collected by displaying the reference shape in the area separate from the sketching area. We refer to this dataset as *FVRS* – a dataset of Freehand VR Sketches.

We use the same split for testing, training and validation sets as was proposed in [23], comprising 202, 702 and 101 sketches, respectively. During testing we perform retrieval from 5,794 shapes unseen during training and validation. The test set contains all the sketches of 5 randomly selected participants (50 sketches), and a subset of sketches of each of the remaining 45 participants (152 sketches). The 45 participants' sketches are split proportionally across the training, validation and test sets. We represent our sketches and shapes by uniformly sampling 1,024 points.

As in our envisioned application scenario a human is sketching from memory or imagination, we collected a small set of sketches from memory to (1) test the performance on such data and (2) being able to analyze the differences with previously collected data from observation. The participants could only sketch when the reference shape was invisible, ensuring that they needed to memorize the reference. They were instructed to sketch the features of the reference shape that discriminate it from others. They could go back to the reference at any time, as the task of memorizing an arbitrary shape is somewhat artificial. This allows to better approximate the scenario where people draw some concrete shapes from memory. For a fair evaluation, we collect additional sketches *for the same* 202 *shapes as the 202 sketches in the test set of the FVRS datasset*. We hire 10 participants, where 8 sketch 20 shapes each and 2 sketch 21 shapes. We refer to this new dataset as *FVRS-M* – a dataset of Freehand VR Sketches by sketching from Memory. The analysis and comparison of sketches from *FVRS-M* and *FVRS* is provided in Fig. 5 and its caption.
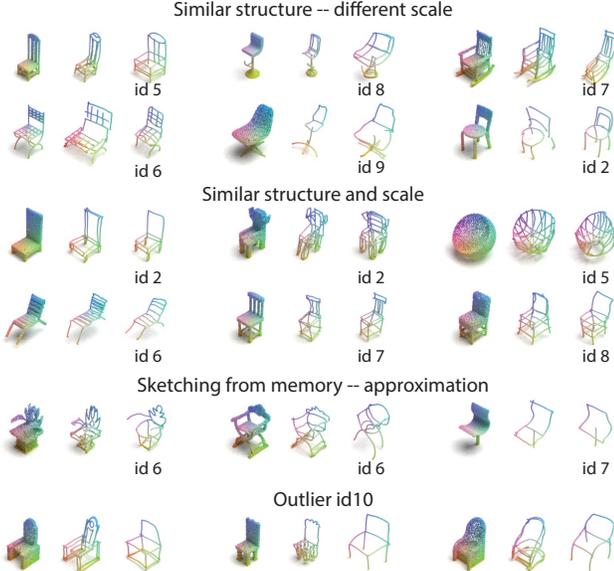
Figure 5: *FVRS-M* vs. *FVRS*: In each triplet of images, the 1st image is a reference shape, the 2nd is a sketch from *FVRS*[23], and the 3rd is a sketch from our new dataset *FVRS-M*. "id" denotes each participant's unique identifier. The 1st row shows examples where the participants sketching from the visible reference preserved the overall proportions better, while the 2nd row shows cases where the participants sketching from memory reproduced the overall proportions better. *Large difference in proportions in these sketches further support our goal of designing a structure-aware retrieval method.* The 3rd and 4th rows show examples in which participants in both sketching scenarios accurately reproduced structure and shape dimensions. The 5th row shows that when sketching from memory, participants draw complex details approximately. Finally, one participant drew too quick and non-detailed sketches (row 6).

5.2. Evaluation criteria

We evaluate the accuracy of the retrieval using 4 different metrics. In all cases, our sketches, source shapes, deformed source and target shapes are represented as point clouds, and points-to-points distances are used for the evaluation. First, we evaluate the results in terms of Top-k accuracy – the standard fine-grained retrieval metric. Top-k accuracy computes the percentage of queries for which the ground-truth shape is among the top k ranked retrieval results. In addition, we compute 3 additional criteria allowing us to make more informed conclusions on how similar the shapes are to the ground-truth among the top k ranked retrieval results. First, we evaluate the similarity in terms of the average Chamfer distance (*Avg CD*) between each of the shapes in the top-k ranked results and the ground-truth. Then, we evaluate the structural similarity (*Avg* $\delta_{CD}^{A/S}$) based

on the average asymmetric/symmetric fitting gap $\delta_{CD}^{A/S}$ between each of the shapes in the top-k ranked results and the ground-truth.

