This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: https://orca.cardiff.ac.uk/id/eprint/114997/

This is the author’s version of a work that was submitted to / accepted for publication.

Citation for final published version:

Publishers page: https://doi.org/10.1145/3272127.3275028

Please note:
Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher’s version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See http://orca.cf.ac.uk/policies.html for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.
Automatic Unpaired Shape Deformation Transfer

LIN GAO, Institute of Computing Technology, Chinese Academy of Sciences
JIE YANG, Institute of Computing Technology, Chinese Academy of Sciences and University of Chinese Academy of Sciences
YI-LING QIAO, Institute of Computing Technology, CAS and University of Chinese Academy of Sciences
YU-KUN LAI, Cardiff University
PAUL L. ROSIN, Cardiff University
WEIWEI XU, Zhejiang University
SHIHONG XIA, Institute of Computing Technology, Chinese Academy of Sciences

Fig. 1. Deformation transfer from a fit person to a fat person, both from the MPI DYNA dataset [Pons-Moll et al. 2015]. First row: source fit person shapes, second row: our results of deformation transfer to a fat person. Our method automatically transfers rich actions across shapes with substantial geometric differences without the need for specifying correspondences or shape pairs between source and target.

Transferring deformation from a source shape to a target shape is a very useful technique in computer graphics. State-of-the-art deformation transfer methods require either point-wise correspondences between source and target shapes, or pairs of deformed source and target shapes with corresponding deformations. However, in most cases, such correspondences are not available and cannot be reliably established using an automatic algorithm. Therefore, substantial user effort is needed to label the correspondences or to obtain and specify such shape sets. In this work, we propose a novel approach to automatic deformation transfer between two unpaired shape sets without correspondences. 3D deformation is represented in a high-dimensional space. To obtain a more compact and effective representation, two convolutional variational autoencoders are learned to encode source and target shapes to their latent spaces. We exploit a Generative Adversarial Network (GAN) to map deformed source shapes to deformed target shapes, both in the latent spaces, which ensures the obtained shapes from the mapping are indistinguishable from the target shapes. This is still an under-constrained problem, so we further utilize a reverse mapping from target shapes to source shapes and incorporate cycle consistency loss, i.e. applying both mappings should reverse to the input shape. This VAE-Cycle GAN (VC-GAN) architecture is used to build a reliable mapping between shape spaces. Finally, a similarity constraint is employed to ensure the mapping is consistent with visual similarity, achieved by learning a similarity neural network that takes the embedding vectors from the source and target latent spaces and predicts the light field distance between the corresponding shapes. Experimental results show that our fully automatic method is able to obtain high-quality deformation transfer results with unpaired data sets, comparable or better than existing methods where strict correspondences are required.

CCS Concepts: • Computing methodologies → Shape modeling; Animation;

Additional Key Words and Phrases: Deformation transfer, generative adversarial network, cycle consistency, visual similarity
1 INTRODUCTION

Shape deformation is widely used in computer graphics. It is useful in geometric modeling for generating new shapes from existing ones and in computer animation for producing smooth animation sequences. However, producing realistic animation is time-consuming and requires artistic expertise. Deformation transfer, i.e. transferring deformation of one shape to another, provides a cost-effective solution to producing new deformation results by reusing existing ones.

Given two sets of deformable shapes (source and target) and a new deformed source shape, deformation transfer aims to produce realistic deformation of the target shape, visually corresponding to the given deformed source shape. It is a challenging task, since the source and target shapes can differ significantly. Existing state-of-the-art methods [Chu and Lin 2010; Sumner and Popović 2004] for surface-based deformation transfer rely on point-wise correspondences between source and target shapes. In general cases, there is no reliable automatic method to achieve this, so existing methods require the user to specify a sufficient number of corresponding points such that point-wise correspondences can be deduced. This process is tedious and often requires trial-and-error to ensure specified corresponding points provide sufficient constraints. An alternative approach considers semantic deformation transfer [Baran et al. 2009]. The method does not require point-wise correspondence between source and target shapes. However, it takes paired source and target shapes, assuming that each model in the source set is semantically related to the corresponding shape in the target set. In practice, however, if the source and target shape datasets are constructed independently, this property is unlikely to be satisfied.

In this work, we aim to develop a fully automatic algorithm to deform target shapes in a way as similar as possible to the source deformed shapes, which none of the existing deformation transfer methods can achieve. To make this seemingly impossible task a reality, we exploit the learning capability of deep neural networks to learn how shapes deform naturally from a given dataset, providing a differentiable metric measuring visual similarity and build reliable mapping between the latent spaces with cycle-consistency. This is inspired by how humans perform this task, by observing the deformed shapes to learn their characteristics, considering the similarity between source and target shapes, and thinking about how the target shapes should deform to resemble source shapes. An example of our method is shown in Fig. 1 where the deformation of a fit person is automatically transferred to that of a fat person, with substantial body shape differences. Unlike previous methods, we do not require point-wise correspondences between source and target shapes, or paired source and target shapes as input. Instead, source and target shape sets may contain different deformations, as long as they are both sufficient to cover the relevant deformation spaces. This greatly reduces user efforts and allows using two independent shape deformation datasets. To achieve this, we propose a cycle-consistent Generative Adversarial Network (GAN) architecture for mesh deformation transfer. To ensure learning efficiency with a relatively small number of training examples, especially because plausible deformations form a much lower dimensional manifold in the high-dimensional deformation space, we introduce a convolutional autoencoder to represent shape deformations in a compact latent space. To ensure effective transfer, we further propose a neural network to measure visual similarity. The main contributions of this work are summarized as follows:

- This is the first automatic work to transfer deformation between unpaired shape datasets. To achieve this, we present an efficient and differentiable method, which is composed of a variational autoencoder to encode shape deformations, a differentiable network to measure visual similarity, and a cycle-consistent GAN for reliable mapping between latent spaces.
- We propose a novel neural network to measure the visual similarity between deformed shape pairs, which is differentiable and efficiently approximates light field distances. This network is the key to make the whole approach differentiable and trainable.
- We also propose a novel mesh-based convolutional variational autoencoder (VAE) to encode a shape set with flexible deformations in a compact latent space, which copes well with large deformations, supports generating new shapes in the space, and has good generalizability.

