Designing Deep Learning Architectures for Multiview 3D Shape Estimation Using Image Transformers

Kanika Singla and Parmanand Astya

Abstract The task of 3D shape generation for realistic data is an important challenge that needs to be addressed in the domain of computer vision, robotics, and graphics which serve as a building block for many real-time applications like autonomous driving or 3D modeling, etc. Estimating shapes from a few 2D images are fundamentally ill-posed as numerous 3D shapes can be explained by a few images. In the absence of complete information, recently, deep learning has been used to fill in the gap by leveraging data driven category level priors. In this work, we propose a novel 3D shape estimation network that uses an image transformer to better encode the shape features into a latent representation which is later decoded using a multilayer perceptron. Our experiments show that image transformers are better than convolution-based encoders due to their wide attention capability. We perform both qualitative and quantitative experiments to demonstrate the effect of new architecture on shape quality and detail.

Keywords Computer vision · Multiview shape · 3D reconstruction · Shape estimation · Multiview images · Image transformers

1 Introduction

We are particularly interested in the problem of 3D shape estimation, which involves estimating the complete structure of objects from multiview images of the target. Many relevant real-world applications need this as prerequisite. 3D shape estimation, for example, will help autonomous vehicles track objects [1], and robots find out the best grasping position [2]. Humans can naturally approximate the shapes of objects using only the most basic information. Human eyes can easily perceive the 3D structure from the limited, ambiguous, and even occluded 2D details. However, this task becomes particularly challenging, when this concept is applied to the machines, due to ambiguity generated from single view images, occluded images, and sparse

K. Singla (✉) · P. Astya
Department of Computer Science and Engineering, School of Engineering & Technology, Sharda University, Greater Noida, India

N. Marriwala et al. (eds.), *Emergent Converging Technologies and Biomedical Systems*, Lecture Notes in Electrical Engineering 841,
https://doi.org/10.1007/978-981-16-8774-7_42

point clouds [3]. It is unfair to expect them to predict a deterministic output from an uncertain input [4].

The aim of this paper is to create a geometric representation of the underlying 3D world from a set of images captured from various camera positions. We are estimating the 3D shape of an object using 2D images taken from (digital) cameras by using learned data driven priors of other objects.

With the rising popularity of deep learning, several methods have been proposed to do 3D shape generation from single [5] and multiple images [6]. At the core of this, development is the successes in designing convolutional kernels in 2D/3D space that can learn meaningful features. Frequently, these convolutional neural networks are designed to have several consecutive layers making them "deep" with the objective of increasing their receptive field that is found to aid in learning richer features [6]. Recent research efforts have, however, highlighted the limitations of convolutional neural networks in learning long range relationships in spatial [7] and temporal [7] dimensions. To resolve these limitations, several methods have been investigating transformer networks to learn richer representations that additionally encode long range spatio-temporal dependencies [8].

In this work, we exploit the capability of image transformers to learn such long range dependencies to improve the task of multiimage 3D shape generations. In particular, we show that compared to the popular ResNet-18 [9] encoder, an image transformer such as data efficient transformer (DiT) [7] can capture long range dependencies in the input image and consequently learn richer latent shape representations. We also propose a novel architecture that uses this image transformer and a point-based multilayer perceptron to generate 3D shapes.

While recurrent neural networks (RNNs) have been used in recent approaches to learn object's mapping between distinct views, [10]. These designs are inefficient in terms of computation, and the RNN model's input views are sensitive to the order of permutation [11] which makes it difficult to work with a collection of different unordered acquired views. In contrast, we use Max pooling operation to fuse latent representations from multiple views, thus making our approach permutation invariant.

This paper has been broken down into different sections. The first section of this paper includes the paper's introduction, as well as the problem statement's goal, inspiration, and objectives. Section 2 discusses the literature review of the concepts used in the study. The image transformers' background is detailed in Sect. 3. Section 4 includes qualitative and quantitative experiments analysis. Section 5 covers the interpretation and discussion of the results, as well as the work's contribution to the previous studies. This segment also discusses the potential scope of the work.

2 Related Work

Generating the shape of a 3D object from a few images is an ill-posed problem. We now review the relevant literature in both traditional and learning-based 3D

reconstruction. We then briefly review the recent transformer literature as it relates to this work.