5.3. Results

All models are evaluated by training for 500 epochs. We evaluate using cross-validation. We keep the test set fixed and partition the rest of the data (803 sketch-shape pairs) into 5 different subsets. For each of the 45 sketchers, we split their sketches proportionally, randomly, between the validation and training sets. We select 101 sketches as the validation set and 702 sketches as the training set. The results in the paper are obtained by the last checkpoint. Results chosen by other metrics on validation set can be found in supplemental materials.

5.3.1 Adaptive vs. fixed margin

First, we compare our model against the state-of-the-art, comprising training with a triplet loss based model with a constant margin value [23] with and without data augmentation (Tab. 1). As data augmentation, we apply global sketch/shape distortions as was proposed by Luo *et al.* [23].

We tested three constant margin values in the range $[0.3, 0.9]$ and found the value of 0.6 to give the best top-k retrieval accuracy and 0.3 to perform better according to other criteria. When the data augmentation is used, the margin value of 0.6 consistently outperforms other constant margin values settings. Please see the supplemental for the detailed comparison. The results in Tab. 1 are computed with the margin value of 0.6.

We set our method hyper-parameters α and β in Eq. (2) to 0.3 and 1.2, respectively. The ratio on the right in Eq. (2) takes values in the range $(0, 1]$, so the margin value is in the range $(0.3, 1.2]$. Our ablation study in the supplemental shows that our method is relatively robust to small variations in the values of these parameters, and consistently outperforms the fixed margin setting.

Tab. 1 shows that the proposed adaptive margin scheme results in better performance according to all measures evaluating the average similarity (*Avg X*) for the top-$1, 5, 10$ ranked retrieval results – retrieving the 3D shapes with better structural similarities to the sketch query's ground-truth 3D shape (Tab. 1). Moreover, the fitting gap based baselines (the last 4 rows in Tab. 1a, Tab. 1b) also have higher top-k retrieval accuracy.

Importantly, we observe that for the models with the fixed margin value, the average fitting gap value (*Avg* $\delta_{CD}^{A/S}$) of the top-5 ranked retrieved results on validation set starts to increase after roughly the 30th epoch, and only the fitting gap value for the top-1 ranked retrieved results decreases. This may not be desirable if a user is interested in finding a set of similar shapes, rather than only one.

Table 1: Quantitative comparison (Sec. 5.3) of the proposed method against state-of-the-art [23]. The usage of data augmentation with sketch and shape distortions is denoted by ∗. The best result in each column is highlighted in bold, and the second best result is underlined. *Acc* denotes the top-k retrieval accuracy – the percentage of queries that contain a ground-truth among its top k ranked retrieval results. *Avg X* denotes the average distance between the ground-truth 3D shape and the ranked top-k retrieval results according to the criterion X: CD is the Chamfer distance, $\delta^{A/S}$ is the asymmetric/symmetric fitting gap, $\delta_{CD}^{A/S}$ is defined in Eq. (3) and Eq. (4), for the $\delta_F^{A/S}$ the distance between the shapes after deformation is computed as an F-score with the threshold set to 0.01. *Avg X* values are multiplied by $1e2$

(a) On the sketches from the *FVRS* dataset [23].

