# Unsupervised Learning of Disentangled 3D Representation from a Single Image

Junliang Lv, Haiyong Jiang and Jun Xiao*

University of Chinese Academy of Sciences, China
xiaojun@ucas.ac.cn

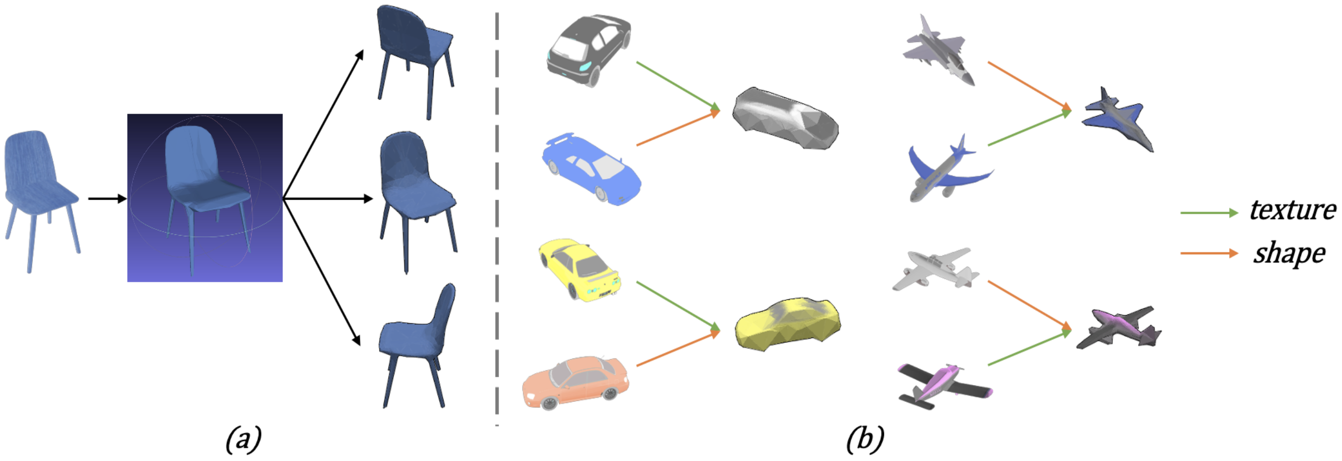

**Figure 1:** *Applications of our method.* ***(a)*** *Our method can reconstruct textured shape of an input image and use it to generate consistent novel view images.* ***(b)*** *Our method can help image editing by explicitly swapping the shape factors or texture factors so that images with novel textures and shapes can be generated.*

**Abstract**

*Learning 3D representation of a single image is challenging considering the ambiguity, occlusion, and perspective project of an object in an image. Previous works either seek image annotation or 3D supervision to learn meaningful factors of an object or employ a StyleGAN-like framework for image synthesis. While the first ones rely on tedious annotation and even dense geometry ground truth, the second solutions usually cannot guarantee consistency of shapes between different view images. In this paper, we combine the advantages of both frameworks and propose an image disentanglement method based on 3D representation. Results show our method facilitates unsupervised 3D representation learning while preserving consistency between images.*

**CCS Concepts**

• ***Computing methodologies*** → ***Image representations;*** *Reconstruction; Mesh models;*

## 1. Introduction

Learning 3D representation of an image, e.g. the 3D shape and the texture, is important for many real-world applications in image synthesis and augmented reality. For example, it is desirable for a photographer to edit objects in an existing image so that better creative intent can be achieved. When buying a furniture, it is also important for a customer to know how well the product will look like in the room. To this end, we need a full understanding of the geometry and texture of an object.