A. **Traditional Multiview 3D Reconstruction:** In geometric processing, shape completion has a long history. Many relevant real-world applications need this as a prerequisite such as in tracking objects for autonomous vehicles [1] grasping for autonomous robotic manipulators [2]. For dense point clouds, a common technique to convert them to meshes is Poisson surface reconstruction [12]. Other classical techniques resort to 3D reconstruction from 2D images by leveraging multiview consistency [13]. While more broadly, structure from motion is performed to do large-scale reconstruction with both posed [14] and in the wild images [15], they are often plagued by non-lambertian surfaces, occlusions, small baselines causing degeneracy, and changes in illuminations. Thus, only by collecting millions of images [16] and using hand-crafted edits by artists can such techniques be used reliably. These limitations motivate explorations into data driven methods that do not suffer from such issues.

B. **Deep Learning on 3D Shapes:** Deep learning allows use of data driven priors for resolving shape ambiguities and thus enabling complete shape generation. Broadly, they can be characterized based on the type of 3D representation that is regressed. A mesh-based representation [17] stores the surface information as a list of vertices and faces. Choy et al. [10] put forth a deep generative model for modeling voxelized 3D shapes that leverage 3D convolution kernels for shape generation. To address the drawbacks of the voxel representation, authors argued for generating point clouds [18] instead using a single image. Rich literature on implicit function learning [19, 20] for shape representation and reconstruction tasks has been done by the researchers in the computer vision and graphics community.

C. **Transformers:** Transformer models have excelled at a number of tasks in natural language processing, including computer translation, document classification [21]. The core part of a transformer is its multi-head self-attention mechanism that combines the characteristics of each token pair in the embedding sequence. Transformer has recently been applied to the domain of computer vision with great success [8, 22]. Impactful and promising applications have been shown by [23]. ViT [8] applies transformer to sequences of image-patches for the task of image classification, without using CNN features, when pre-trained on a large-scale dataset, and demonstrates comparable and higher classification accuracy. These have significant advantages over their CNN counterparts when it comes to attending to long range spatio-temporal dependencies. This is key to learning much richer representations opening several avenues for future research not just in 3D reconstruction but also video understanding [24], scene understanding [25].

3 Method, Background, and Notations

We now describe our image transformer-based multiview 3D object reconstruction method, which has been given a set of images (3 in this case) extracts per image latent vectors and fuses them using a Max pooling operation. The pooled representation is permutation invariant and is later used to extract complete 3D shape via a point-based multilayer perceptron network. We will first provide a brief background on the architecture of a vision transformer [8] and then elaborate our proposed pipeline.

3.1 Background on Vision Transformer

Transformers were first introduced in [26] as a two-part architecture (encoder and decoder) that allows you to turn one series into another. However, it differs from existing sequence models like RNN and LSTM in a sense that it does not include recurrent networks. Figure 1 shows the visualization of transformers. The encoder is located on the left, while the decoder is located on the right. Encoder and decoder are both made up of components that can be placed on top of each other several times, as shown by **Nx** in the diagram. As shown in the Fig. 1, the modules are primarily made up of multi-head attention and feedforward layers.

Positional encoding is another crucial component of the model. Since a sequence relies on the order of its components, and we do not have any recurrent networks that can remember how sequences are fed into a model, we need to allocate each component of our sequence a relative place. These coordinates are added to the embedded representation of each letter (n-dimensional vector).

For a sequence of $\mathbf{Y}$ query vectors (packed into $\mathbf{Z} \in \mathbf{R}^{\mathbf{Y}\times\mathbf{d}}$), it produces an output matrix (of size $\mathbf{Y} \times \mathbf{d}$):

$$\textbf{Attention}(\mathbf{Z},\ \mathbf{K},\ \mathbf{V}) = \textbf{Softmax}\left(\mathbf{Z}\mathbf{K}^{\mathbf{T}}/\sqrt{\mathbf{d}}\right)\mathbf{V}, \tag{1}$$

where the Softmax function is applied over each row of the input matrix and the $\sqrt{\mathbf{d}}$ term is used to normalize the result.