Model	Acc ↑			Avg $CD(T_S^p,T_S^n)$ ↓			Avg $\delta_{CD}^A(T_S^p,T_S^n)$ ↓			Avg $\delta_{CD}^S(T_S^p,T_S^n)$ ↓			Avg $\delta_F^A(T_S^p,T_S^n)$ ↑			Avg $\delta_F^S(T_S^p,T_S^n)$ ↑		
	t-1	t-5	t-10	t-1	t-5	t-10	t-1	t-5	t-10	t-1	t-5	t-10	t-1	t-5	t-10	t-1	t-5	t-10
[23]	24.55	43.76	52.57	3.93	4.88	5.17	1.27	1.59	1.69	1.23	1.56	1.67	85.29	81.44	80.18	85.69	81.67	80.35
[23]*	24.95	44.46	53.37	3.50	4.64	4.99	1.15	1.51	1.63	1.12	1.50	1.62	86.58	82.36	80.94	86.91	82.50	81.00
CD	23.47	41.49	49.01	3.18	3.98	4.22	1.10	1.43	1.52	1.07	1.40	1.50	86.92	82.95	81.84	87.22	83.19	82.02
CD*	24.65	43.86	53.27	2.89	3.95	4.21	1.09	1.40	1.49	1.06	1.37	1.47	87.23	83.50	82.25	87.51	83.64	82.37
δ_{CD}^A	24.55	43.86	52.67	3.23	4.16	4.49	1.10	1.40	1.50	1.08	1.39	1.48	86.96	83.26	82.09	87.27	83.43	82.23
δ_{CD}^A *	27.23	47.13	55.25	3.05	4.22	4.54	1.05	1.38	1.49	1.03	1.36	1.47	87.48	83.57	82.32	87.73	83.73	82.45
δ_{CD}^S	25.45	43.96	53.37	3.07	4.13	4.38	1.09	1.41	1.50	1.06	1.38	1.47	87.18	83.21	82.14	87.59	83.49	82.37
δ_{CD}^S *	26.53	46.14	55.15	2.93	4.13	4.49	1.06	1.39	1.48	1.03	1.36	1.45	87.63	83.71	82.39	88.01	83.95	82.57

(b) On the sketches from the *FVRS-M* dataset.

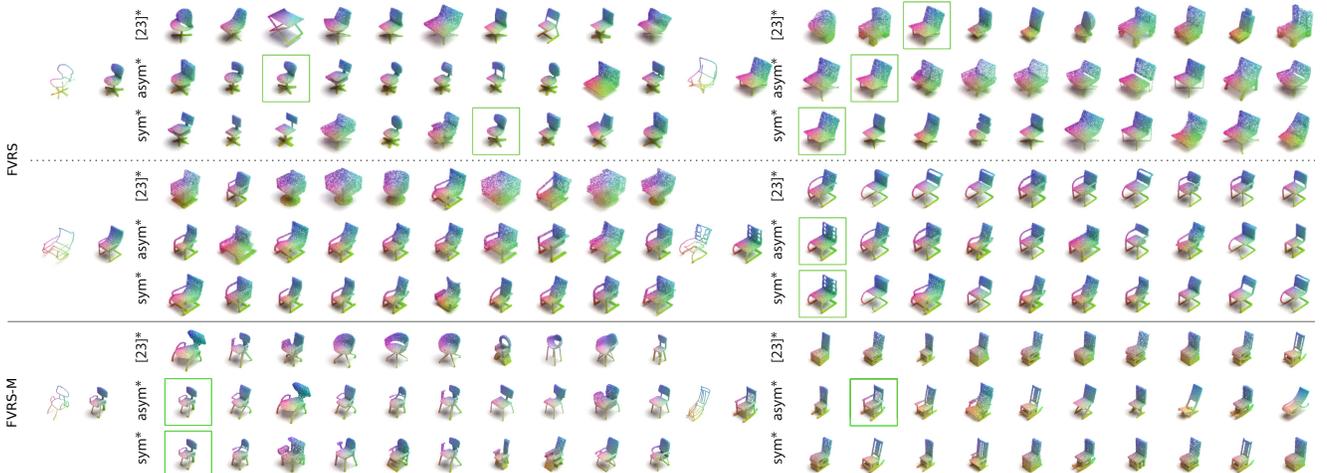| Method	Acc ↑			Avg $CD(T_S^p,T_S^n)$ ↓			Avg $\delta_{CD}^A(T_S^p,T_S^n)$ ↓		
	top-1	top-5	top-10	top-1	top-5	top-10	top-1	top-5	top-10
[23]	15.94	34.55	43.76	4.63	5.27	5.55	1.43	1.64	1.74
[23]*	17.92	36.63	44.36	4.18	5.04	5.39	1.32	1.6	1.71
δ_{CD}^A	17.43	35.35	43.66	3.81	4.58	4.87	1.24	1.46	1.55
δ_{CD}^A *	19.41	38.42	47.33	3.91	4.6	4.87	1.22	1.43	1.51
δ_{CD}^S	16.63	34.36	43.17	3.91	4.6	4.87	1.26	1.49	1.55
δ_{CD}^S *	18.51	36.53	45.45	3.8	4.57	4.83	1.22	1.43	1.52