All the network components work together tightly and each is indispensable. The mesh-based convolutional autoencoder is used to learn and describe the plausible deformation space for generating natural shapes; the similarity metric provides a differentiable approximation to the Light Field Distance (LFD) [Chen et al. 2003], enabling our whole deep learning architecture to perceive intricate visual similarities across 3D model domains. Our method also benefits from the cycle consistency applied to the GAN network [Zhu et al. 2017] to build reliable mapping between two spaces. We further consider a simple extension of our method to semantic deformation transfer by utilizing a small number of paired shapes for learning a semantic similarity metric which cannot be characterized by visual similarity.

In Sec. 2, we review the work most related to ours. We then give the detailed description of our method, including overall architecture and loss functions in Sec. 3. Implementation details are presented in Sec. 4. We present experimental results, including extensive comparisons with state-of-the-art methods in Sec. 5, and finally, we draw conclusions in Sec. 6.

2 RELATED WORK

Shape deformation is an active research topic. A comprehensive survey of relevant techniques can be found in [Gain and Bechmann 2008]. We now review techniques for deformation transfer and deep learning which are related to this work.

Mesh Deformation Transfer. Sumner et al. [2004] performed pioneering work for mesh deformation transfer. The method requires point-wise correspondences between source and target shapes. Local

ACM Transactions on Graphics, Vol. 0, No. 0, Article 0 (2018), 15 pages.
https://doi.org/10.1145/3272127.3275028
shape deformation is then represented using deformation gradients, which are transferred from the source to the target shapes. This method however relies on specifying typically a large number of correspondences to cope with the differences between shapes since deformation gradients are local. Moreover, the method may transfer geometric details from the source shape to the target, which is undesirable. To address this, Chu and Lin [2010] proposed a method that further projected the deformed shape to the manifold of the target shape space, under the assumption that the target shape set provides sufficient coverage of plausible deformations. To reduce user effort, Yang et al. [2018] developed a method to automatically choose a set of suitable key points on the source shape, although the corresponding points on the target shape still require to be manually specified. Instead of specifying point-wise correspondences, Baran et al. [2009] proposed a different approach where the input is a set of source shapes and the same number of target shapes, where the corresponding pairs of source and target shapes have related semantic meaning. Their method achieves semantic deformation transfer by representing each deformed shape in the source sequence using a combination of given shapes in the source set, and producing the target deformed shape by utilizing the combination weights with shapes in the target set. The method produces interesting results, although the required input is not generally available if the two shape sets are obtained independently, which restricts its use. Our method is fundamentally different from these works: by utilizing the learning capability of a novel GAN-based architecture we are able to take unpaired source and target shape sets and do not require point-wise or shape-wise correspondences between the sets.

For shapes which are not manifold triangle meshes, e.g. triangle soups or tetrahedra, methods have been developed [Ben-Chen et al. 2009; Chen et al. 2010] for transferring deformation using cages that enclose the shapes to be transferred. However, effort is needed to construct cages, and such methods may erroneously deform spatially adjacent regions if they happen to fit in the same cage. To cope with shapes involving multiple connected components, Zhou et al. [2010] developed a method based on a graph structure. These methods similarly require input for correspondences. Recent work [Corman et al. 2017; Rustamov et al. 2013] develops effective approaches to measuring shape differences, which are used for embedding shape collections. Given a shape in the first collection, these methods can be used to find a similar shape in the second collection without known correspondence. However, such methods do not synthesize new shapes, and therefore may not always be able to find suitable corresponding shapes.

Deep Learning for 3D Shape Processing. We exploit the learning capability of neural networks in this work, which have achieved great success in 2D image processing. In recent years, effort has been made to process 3D shapes, which are more challenging due to the higher dimension and irregular connectivity.

For 3D shape recognition and analysis, shapes can be represented using multi-view projection images along with 2D-CNNs for classification [Shi et al. 2015]. Such approaches are used in [Huang et al. 2018] to learn local shape descriptors useful for shape correspondence and segmentation. Shapes can also be represented using voxels with 3D-CNNs extended from 2D [Maturana and Scherer 2015]. Tulsiani et al. [2017] use this representation to approximate shapes with cuboids, giving an abstract representation. To improve efficiency, Wang et al. [2017a] propose an octree structure to represent 3D shapes and perform convolutional operations on the octree to build CNNs. Alternatively, meshes can be treated as irregular graphs, and CNNs are extended to handle such graphs either in the spectral [Bruna et al. 2013; Defferrard et al. 2016; Henaff et al. 2015] or the spatial domain [Duvenaud et al. 2015; Niepert et al. 2016]. These representations are used for shape correspondences [Boscaini et al. 2016a,b] and shape segmentation [Yi et al. 2017a]. Maron et al. [2017] parameterize a sphere-type shape to a planar flat-torus with a well-defined convolutional operator to build CNN models.

For 3D shape synthesis, methods have been developed using voxel-based 3D CNNs, including deep belief networks [Wu et al. 2015] and GANs [Wu et al. 2016]. The latter is pioneering work that uses a GAN to generate 3D shapes. However, it uses a voxel representation with limited resolution, whereas our method aims to automatically transfer mesh deformations with rich geometry details. Moreover, [Wu et al. 2016] needs paired images and 3D models, while our method is fully unsupervised. Liu et al. [2017] extend a 3D GAN to support interactive editing, where a projection operator is provided to map user designed voxels to more detailed shapes. Sharma et al. [2016] use an unsupervised voxel-based autoencoder for applications such as denoising. The voxel-based representation has high space complexity due to its cubic nature, and therefore can only be practically used to synthesize coarse shapes. An alternative approach uses geometry images and an image-based ResNet architecture to synthesize 3D shapes [Sinha et al. 2017]. Geometry images allow details to be better preserved, but they also have unavoidable distortions and seams, and require shapes to be aligned to facilitate processing. Taking aligned meshes with consistent segmentation as input, neural networks are also used to synthesize 3D shapes with pre-segmented parts [Li et al. 2017; Nash and Williams 2017]. These methods focus on man-made objects and do not support non-rigid deformation.