3.2 Architecture Details

We now provide details about the network architecture. In the Fig. 2, we show how a set of images (set of 3 in this case) is fed to a ResNet-18 encoder R_ϕ, with shared network weights. This results in view dependent latent vectors L1, L2, L3. In order to fuse them together while maintaining permutation invariance, we use the "Max" operator that takes the max(L1, L2, L3) along the views and results in the global latent code L. This latent code is then decoded via a set of MLP layers Q_θ to generate

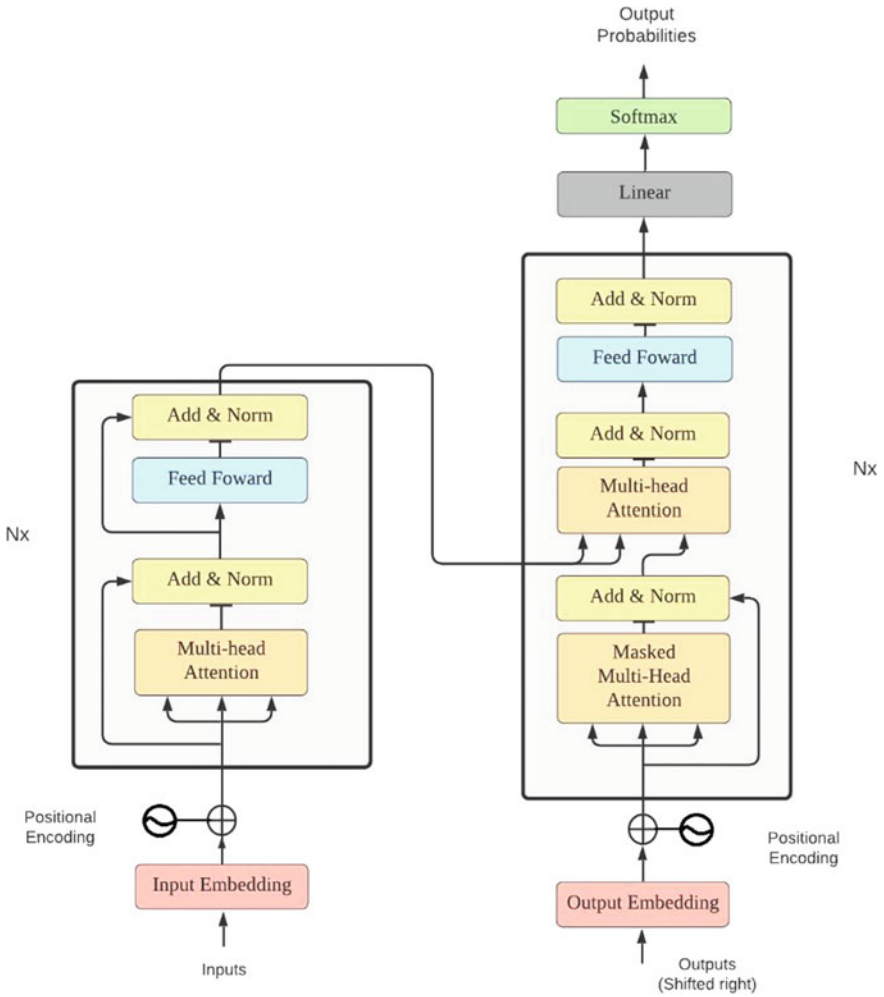


Fig. 1 Transformer model

a point cloud of 1024 points. The MLP being used here has the architecture as [1024, 1000, 1000, 1000, 3072]. While this architecture is plausible and gives us a point cloud based on input views of the object, we propose to further improve the results by means of transformer layers. As discussed above, transformers are able to better capture the long range dependencies in images and as such provide much richer latent codes. Here, the data efficient transformer [8] T_ϕ is used to encode the image set I, which is passed through the max operator to get global latent code L and subsequent point cloud via the decoder Q_θ (same as above). We train both these architectures on