Figure 6: Qualitative evaluation: The lines in each sub-figure show top-10 ranked retrieval results when data augmentation is used for [23], δ_{CD}^A (asym*) and δ_{CD}^S (sym*) models. The green square denotes the ground-truth matching shape.

On the newly collected sketches from memory *FVRS-M* we obtain lower retrieval accuracy on average (Tab. 1b) as compared to performance on sketches from *FVRS* (Tab. 1a). Our adaptive margin scheme results in better performance according to all criteria: top-k retrieval accuracy and measures evaluating the similarity of the top-k retrieval results.

To demonstrate that our approach is not limited to a chair category, we generate synthetic 3D VR sketches for two additional categories: 'lamp' and 'airplane'. We rely on the method proposed in [22] and set the abstractness level to 1.0, as it was demonstrated to match freehand sketches best. However, as we demonstrate in the supplemental their method still produces very dense sketches for the shapes from the 'airplane' category, closely resembling the 3D shapes. We show in Tab. 2 that our method with an adaptive margin outperforms constant margin setting in terms of

Table 2: Quantitative comparison of the proposed method against state-of-the-art on lamp and airplane shapes. Please see Tab. 1 for the notations.

Category	Model	Acc ↑			Avg $CD(T_S^p, T_S^n)$ ↓			Avg $\delta_{CD}^A(T_S^p, T_S^n)$ ↓			Avg $\delta_{CD}^S(T_S^p, T_S^n)$ ↓			Avg $\delta_F^A(T_S^p, T_S^n)$ ↑			Avg $\delta_F^S(T_S^p, T_S^n)$ ↑		
		t-1	t-5	t-10	t-1	t-5	t-10	t-1	t-5	t-10	t-1	t-5	t-10	t-1	t-5	t-10	t-1	t-5	t-10
lamp	[23]*	63.87	83.23	86.45	1.20	6.37	8.82	0.39	1.54	1.98	0.43	1.86	2.43	95.38	82.57	78.09	94.96	81.55	76.51
	δ_{CD}^A*	67.74	87.10	92.26	1.11	4.00	5.08	0.33	1.24	1.45	0.32	1.35	1.64	96.11	85.29	82.67	95.99	85.02	82.06
	δ_{CD}^S*	65.81	85.16	91.61	0.98	3.78	4.98	0.34	1.21	1.46	0.34	1.31	1.64	95.95	85.71	82.75	95.96	85.54	82.24
airplane	[23]*	81.33	92.00	93.33	0.24	1.34	1.69	0.16	0.59	0.68	0.16	0.62	0.72	99.27	93.58	92.40	99.24	93.29	92.03
	δ_{CD}^A*	82.22	93.78	95.56	0.20	1.11	1.40	0.14	0.52	0.61	0.15	0.53	0.63	99.44	94.69	93.48	99.39	94.54	93.25
	δ_{CD}^S*	82.67	96.00	96.00	0.23	1.08	1.33	0.14	0.49	0.57	0.15	0.51	0.59	99.38	94.97	93.93	99.33	94.83	93.75

Figure 7: Qualitative comparison of the regression loss (RL) [32] and the triplet loss (TL) with an adaptive margin (Eqs. (1) and (2)). The lines in each sub-figure show top-10 ranked retrieval results for (1) TL: δ_{CD}^A with triplet loss , and (2) RL: δ_{CD}^A with regression loss. The green square denotes the ground-truth matching shape.