Some works address relevant but different tasks from synthesizing general 3D shapes. Han et al. [2017] use a CNN to model 3D faces with sketches. In [Sung et al. 2017], neural networks are employed for assembly-based modeling with suggestions of complementary components and their placement. Other research considers joint embeddings of 2D images and 3D shapes [Li et al. 2015], and maps 2D images to corresponding 3D shapes [Choy et al. 2016; Fan et al. 2017; Girdhar et al. 2016; Yan et al. 2016].

None of these works address the problem we study here, namely deformation transfer.

Deep Learning for Non-Rigid Deformation. Our method deals with deformable mesh datasets. To analyze them, Tan et al. [2018a] firstly proposed a mesh variational autoencoder network with fully connected layers to encode meshes using a recent rotation-invariant representation [Guo et al. 2016], with applications to shape embedding and synthesis. However, the use of fully connected layers restricts its generalizability. An alternative mesh-based convolutional autoencoder was proposed in [Litany et al. 2017] for completion of deformable shapes. Their method however takes Euclidean coordinates directly and thus cannot handle large rotations well. The work
[Tan et al. 2018b] proposed a convolutional autoencoder to encode deformable mesh sets with sparse constraints to extract localized deformation components. Their architecture is not variational and thus not suitable for generating new data. Our method proposes a new convolutional variational autoencoder with a representation that handles large rotations, which effectively embeds deformable shapes to a compact latent space.

Image Transfer using GAN. Synthesizing new images by transferring information from existing ones has been an active topic for decades. Recent work uses the GAN-based architecture, where the joint training of a generator and a discriminator helps ensure that the synthesized images have characteristics indistinguishable from the target sets. This is related to our work, although we deal with deformation transfer between shape sets. The work pix2pix [Isola et al. 2017] uses a conditional GAN for paired image-to-image translation, without the requirement of tuning loss functions. DualGAN [Yi et al. 2017b] uses two identical GANs to achieve image-to-image translation in an unsupervised manner. The method CycleGAN [Zhu et al. 2017] allows image transfer with unpaired training examples, where the key idea is to introduce a cycle consistency loss to regularize the mapping. We instead perform deformation synthesis in the space of 3D shapes utilizing a GAN, since GANs are proved to be capable for such tasks [Wang et al. 2017b]. Unlike existing CycleGAN works that focus on images, we propose the first work for automatic mesh deformation transfer with unpaired source and target shapes, by generalizing the CycleGAN architecture to deal with challenging 3D shapes.

3 METHOD
Given two sets of unpaired shapes, a source shape set \( S \) and a target shape set \( T \), as well as a deformed source shape \( s \), our aim is to produce a deformed target shape \( t \) which has visually similar deformation as \( s \). Shapes in the same set \( S \) or \( T \) have the same connectivity. Many shape datasets exist: they are either obtained by deforming a mesh model, or fitting a template mesh model to deformed shapes.

We do not assume shapes in \( S \) correspond to specific shapes in \( T \), although we assume that \( S \) and \( T \) provide sufficient coverage of typical deformations of the relevant shapes. We learn a deep model with \( S \) and \( T \) as training examples. Once the deep model is trained, a shape \( t \) is generated for each input \( s \).

3.1 Overview
The overall architecture of the VAE-Cycle GAN (VC-GAN) is illustrated in Fig. 2. Since mesh models typically contain thousands of vertices and mesh datasets may have less than a hundred shapes, establishing a mapping from source shapes \( S \) to target shapes \( T \) can be underdetermined. Therefore, before setting up the mapping functions, we employ variational autoencoders to encode the source and target shape sets into compact latent spaces \( \hat{S} \) and \( \hat{T} \). This not only makes the training process better constrained, but also ensures generated shapes conform to the deformation space. Denote by \( Enc_S \) and \( Dec_S \) the encoder and decoder for the source shape set. \( Enc_T \) and \( Dec_T \) are similarly defined for the target shape set.

See Sec. 3.2 for details of our convolutional variational autoencoder (VAE).

Our mapping functions \( G: \hat{S} \rightarrow \hat{T} \) and \( F: \hat{T} \rightarrow \hat{S} \) are defined between the latent spaces. To ensure the mapping is meaningful, three types of regularization terms are employed including adversarial loss, cycle-consistency loss and visual similarity loss. Measuring the visual similarity between shapes with different topologies is not trivial and we introduce a dedicated neural network to predict this.

The first loss is the adversarial loss (Sec. 3.3) which discriminates the generated shape \( G(\cdot) \) from the target shape sets. In the joint training process, it ensures the generated shape belongs to the target space. As we will show later, since the latent space from the
convolutional VAE $\mathcal{V}$ has a Gaussian distribution, the discriminator cannot be effectively defined in the latent space to differentiate genuine and synthesized shapes. Instead, we define the discriminator in the decoded space. Let $D_S$ and $D_T$ be the discriminators in the source $\mathcal{S}$ and target $\mathcal{T}$ shape spaces. $D_T$ and $D_S$ take $\text{Dec}_T(G(s))$ and $\text{Dec}_S(F(t))$ as input respectively, where $s$ and $t$ are source and target shapes encoded in the latent spaces; see Fig. 3. The discriminator is a mesh-based convolutional neural network whose task is to discriminate between the transferred meshes (fake) and existing meshes in the target dataset (real).

The second loss is the cycle-consistency loss (Sec. 3.4), which ensures that the mapping from one domain to the other domain followed by the reverse mapping returns to the starting point. The cycle-consistency loss includes the forward cycle-consistency loss: $s \rightarrow G(s) \rightarrow F(G(s)) \approx s$ and backward cycle-consistency loss: $t \rightarrow F(t) \rightarrow G(F(t)) \approx t$; see Fig. 3.

For the deformation transfer task, the transferred target shape should be similar to the source shape being transferred. The visual similarity loss is employed to measure the visual similarity between the source shape and the transferred target shape. The forward and backward visual similarity losses are defined as $V(G(s), t)$ and $V(F(t), s)$, where $V(\cdot)$ is a fully connected neural network to measure the visual similarity between two shapes represented in the latent spaces; see Sec. 3.5 for more details.