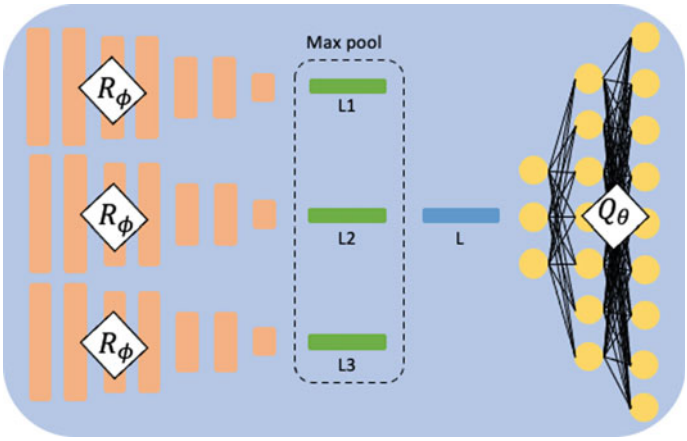


Fig. 2 ResNet-18 as encoder

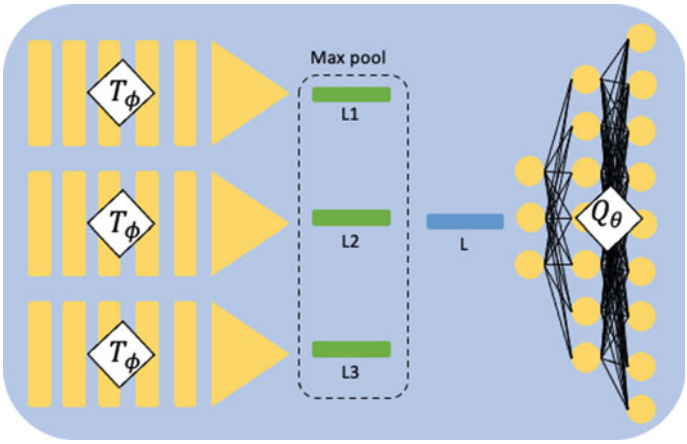


Fig. 3 Image transformer encoding the imageSet

ShapeNet dataset [27] and use the renderings provided by [10] for 120 epochs with a learning rate of 1e-3 and weight decay of 0.98 after every 250 iterations. In total, this takes roughly 30 h on a single NVIDIA 2080Ti GPU (Fig. 3).

4 Experiments and Result Analysis

We now provide some quantitative and qualitative results and comparisons between ResNet-18 encoder and image transformer to highlight benefits of using image transformers for the task of multiview 3D reconstruction.

Table 1 Comparing chamfer distance for multiview reconstruction on resnet-18 and image transformer

Encoder type	Resnet-18	Image transformer (ours)
Chamfer distance	0.83	**0.78**

A. Quantitative Results

For quantitative comparisons, we evaluate the bidirectional chamfer distance (Eq. 1) of the generated point clouds. Here, lower values are better. Chamfer distance is a common metric that quantifies distance of two point clouds, from p to q, which is defined as

$$L(\Theta) = \sum \min_{q \in MG} ||p - q||^2 + \sum \min_{p \in MP} ||p - q||^2 \quad (2)$$

where **MG** and **MP are** ground mesh and predicted mesh, respectively (Table 1).

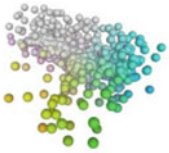
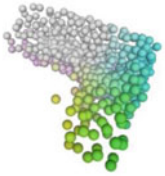
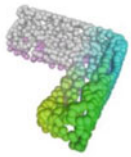
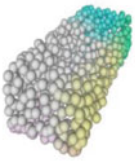
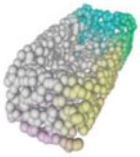
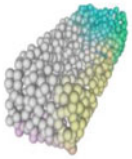
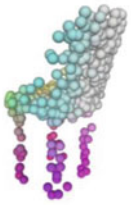
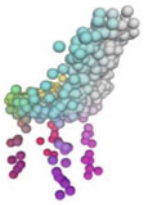
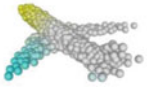
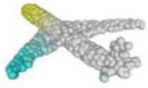
B. Qualitative Results

We now visualize some examples of the reconstructions obtained from ResNet-18 and image transformer encoders, respectively, based on the input image triplets (Table 2).

5 Conclusion and Future Scope

In this work, we presented a novel architecture for 3D shape generation from multiple point clouds and demonstrated that image transformers achieve significantly better performance for this task compared to their ResNet counterparts. We provide both qualitative and quantitative results to establish this claim. In future, we would like to investigate the effects of using a transformer-based decoder (instead of an MLP). I believe that the gains we see by replacing convolutional encoders with transformers will also translate to the decoder side and would hopefully result in much higher fidelity of reconstructions.