Table 3: Comparison of the regression loss (RL) [32] and the triplet loss (TL) with an adaptive margin (Eqs. (1) and (2)). Please see Tab. 1 for the notations.

		Acc ↑			Avg $CD(T_S^p, T_S^n)$ ↓			Avg $\delta_{CD}^A(T_S^p, T_S^n)$ ↓			Avg $\delta_{CD}^S(T_S^p, T_S^n)$ ↓			Avg $\delta_F^A(T_S^p, T_S^n)$ ↑			Avg $\delta_F^S(T_S^p, T_S^n)$ ↑		
On the sketches from the *FVRS* dataset [23].																			
Model		t-1	t-5	t-10	t-1	t-5	t-10	t-1	t-5	t-10	t-1	t-5	t-10	t-1	t-5	t-10	t-1	t-5	t-10
δ_{CD}^A TL		24.55	43.86	52.67	3.23	4.16	4.49	1.10	1.40	1.50	1.08	1.39	1.48	86.96	83.26	82.09	87.27	83.43	82.23
δ_{CD}^A RL		15.35	31.98	39.11	3.62	4.12	4.28	1.22	1.35	1.41	1.23	1.38	1.44	84.85	83.42	82.81	84.88	83.31	82.68
On the sketches from the *FVRS-M* dataset.																			
Model		t-1	t-5	t-10	t-1	t-5	t-10	t-1	t-5	t-10	t-1	t-5	t-10	t-1	t-5	t-10	t-1	t-5	t-10
δ_{CD}^A TL		17.43	35.35	43.66	3.81	4.58	4.87	1.24	1.46	1.55	1.26	1.49	1.59	85.07	82.54	81.45	85.19	82.53	81.41
δ_{CD}^A RL		9.01	21.88	28.32	4.35	4.57	4.70	1.36	1.43	1.46	1.40	1.49	1.53	83.01	82.32	81.99	82.99	82.09	81.71

top-k accuracy and all distance metrics.

5.3.2 Comparison with regression loss

For 3D shape to 3D shape retrieval, Uy et al. [32] proposed a regression loss, to retrieve a 3D shape best matching a query 3D model after deformation. Here, we adopt this loss to our problem, and use our definition of the asymmetric fitting gap as the distance between 3D shapes, and the L_2 distance as the distance in the feature space. Please see the supplemental for details.

Tab. 3 shows that, in certain cases, Regression Loss (RL) results in slightly better criteria, measuring the average similarity of the top-5/10 retrieval results to the ground-truth. However, RL results in significantly worse retrieval performance according to top-k retrieval accuracy (Acc) and the criteria evaluating the average similarity of the top-1 retrieval results. For additional evaluation and details please refer to the supplemental.

RL encourages the distribution of distances in the latent space to approximate the distribution of distances between 3D shapes. Thus, many similar shapes are encoded close to each other in the latent space. Meanwhile, triplet loss's primary goal is to ensure that the distance between similar shapes is smaller than the distance between dissimilar

shapes. Our formulation with the adaptive margin allows to achieve a desired trade off (Fig. 7).

6. Conclusion and future work

In this work, we propose an *adaptive margin setting* allowing for *structural similarities of shapes* to play an important role in ranking retrieval results. We further show that, due to inaccuracies inherent to sketches, top-k retrieval accuracy, the standard for retrieval evaluation, might be insufficient to drive out conclusions on retrieval performances. We deploy a recent off-the-shelf neural deformation operator to compute the fitting gap that measures the similarity between two 3D shapes under deformation. We conduct a user study that shows the superiority of this fitting gap over the Chamfer distance in measuring structural similarities. We deploy fitting gap as one of the criteria to measure retrieval results and as a distance metric for the adaptive margin scheme. We show that our results outperform state-of-the-art retrieval methods for sketch-based 3D shape retrieval. Our perceptual studies show that both symmetric and asymmetric fitting gap can lead to sub-optimal rankings. In the future, we would like to improve the robustness of the fitting gap driven retrieval by considering not only the distances after the deformation, but also how much the shapes need to be deformed.