### 3.2 Convolutional Variational Autoencoder

To cope with large deformations, we represent each deformed shape using a recently proposed ACAP (as consistent as possible) shape deformation representation [Gao et al. 2017], which handles large rotations well, and has features defined only on vertices, making convolutional operators easier to define. This is different from alternative representations such as [Gao et al. 2016] where features are also associated with edges. Take $\mathcal{S}$ for example, and the same process is applied to $\mathcal{T}$. Assume $\mathcal{S}$ contains $N$ shapes with the same connectivity, each denoted as $S_m$, $p_{m,i} \in \mathbb{R}^3$ is the $i^{th}$ vertex on the $m^{th}$ mesh. The deformation gradient $T_{m,i} \in \mathbb{R}^{3 \times 3}$ representing local shape deformations can be obtained by minimizing:

$$\arg \min_{T_{m,i}} \sum_{j \in N_i} c_{ij} \| (p_{m,i} - p_{m,j}) - T_{m,i}(p_{i,j} - p_{i,j}) \|^2_2,$$

where $c_{ij}$ is the cotangent weight and $N_i$ represents the 1-ring neighbors of the $i^{th}$ vertex. $T_{m,i}$ can be decomposed into $R_{m,i}S_{m,i}$ where $R_{m,i}$ is the rotation and $S_{m,i}$ is the scale/shear deformation. The rotation matrix $R_{m,i}$ can be represented by the rotation axis $\omega_{m,i}$ and the associated rotation angle $\theta_{m,i}$. The mapping from the rotation matrix to rotation axis and angle is one to many, and the possible solutions are in the set $\Omega_{m,i}$:

$$\Omega_{m,i} = \{ (\omega_{m,i}, \theta_{m,i} + t \cdot 2\pi), (-\omega_{m,i}, -\theta_{m,i} + t \cdot 2\pi) \}$$

(1)

For shapes with large-scale rotations, the adjacent vertices may have inconsistent rotation angles and rotation axes, which will lead to artifacts during shape blending and synthesis, as shown in [Gao et al. 2017]. To solve this problem, [Gao et al. 2017] proposes a method based on global integer programming to resolve rotation ambiguities and ensure consistency. Two global integer programming optimizations are used to solve for rotation axes and angles to make them as-consistent-as possible. For each vertex, this gives a feature vector $q_{m,i} \in \mathbb{R}^9$. The components of $q_{m,i}$ come from the combination of the rotation matrix $R_{m,i}$ and scale/shear matrix $S_{m,i}$ of the vertex. The logarithm of $R_{m,i}$ is a skew-symmetric matrix and $S_{m,i}$ is a symmetry matrix, so we extract a 3-dimensional vector for $R_{m,i}$ and a 6-dimensional vector for $S_{m,i}$ using non-duplicated entries and concatenate them to a 9-dimensional vector. As shown in [Tan et al. 2018b], this ACAP representation is appropriate for mesh-based convolutional neural networks. Following [Tan et al. 2018b], we rescale each dimension of $q_{m,i}$ independently to $[-0.95, 0.95]$ before feeding into the convolutional VAE, and scale it back from the output of the convolutional VAE, such that the $\text{tanh}$ activation function can be applied.

The overall architecture of our convolutional VAE is illustrated in Fig. 4. For meshes with $v$ vertices, the input $s$ to the VAE is $9 \times v$ dimensional. Unlike [Tan et al. 2018a] which uses fully connected layers, we use convolutional layers which have better generalizability. As illustrated in Fig. 5, we use a mesh-based convolution operator [Duvenaud et al. 2015; Tan et al. 2018b] where the output at a vertex is obtained as a linear combination of input in its 1-ring neighbors along with a bias. The output of the operator $y_i$ for the $i^{th}$ vertex is defined as follows:

$$y_i = W_{\text{point}}s_i + W_{\text{neighbor}} \sum_{j=1}^{D_i} s_{n_{ij}} + b,$$

(2)

where $s_i$ is the input feature vector for the $i^{th}$ vertex, $D_i$ is the degree of the $i^{th}$ vertex, $n_{ij} (1 \leq j \leq D_i)$ is the $j^{th}$ neighbor of the $i^{th}$ vertex. $W_{\text{point}}, W_{\text{neighbor}} \in \mathbb{R}^{9 \times 9}$ and $b \in \mathbb{R}^9$ are the weights and bias. Following a convolutional neural network, these weights and the bias are shared by all the neighborhoods within each convolutional layer and learned during training. Since shapes in the same dataset contain complex deformations, their intrinsic and extrinsic geometries can change substantially. Therefore, uniform weights as given in the definition above, whereby neighboring vertices contribute equally in the convolution operator, are beneficial and used in our experiments as they only depend on the topology. As shown in Table 1, using uniform weights gives better performance than using cotangent weights [Meyer et al. 2003] commonly used in mesh processing.

After connecting several convolutional layers with the same size, these nodes are further connected with a fully connected layer. The architecture of this convolutional variational autoencoder is shown in Fig. 4. The work [Tan et al. 2018b] is not a variational

### Table 1. Mean RMS (root mean square) reconstruction errors of applying our method to generate unseen data using uniform or cotangent [Meyer et al. 2003] weight matrices for the convolutional operator on the ball dataset [Rustamov et al. 2013] (see Fig. 11(a)). We randomly choose 75% of the dataset as the training set and the remaining 25% of the dataset as the test set.

<table>
<thead>
<tr>
<th>Method</th>
<th>Uniform Weight</th>
<th>Cotangent Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Balls</td>
<td>0.0059</td>
<td>0.0080</td>
</tr>
</tbody>
</table>

ACM Transactions on Graphics, Vol. 0, No. 0, Article 0. Publication date: 2018.
variable with a Gaussian distribution with \( \text{mean and unit variance.} \)
The encoder encodes shapes from shape set \( S \) in a latent space \( \tilde{S} \), and the decoder recovers the shape \( S' \).