Table 2 Qualitative results on 3d reconstruction

Image	ResNet-18	Image transformer (ours)	Ground truth
			
			
			
			

References

1. Giancola S, Zarzar J, Ghanem B (2019) Leveraging shape completion for 3d siamese tracking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1359–1368
2. Varley J, DeChant C, Richardson A, Ruales J, Allen P (2017) Shape completion enabled robotic grasping. In: 2017 IEEE/RSJ international conference on intelligent robots and systems (IROS), pp 2442–2447. IEEE
3. Qi CR, Su H, Mo K, Guibas LJ (2017) Pointnet: deep learning on point sets for 3d classification and segmentation. In Proc CVPR
4. Mandikal P, Navaneet KL, Agarwal M, Babu RV (2018) 3D-LMNet: latent embedding matching for accurate and diverse 3D point cloud reconstruction from a single image. arXiv preprint arXiv:1807.07796
5. Niu C, Yu Y, Bian Z, Li J, Xu K (2020). Weakly supervised part-wise 3D shape reconstruction from single-view RGB images. In Computer graphics forum, vol 39, No 7, pp 447–457

6. Choy CB, Xu D, Gwak J, Chen K, Savarese S (2016) 3d-r2n2: a unified approach for single and multi-view 3d object reconstruction. In European conference on computer vision (ECCV)
7. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H (2020) Training data-efficient image transformers & distillation through attention. arXiv preprint arXiv:2012.12877
8. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, J´egou H (2020) Training data-efficient image transformers & distillation through attention. arXiv preprint arXiv:2012.12877
9. Ou X, Yan P, Zhang Y, Tu B, Zhang G, Wu J, Li W (2019) Moving object detection method via ResNet-18 with encoder–decoder structure in complex scenes. IEEE Access 7:108152–108160
10. Choy CB, Xu D, Gwak J, Chen K, Savarese S (2016) 3d-r2n2: a unified approach for single and multi-view 3d object reconstruction. In Proceedings of the European conference on computer vision (ECCV)
11. Vinyals S Bengio, Kudlur M (2016) Order matters: sequence to sequence for sets. In International Conference on Learning Represen-tations (ICLR)
12. Kazhdan M, Bolitho M, Hoppe H (2006) Poisson surface reconstruction. In Proceedings of the fourth Eurographics symposium on Geometry processing, vol 7
13. Hartley R, Zisserman A (2003) Multiple view geometry in computer vision (cambridge university). *C1 C3*, *2*
14. Yingze Bao S, Chandraker M, Lin Y, Savarese S (2013) Dense object reconstruction with semantic priors. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1264–1271
15. Wang N, Zhang Y, Li Z, Fu Y, Liu W, Jiang YG (2018) Pixel2mesh: generating 3d mesh models from single rgb images. In Proceedings of the European conference on computer vision (ECCV), pp 52–67
16. Agarwal S, Furukawa Y, Snavely N, Simon I, Curless B, Seitz SM, Szeliski R (2011) Building Rome in a day. Commun ACM 54(10):105–112
17. Achlioptas P, Diamanti O, Mitliagkas I, Guibas L (2018) Learning representations and generative models for 3d point clouds. In International conference on machine learning, pp 40–49. PMLR
18. Liu Y, Fan B, Meng G, Lu J, Xiang S, Pan C (2019) Densepoint: learning densely contextual representation for efficient point cloud processing. In Proceedings of the IEEE/CVF international conference on computer vision, pp 5239–5248
19. Park JJ, Florence P, Straub J, Newcombe R, Lovegrove S (2019) Deepsdf: learning continuous signed distance functions for shape representation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 165–174
20. Michalkiewicz M, Pontes JK, Jack D, Baktashmotlagh M, Eriksson A (2019) Deep level sets: implicit surface representations for 3d shape inference. arXiv preprint arXiv:1901.06802
21. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, ... Amodei D (2020) Language models are few-shot learners. arXiv preprint arXiv:2005.14165
22. Devlin M-W, Chang K Lee, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXivpreprint arXiv:1810.04805
23. Dosovitskiy L, Beyer A, Kolesnikov D, Weissenborn X, Zhai T, Unterthiner M, Dehghani M, Minderer G, Heigold S, Gelly J Uszkoreit, Houlsby N (2021) An image is worth 16 x 16 words: Trans-formers for image recognition at scale. In International Conference on Learning Representations (ICLR)
24. Kwon H, Kim M, Kwak S, Cho M (2020) Motionsqueeze: neural motion feature learning for video understanding. In European conference on computer vision, pp 345–362. Springer, Cham
25. Guo Z, Huang Y, Hu X, Wei H, Zhao B (2021) A survey on deep learning based approaches for scene understanding in autonomous driving. Electronics 10(4):471
26. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser LU, Polosukhin I (2017) Attention is all you need. In Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems, vol 30. Curran Associates, Inc
27. Chang AX, Funkhouser T, Guibas L, Hanrahan P, Huang Q, Li Z, ... Yu F (2015) Shapenet: an information-rich 3d model repository. arXiv preprint arXiv:1512.03012