References

[1] Harry G Barrow, Jay M Tenenbaum, Robert C Bolles, and Helen C Wolf. Parametric correspondence and chamfer matching: Two new techniques for image matching. Technical report, 1977.

[2] Wenming Cao, Zhiyue Yan, Zhiquan He, and Zhihai He. A comprehensive survey on geometric deep learning. *IEEE Access*, 2020.

[3] Nenglun Chen, Lingjie Liu, Zhiming Cui, Runnan Chen, Duygu Ceylan, Changhe Tu, and Wenping Wang. Unsupervised learning of intrinsic structural representation points. In *CVPR*, 2020.

[4] John Collomosse, Tu Bui, Michael J Wilber, Chen Fang, and Hailin Jin. Sketching with style: Visual search with sketches and aesthetic context. In *ICCV*, 2017.

[5] Guoxian Dai, Jin Xie, Fan Zhu, and Yi Fang. Deep correlated metric learning for sketch-based 3D shape retrieval. In *AAAI*, 2017.

[6] Mathias Eitz, Ronald Richter, Tamy Boubekeur, Kristian Hildebrand, and Marc Alexa. Sketch-based shape retrieval. *ACM Trans. Graph.*, 2012.

[7] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *CVPR*, 2017.

[8] Matheus Gadelha, Giorgio Gori, Duygu Ceylan, Radomír Měch, Nathan Carr, Tamy Boubekeur, Rui Wang, and Subhransu Maji. Learning generative models of shape handles. In *CVPR*, 2020.

[9] Daniele Giunchi, Stuart James, and Anthony Steed. 3D sketching for interactive model retrieval in virtual reality. In *Expressive*, 2018.

[10] Daniele Giunchi, Alejandro Sztrajman, Stuart James, and Anthony Steed. Mixing modalities of 3d sketching and speech for interactive model retrieval in virtual reality. *CoRR*, 2021.

[11] Alexander Grabner, Peter M Roth, and Vincent Lepetit. 3D pose estimation and 3D model retrieval for objects in the wild. In *CVPR*, 2018.

[12] Mai Lan Ha and Volker Blanz. Deep ranking with adaptive margin triplet loss. *arXiv preprint arXiv:2107.06187*, 2021.

[13] Xinwei He, Yang Zhou, Zhichao Zhou, Song Bai, and Xiang Bai. Triplet-center loss for multi-view 3D object retrieval. In *CVPR*, 2018.

[14] Tomas Jakab, Richard Tucker, Ameesh Makadia, Jiajun Wu, Noah Snavely, and Angjoo Kanazawa. Keypointdeformer: Unsupervised 3d keypoint discovery for shape control. In *CVPR*, 2021.

[15] Bo Li, Yijuan Lu, Fuqing Duan, Shuilong Dong, Yachun Fan, Lu Qian, Hamid Laga, Haisheng Li, Yuxiang Li, Peng Liu, et al. SHREC'16 Track: 3D Sketch-Based 3D Shape Retrieval. In *Proceedings of the Eurographics 2016 Workshop on 3D Object Retrieval*, 2016.

[16] Bo Li, Yijuan Lu, Azeem Ghumman, Bradley Strylowski, Mario Gutierrez, Safiyah Sadiq, Scott Forster, Natacha Feola, and Travis Bugerin. 3D sketch-based 3D model retrieval. In *ICMR*, 2015.

[17] Bo Li, Yijuan Lu, Azeem Ghumman, Bradley Strylowski, Mario Gutierrez, Safiyah Sadiq, Scott Forster, Natacha Feola, and Travis Bugerin. Kinectsbr: A kinect-assisted 3d sketch-based 3d model retrieval system. In *ICMR*, 2015.

[18] Bo Li, Yijuan Lu, Chunyuan Li, Afzal Godil, Tobias Schreck, Masaki Aono, Martin Burtscher, Hongbo Fu, Takahiko Furuya, Henry Johan, et al. Shrec'14 track: Extended large scale sketch-based 3d shape retrieval. In *Eurographics workshop on 3D object retrieval*, 2014.