Let \( Enc(\cdot) \) and \( Dec(\cdot) \) be the encoder and decoder of our VAE network. \( s \) represents the input shape from dataset \( S \), \( \tilde{s} = Enc(s) \) is the encoded latent vector and \( s' = Dec(\tilde{s}) \) is the reconstructed shape. Our convolutional VAE minimizes the following loss:

\[
L_{VAE} = L_{\text{recon}} + \alpha_1 L_{KL} + \alpha_2 L_{\text{RegVAE}}
\]

where \( \alpha_1 \) and \( \alpha_2 \) are relative weights of different loss terms, \( L_{\text{recon}} = \frac{1}{|S|} \sum_{s \in S} ||s - s'||_2^2 \) denotes the MSE (mean square error) reconstruction loss to ensure faithful reconstruction, \( L_{KL} = D_{KL}(q(\tilde{s}|\tilde{s})||p(\tilde{s})) \) is the KL divergence to promote Gaussian distribution in the latent space, where \( q(\tilde{s}|\tilde{s}) \) is the posterior distribution given input shape \( s \), and \( p(s) \) is the Gaussian prior distribution. \( L_{\text{RegVAE}} \) is the squared \( \ell_2 \) norm regularization term of the network parameters used to avoid overfitting. The Gaussian distribution makes it effective to generate new shapes by sampling in the latent space, which is used for training the GAN model, as we will explain later. The minimization of the above loss \( L_{VAE} \) is performed by gradient descent using the ADAM (adaptive moment estimation) solver. The parameters and types of layers of this convolutional variational autoencoder are evaluated in the supplementary material.

### 3.3 Adversarial Loss

The adversarial losses are applied to both mapping functions \( G \) and \( F \) between the latent spaces. For the mapping function \( G : \tilde{S} \rightarrow \tilde{T} \), it is defined with the discriminator neural network \( D_{\tilde{T}} \), as follows:

\[
L_{GAN-G}(G, D_{\tilde{T}}, \tilde{S}, \tilde{T}) = \mathbb{E}_{\tilde{s} \sim p_{\text{data}}(\tilde{s})} [\log D_{\tilde{T}}(\tilde{s})] + \mathbb{E}_{\tilde{s} \sim p_{\text{data}}(\tilde{s})} [\log (1 - D_{\tilde{T}}(Dec_F(G(\tilde{s}))))],
\]

where \( p_{\text{data}}(\cdot) \) represents the data distribution. \( \mathbb{E} \) is the expected value of the distribution. The mapping \( G \) is used to generate \( G(\tilde{S}) \) that has similar deformations in the latent space of \( \tilde{T} \). The discriminator \( D_{\tilde{T}} \) aims to distinguish the generated deformation shapes.
As illustrated in Fig. 3, for each shape feature \( \tilde{D} \), the adversary aims to maximize the objective function while the mapping function \( G \) aims to minimize it. This is equivalent to \( \min_G \max_{D_T} L_{GAN-G}(G, D_T, \tilde{S}, \tilde{T}) \). Similarly, the inverse mapping function \( F \) and its associated discriminator \( D_S \) are optimized by \( \min_F \max_{D_T} L_{GAN-F}(F, D_S, \tilde{F}, \tilde{S}) \). We define \( L_{GAN} \) to be the sum of both GAN losses:

\[
L_{GAN}(G, F, D_S, D_T) = L_{GAN-G} + L_{GAN-F}.
\]

### 3.4 Cycle Consistency Loss

Cycle-consistency loss is known to be effective to better constrain the network to produce more stable results and avoid over-fitting. As illustrated in Fig. 3, for each shape feature \( \tilde{s} \) in the latent space \( \tilde{S} \), applying \( G \) followed by \( F \) should bring it back to the original feature \( s \rightarrow G(\tilde{s}) \rightarrow F(G(\tilde{s})) = \tilde{s} \approx s \). Similarly for \( t \), it satisfies the following cycle consistency: \( t \rightarrow F(\tilde{t}) \rightarrow G(F(\tilde{t})) = \tilde{t} \approx t \). The cycle-consistency loss \( L_{Cycle} \) is defined as:

\[
L_{Cycle}(G, F) = \mathbb{E}_{s \sim P_{data}(\tilde{S})} \|F(G(\tilde{s})) - \tilde{s}\|_1 + \mathbb{E}_{t \sim P_{data}(\tilde{T})} \|G(F(\tilde{t})) - \tilde{t}\|_1.
\]

Using \( \ell_1 \) loss above gives near identical performance to an alternative negative log likelihood definition, so is used in our experiments.

As shown in Fig. 6, with the cycle consistency loss, given an input shape \( \tilde{s} \) in the latent space, the recovered shape \( s \) after applying both mappings \( G \) and \( F \) is visually identical to \( \tilde{s} \) after applying the decoder. This demonstrates the effectiveness of the cycle consistency loss.

### 3.5 Visual Similarity Loss

One aim of deformation transfer is to ensure visual similarity between the source and target shapes. In [Zhu et al. 2017] the visual similarity metric between two images can be measured by the squared \( \ell_2 \) norm in the image domain since images are naturally aligned, However, this similarity metric cannot be generalized to the 3D shape domain.

In this work, the light field distance (LFD) [Chen et al. 2003] is employed to measure the visual similarity due to its robustness and accuracy. A 3D shape is projected to multiple views and image features are calculated based on projected images. When calculating distances between two shapes, global rotation is further considered to minimize the image feature differences. LFD is known to be an effective feature for shape retrieval [Shilane et al. 2004].

However, LFD is not differentiable and cannot be used in the loss function. We propose to use a neural network SimNet to learn this similarity measure. Since latent space is more compact, SimNet takes two vectors \( \tilde{s} \) and \( \tilde{t} \) from the latent space of \( \tilde{S} \) and \( \tilde{T} \) respectively, and predicts the LFD between the decoded shapes. Since there is no spatial relationship in the latent space, a fully connected neural network is employed, as shown in Fig. 7. The output of the neural network is denoted as \( V(\tilde{s}, \tilde{t}) \approx LFD(Dec_S(\tilde{s}), Dec_T(\tilde{t})) \). The network is trained by minimizing the following loss:

\[
L_{SimNet}(V) = L_{Dist}(V) + \beta L_{RegSim},
\]

where \( \beta \) is the weight, \( L_{Dist} \) is the average of the absolute difference \( V(\tilde{s}, \tilde{t}) - LFD(Dec_S(\tilde{s}), Dec_T(\tilde{t})) \), and \( L_{RegSim} \) is the squared \( \ell_2 \) norm regularization of network parameters to avoid overfitting. Our architecture is not restricted to the light field distance, which could be replaced with other advanced shape features.

Using the visual similarity measure, the loss is defined as follows:

\[
L_{Sim}(G, F) = \mathbb{E}_{s \sim P_{data}(\tilde{S})} [V(s, G(s))] + \mathbb{E}_{t \sim P_{data}(\tilde{T})} [V(F(\tilde{t}), \tilde{t})].
\]

### 3.6 Overall Loss Function for Cycle GAN

The overall loss for the CycleGAN is:

\[
L_{CycleGAN}(G, F, D_S, D_T) = L_{Sim}(G, F) + \gamma_1 L_{Cycle}(G, F) + \gamma_2 L_{GAN}(G, F, D_S, D_T),
\]

which is a linear combination of visual similarity, cycle consistency and adversarial losses with \( \gamma_1 \) and \( \gamma_2 \) being relative weights.

Our CycleGAN network is optimized as follows:

\[
G^*, F^* = \arg\min_{G,F} \max_{D_S,D_T} L_{CycleGAN}(G, F, D_S, D_T),
\]

where generators \( G \) and \( F \) aim to minimize the total loss while the discriminators \( D_S \) and \( D_T \) aim to minimize the loss by identifying synthetic shapes.

### 4 IMPLEMENTATION

We now give details of our network architecture and training process.
4.1 Network Architecture

Our network consists of three components, convolutional VAE for encoding shapes in latent spaces, SimNet for calculating visual similarity between two shapes (from $S$ and $T$ respectively) both in the latent spaces, and CycleGAN for deformation transfer.

As illustrated in Fig. 4, our proposed VAE architecture uses convolutional layers for improved generalization capability. The encoder takes as input the features defined on vertices, followed by two convolutional layers with $tanh$ as the activation function. In the last convolutional layer we abandon the non-linear activation function, similar to [Tan et al. 2018b]. The output of the convolutional layer is then reshaped to a vector and mapped into a 128-dimensional latent space by a fully connected layer. The decoder has a symmetric architecture, sharing the weights and biases with the encoder. We train one VAE for the source shape set $S$ and one for the target shape set $T$.

For SimNet, its input includes latent vectors from both domains $S$ and $T$. Since dimensions of the latent space do not have spatial relationship, its hidden layers are fully connected with dimensions of 2048, 1024, 512 and 256 respectively, each of them having Leaky ReLU as the activation function, as illustrated in Fig. 7.

Our CycleGAN architecture is similar to [Zhu et al. 2017] although the generators map vectors in the latent spaces. Since the latent space does not have a clear spatial relationship, the generators $G$ and $F$ are fully connected networks with four hidden layers of 512, 1024, 2048, 1024 dimensions respectively, mapping the latent vector of one model to another (see Fig. 8). The discriminators are defined in the feature space (i.e. after applying the decoder), and aim to classify whether the shape is genuine or synthesized. Discriminators $D_S$ and $D_T$ have three convolutional layers and a fully connected layer, similar to the architecture of the encoder (see Fig. 9). More detailed analysis of network architecture, parameters and input characteristics (noise, topological changes etc.) is given in the supplementary material.

4.2 Training Details

In the experiments, we fix the weight parameters $\alpha_1 = 1$, $\alpha_2 = 0.01$, $\beta = 0.01$, $y_1 = 2$, $y_2 = 2$. We train the whole network in three steps (VAE, SimNet and CycleGAN). The Adam solver [Kingma and Ba 2014] is used for all three training steps. Here we train each network separately (i.e. VAE and SimNet, followed by CycleGAN) rather than in an end-to-end manner. Compared with end-to-end training, training the networks separately not only saves memory during training, but also results in smaller loss, especially the SimNet loss, as shown in Table 2, which is based on transferring between a fat person (ID: 50002) and a fit person (ID: 50009) from the MPI DYNA dataset.

The main reason is that starting training the SimNet and GAN before obtaining a well-trained VAE may lead to wrong optimization directions in early iterations, which ultimately results in optimization stuck at a poor local minimum. Moreover, the visual quality of the deformation transfer results is also worse, in terms of pose, than for separate training, as shown in the example in Fig. 10.

The VAE is trained with 5,000 iterations, SimNet with 12,000 iterations and CycleGAN with 7,000 iterations, by minimizing loss functions $L_{VAE}$ (Eq. 3), $L_{SimNet}$ (Eq. 7) and $L_{CycleGAN}$ (Eq. 9), respectively. For both VAE and CycleGAN, we set the batch size to 128 and learning rate to 0.001. For training SimNet, we set the batch size and learning rate to 512 and 0.001 respectively for the first 2,000 iterations and change them to 128 and 0.00005 for the remaining 10,000 iterations. Training batches for the VAE are randomly sampled from the training shape set. For SimNet training, we randomly select pairs of shapes from the two shape sets for training. For CycleGAN, training batches are sampled randomly from the latent space with a Gaussian distribution.

5 RESULTS AND EVALUATIONS

Our experiments were carried out on a computer with an i7 6850K CPU, 16GB RAM, and a Titan Xp GPU. The code is available at http://www.geometrylearning.com/ausdt/. The training time for each network is 30-45min for VAE (depending on the mesh size), 54min for SimNet and 65min for CycleGAN. Once the deep model is trained, transferring a source shape to a target shape is real-time.
We first analyze the effectiveness of our network components. We compare our convolutional VAE with a state-of-the-art VAE architecture for encoding shapes. Since [Litany et al. 2017] operates on Euclidean coordinates, it does not cope well with large rotations, so we compare with [Tan et al. 2018a] which uses a rotation-invariant representation. Our use of convolutional layers better exploits spatial relationships. For each of the three datasets Camel, Horse and SCAPE, we randomly select half the shapes in the dataset for training, and the test shapes used as our nearest neighbors. To demonstrate the capability of VC-GAN, we use 75% of randomly selected shapes for training, and the test shapes used as our nearest neighbors.