[19] Lei Li, Zhe Huang, Changqing Zou, Chiew-Lan Tai, Rynson WH Lau, Hao Zhang, Ping Tan, and Hongbo Fu. Model-driven sketch reconstruction with structure-oriented retrieval. In *SIGGRAPH ASIA 2016 Technical Briefs*. ACM, 2016.

[20] Shuang Liang, Weidong Dai, and Yichen Wei. Uncertainty learning for noise resistant sketch-based 3d shape retrieval. *IEEE TIP*, 30, 2021.

[21] Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *CVPR*, 2017.

[22] Ling Luo, Yulia Gryaditskaya, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Towards 3d vr-sketch to 3d shape retrieval. In *3DV*, 2020.

[23] Ling Luo, Yulia Gryaditskaya, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Fine-grained vr sketching: Dataset and insights. In *3DV*, 2021.

[24] Anran Qi, Yulia Gryaditskaya, Jifei Song, Yongxin Yang, Yonggang Qi, Timothy M. Hospedales, Tao Xiang, and Yi-Zhe Song. Towards fine-grained sketch-based 3d shape retrieval. *IEEE TIP*, 2021.

[25] A. Qi, Y. Song, and T. Xiang. Semantic embedding for sketch-based 3D shape retrieval. In *BMVC*, 2018.

[26] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *NeurIPS*, 30, 2017.

[27] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Trans. Graph.*, 2016.

[28] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

[29] Jifei Song, Qian Yu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *ICCV*, 2017.

[30] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In *CVPR*, 2019.

[31] Yi Tian, Zhiwei Wen, Weicheng Xie, Xi Zhang, Linlin Shen, and Jinming Duan. Outlier-suppressed triplet loss with adaptive class-aware margins for facial expression recognition. In *ICIP*, 2019.

[32] Mikaela Angelina Uy, Jingwei Huang, Minhyuk Sung, Tolga Birdal, and Leonidas Guibas. Deformation-aware 3d model embedding and retrieval. In *ECCV*, 2020.

[33] Mikaela Angelina Uy, Vladimir G Kim, Minhyuk Sung, Noam Aigerman, Siddhartha Chaudhuri, and Leonidas J

Guibas. Joint learning of 3d shape retrieval and deformation. In *CVPR*, 2021.

[34] Fang Wang, Le Kang, and Yi Li. Sketch-based 3d shape retrieval using convolutional neural networks. In *CVPR*, 2015.

[35] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *CVPR*, 2014.

[36] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, 2016.

[37] Jin Xie, Guoxian Dai, Fan Zhu, and Yi Fang. Learning barycentric representations of 3d shapes for sketch-based 3d shape retrieval. In *CVPR*, 2017.

[38] Yuxiang Ye, Bo Li, and Yijuan Lu. 3D sketch-based 3D model retrieval with convolutional neural network. In *ICPR*, 2016.

[39] Wang Yifan, Noam Aigerman, Vladimir G Kim, Siddhartha Chaudhuri, and Olga Sorkine-Hornung. Neural cages for detail-preserving 3d deformations. In *CVPR*, 2020.

[40] Sang Min Yoon, Maximilian Scherer, Tobias Schreck, and Arjan Kuijper. Sketch-based 3D model retrieval using diffusion tensor fields of suggestive contours. In *IEEE Trans. Multimedia*, 2010.

[41] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales, and Chen Change Loy. Sketch me that shoe. In *CVPR*, 2016.

[42] Qian Yu, Jifei Song, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Fine-grained instance-level sketch-based image retrieval. *IJCV*, 2021.

[43] Sergey Zakharov, Wadim Kehl, Benjamin Planche, Andreas Hutter, and Slobodan Ilic. 3d object instance recognition and pose estimation using triplet loss with dynamic margin. In *IROS*. IEEE, 2017.

[44] Xiaonan Zhao, Huan Qi, Rui Luo, and Larry Davis. A weakly supervised adaptive triplet loss for deep metric learning. In *IEEE/CVF International Conference on Computer Vision Workshops*, 2019.

[45] Fan Zhu, Jin Xie, and Yi Fang. Learning cross-domain neural networks for sketch-based 3d shape retrieval. In *AAAI*, 2016.