5.1 Quantitative and Qualitative Evaluation of Network Components

We first analyze the effectiveness of our network components. We propose SimNet to approximate the light field distance between a pair of shapes, which ensures efficient evaluation and differentiability. To demonstrate its convergence, we take the Horse and Camel shape sets, randomly choose 75% shape pairs for training and the remaining shapes for testing, and in Fig. 13 plot errors over training iterations on both the training and test sets. It clearly shows that our training process is stable, and the learned model has good generalizability as it also works well for the test set containing unseen shapes. The average relative errors on the training and test sets are 7.63% and 8.02% respectively, showing its capability to efficiently calculate shape similarity.

To demonstrate the effectiveness of each loss term in our VC-GAN, we performed an experiment of transferring deformation from Horse to Camel, as shown in Fig. 14. These two datasets have pairs of corresponding shapes. Although our method does not exploit this property, it is useful to provide ground truth to make evaluation easier. To demonstrate the capability of VC-GAN, we use 75% of randomly selected shapes for training, and the test shapes used as
5.2 Deformation Transfer Results and Comparisons

We now demonstrate our system using various deformation transfer examples. Note that our method is the only method that is fully automatic with unpaired shape sets. In comparison, the method [Sumner and Popović 2004] requires point-wise correspondence to be established, which typically needs 70-80 correspondences to be manually specified. The method [Baran et al. 2009] requires building cages, and further specifying 20-40 pairs of corresponding points. The method [Ben-Chen et al. 2009] requires shapes in the source and target datasets have one-to-one correspondences, which is not normally satisfied for independent shape sets. Even if this is possible, it involves user effort to choose semantically related shape pairs. For all the methods compared, when the input is a deformation sequence, we feed in each shape and obtain the output shape to form a transferred deformation sequence. All these methods work well with this simple strategy; see the accompanying video.

Figure 15 shows comparative results of deformation transfer from Flamingo to Crane (human action) datasets. In this example, the source and target shapes differ substantially, so methods that require point-wise correspondences not only need a large number of correspondences, but can also suffer from large, local shape deformations to transform from one shape to the other. The results [Ben-Chen et al. 2009; Sumner and Popović 2004] have clear distortions on the left leg. Moreover, since the Flamingo shape does not have arms, it is not possible to control the arm deformation, even if the target shape set has a suitable arm movement that accompanies a corresponding leg movement. The result of [Baran et al. 2009] does not have artifacts. However, despite carefully choosing 7 pairs of shapes that are semantically similar and have broad coverage of poses, the deformation result is still somewhat dissimilar (e.g. legs) to the input. Our method is not only fully automatic, but also follows the deformation of the source shape, taking into account both shape similarity and plausible deformation of the target shapes.

We compare our method with existing methods requiring manual correspondence specification. The deformation transfer results are shown in Fig. 17. The result of our method (d) is fully automatic and artifact free. Compared with other methods, the VAE component in our network can better fit the data distribution and mitigate self-intersection. The results (b) and (c) are obtained using manually labeled correspondences (15 pairs and 40 pairs respectively) along with the method [Sumner and Popović 2004]. The labeled correspondences are visualized in Fig. 17. It can be seen that (c) is better than (b), with less severe artifacts (see the closeups), however it is more time-consuming to label. It takes a skilled user 7 minutes 36 seconds and 18 minutes 13 seconds to label 15 pairs and 40 pairs of correspondences, respectively. Our fully automatic method produces a better result.

Figure 18 shows deformation transfer from the Lion to Cat datasets. Our automatic method produces similar and sometimes better results than [Sumner and Popović 2004] which requires manually specifying correspondences. Differences between results are highlighted in Fig. 19 using color coding to show the vertex displacements. It clearly highlights the front right leg, where our result looks closer to the source shape.
Fig. 14. The evaluation of each loss term for our VC-GAN model for deformation transfer. (a) input shapes, (b) results with discriminators performed in the latent space, (c) results without cycle consistency loss, (d) results without adversarial loss, (e) results without visual similarity loss, (f) our results, (g) ground truth.

Fig. 15. Comparison of different methods for deformation transfer from Flamingo to Crane (human action) datasets. (a) source shape, (b) deformation transfer result of [Sumner and Popović 2004], (c) deformation transfer result of [Baran et al. 2009], (d) deformation transfer result of [Ben-Chen et al. 2009], (e) our result.

Fig. 16. Two pairs of shapes with manually labeled landmarks. (a)(b) 15 pairs of manually labeled landmarks between the source and target shapes (c)(d) 40 pairs of correspondence points.

In Fig. 20, we also compare our results with alternative methods without correspondences. We would like to stress that neither of these methods are intended for mesh deformation transfer. The results in the second row are obtained using [Rustamov et al. 2013] where each shape is mapped to the nearest shape from the other set in the joint embedding space. The method produces some plausible results, but can also match a shape to an incorrect one (e.g. first column). It also cannot be used to generate deformation sequences since it does not generate new shapes. The results in the third row are generated using a baseline method by representing shapes in a rotation-invariant representation [Gao et al. 2016], applying PCA (Principal Component Analysis) to extract main deformation modes independently for the source and target shape sets and transferring PCA weights from the source shape to the target shape. A similar idea has been used for face deformation transfer [Cosker et al. 2008]. However, for general shapes, it is challenging to identify meaningful basis deformations that can be used to reliably establish the mapping. These are in fact expected due to the nature of the shape sets themselves. We aim to produce plausible shapes sampled from the target space that are visually similar to the source shape. While we wish the deformed target to resemble the source, it should also be consistent with the target samples to ensure it is semantically plausible for the target shape. There is a tradeoff between these two factors, which explains the discrepancies. For example, a person naturally walks in a different way than a Flamingo, and therefore when transferring the walking between them, the poses may look different, but the results are indeed plausible. Figures 21 and 22 show results for our automatic deformation transfer of a galloping elephant to a horse and a collapsing horse to a camel. Our synthesized shapes are visually similar to the target shape space while faithfully following the input deformation. We also evaluate the robustness of our method on the size of the training data. We independently remove 30% of randomly chosen shapes from the input datasets of the collapsing horse and camel, and the deformation transfer results on the reduced shape sets are similar to the transfer results with the whole datasets as shown in Fig. 22. Figure 23 shows an example of transferring hand deformation to a pair of pants, with finger movement effectively transferred to “walking” pants.

Figure 1 shows an example of transferring deformation from a fit person (ID: 50009) to a fat person (ID: 50002) (both from the MPI DYNA dataset). The training datasets contain thousands of shapes with different poses, and our method successfully generates suitable shapes in the shape space, effectively transferring rich actions (punching, running, etc.), despite substantial differences in their body shapes.

Figure 24 shows an example of transferring the deformation from a running person to a robot. The mesh dataset with various poses of the robot is from [Xu et al. 2009]. Our method produces plausible robot deformation following the human action. In this example, the robot mesh is composed of multiple components and our method cannot be directly applied. To cope with this, the first mesh in the robot dataset is converted to a singly connected mesh using a volumetric mesh repair approach [Ju 2004]. The correspondences between the singly connected mesh and the original multi-component mesh
5.3 Extension of Our Method to Semantic Transfer

Our proposed deformation transfer method based on visual similarity is fully automatic. However, sometimes shapes which are semantically related may not be visually similar. To address this, we present a simple extension of our method that takes semantically related pairs as in [Baran et al. 2009] in addition to two unpaired shape sets to perform semantic deformation transfer. Inspired by the Triplet Network [Hoffer and Ailon 2014], we modify the SimNet to two independent fully connected networks, each embedding the source or target latent space to a lower dimensional Euclidean space, where the distance can represent the semantic differences. An example is shown in Fig. 25 where the deformation is transferred from a face to a running female [Pons-Moll et al. 2015]. We expect the semantic pairs in Fig. 25 to express that closing an eye corresponds to lifting a leg. Similar to [Baran et al. 2009], we select 19 pairs of faces and female shapes that are semantically similar as input to train the semantic similarity network. To train it, we use a simple strategy such that the network aims to predict distance 0 for the 19 given pairs of shapes, and the maximum LFD for all the other shape pairs. We compare our method with another method [Baran et al. 2009], which also requires corresponding shape pairs. Results in Fig. 26 show that our method generates more semantically similar shapes than [Baran et al. 2009]. This is because generally two datasets may not have pairs with perfect one-to-one semantic correspondence, where shape pairs selected by a human will inevitably introduce conflicts and confuse the computation of shape bases [Baran et al. 2009]. In such cases, our method can exploit unpaired shapes in the 3D model datasets with VAE to characterize shape distributions in the datasets. In Fig. 26, motions in the second row produced by [Baran et al. 2009] tend to freeze and the legs do not match the eyes. In contrast, our results in the third row have higher semantic similarity.

are then easily established based on nearest points. Using these correspondences as soft constraints, the singly connected mesh can be deformed to approximate the dataset with multi-component meshes. Our automatic unpaired deformation transfer technique is applied to the human and the singly connected robot mesh datasets. Once the deformed robot is obtained, we obtain the rigid rotation and translation of each component based on the vertex correspondences to obtain the deformed multi-component robot shape.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we propose a novel architecture for automatic mesh deformation transfer, which is also flexible, allowing source and target shape sets to be unpaired. Our method achieves state-of-the-art deformation transfer results. However, the method may still be improved. The current visual similarity measure works well when the deformation is visually similar. It may not work as well for semantically similar but visually very different shapes. In the future, we would like to investigate using neural networks to estimate more advanced visual similarity measures. Also, our VC-GAN in the latent space has the potential to be useful in other 3D shape synthesis applications, which we would like to explore in the near future. In cases where the two shape sets have significant visual differences such as a horse and a human (ID: 50009), it is challenging to construct a reliable visual similarity metric between them. An example is shown in Fig. 27. Although both the horse and the person are running, the corresponding body parts between the horse and the person do not move simultaneously. The semantic transfer techniques in Sec. 5.3 could be applied to transfer the deformation.
from the horse to the human in a semantic manner, although automatic methods not requiring user effort would still be preferred. The main problems of the current visual similarity metric are that the light field distance may not be able to measure the differences

Fig. 20. Deformation transfer from Flamingo to Crane (human action) datasets. First row: source flamingo shapes, second row: nearest shapes from the embedding of [Rustamov et al. 2013], third row: shapes obtained by transferring PCA weights [Cosker et al. 2008], fourth row: our deformation transfer results.

Fig. 21. Deformation transfer from elephant to horse. First row: source elephant shapes, second row: our results of deformation transfer to the horse shape.

Fig. 22. Deformation transfer from camel to horse with collapse effect. First row: source camel shapes, second row: our results of deformation transfer to the horse shape, third row: the results obtained by independently removing 30% of randomly chosen shapes in the datasets, and the results are very similar to those in the second row, demonstrating that our method does not require shape sets to have shapes with corresponding poses.

Fig. 23. Deformation transfer from Hand to Pants datasets. First row: source hand shapes, second row: our results of deformation transfer to pants.

Fig. 24. Deformation transfer from a person to a robot. First row: source person shapes, second row: the transferred robot shapes.
in a semantic manner and it is not easy to distinguish small visual differences.

We also perform another experiment where all the pose shapes of two persons (ID: 50020 and 50026) in the MPI DYNA dataset are used as input. As shown in Fig. 28, in general our method reasonably transfers deformation of the first person to the second. However, there are some visual differences between two shapes due to the limitations of the visual similarity metric. In future work, we will develop more powerful visual similarity metrics to better distinguish subtle visual differences and better capture semantic similarities.

ACKNOWLEDGMENTS

We thank the reviewers for their constructive comments and suggestions. We would also like to thank Prof. Mirela Ben-Chen for providing the code of [Ben-Chen et al. 2009], Dr. Bob Sumner for providing the executable of [Sumner and Popović 2004] and the mesh models, Dr. Raif Rustamov for providing the deformed ball shapes of [Rustamov et al. 2013]. This work was supported by National Natural Science Foundation of China (No. 61872440, No. 61828204, No. 61502453 and No. 61772499), Young Elite Scientists Sponsorship Program by CAST (No. 2017QNRC001), Royal Society-Newton Mobility Grant (No. IE150731), CCF-Tencent Open Fund and NVIDIA Corporation with the GPU donation. Weiwei Xu was partially supported by National Natural Science Foundation of China (No. 61732016), Alibaba IDEA Lab and fundamental research fund for the central universities (2017YFB1002600).

REFERENCES