Recent progresses on generative models and deep neural networks have revolted the field of 3D representation learning and image synthesis. For example, StyleGAN [KLA19] creates high-quality 2D images by unsupervised learning from a large corpus of images. Though style mixture and linear separability between different factors are demonstrated, it is hard to change different fac-

DOI: 10.2312/egp.20211030

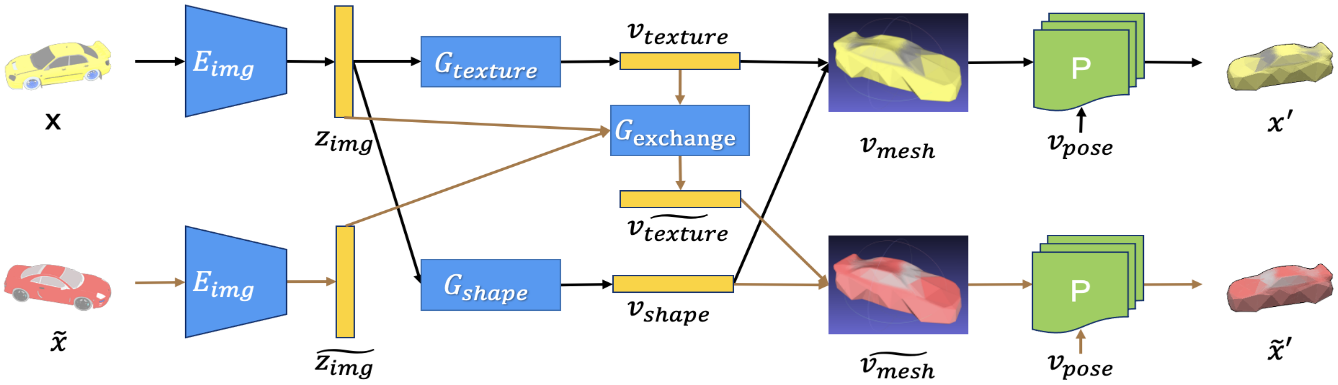

**Figure 2:** *The pipeline of our framework.* ***Black arrows*** *show the reconstruction branch, while* ***brown arrows*** *give the disentangling branch for shape and texture disentanglement.*

tors, respectively, e.g. texture and posture. Recent works explore supervisions on the learning process so that different factors can be disentangled. [KTEM18] uses category-specific 3D models templates and key points in images to learn the disentanglement of different factors, while VON [ZZZ*18] learns 3D representation of an object by explicitly providing 3D supervisions. NGP [CCOCM20] combines neural methods with traditional computer graphics to control image components by explicit attribute disentanglement. Though this work analyzes the internal mechanism of the image more thoroughly, but it also depends on the sampling of 3D generation network. However, annotation of these ground truths, like 3D models and 3D geometry of an image, is tedious and expensive.

In this work, we address the above problems by 3D disentanglement framework. We disentangle an object as shape and texture factors. Then we can generate novel objects by switching factors between two objects. The generated objects should have similar appearance or silhouette with its source or target images, depending on the used factor. We encode these constraints as loss functions and learn the 3D representation in an unsupervised fashion.

## 2. Learning disentangled 3D representation

The framework of our method is illustrated in Figure 2. Our basic assumption is that an object can be represented as texture and geometry shape. Thus we employ an off-the-shelf backbone, i.e. VGGNet, to extract image feature $z_{\text{img}}$ from a given image $x$. Then the shape and texture can be predicted as latent vector $v_{\text{shape}}$ and $v_{\text{texture}}$. The reconstruction branch first recovers 3D shape of an image by combining $v_{\text{shape}}$ and $v_{\text{texture}}$ into a deformed mesh $v_{\text{mesh}}$. And images in different views can be generated by a Soft Rasterizer [LLCL19] for differential rendering. The disentangling branch mixes the shape vector $v_{\text{shape}}$ with the texture vector $\tilde{v}_{\text{texture}}$ of another image $\tilde{x}$ so that novel image $\tilde{x}'$ with the same shape but different textures can be generated. We learn the network through optimizing the reconstruction branch with the typical perceptual loss [JAFF16] and reconstruction loss, and the disentangling branch with the silhouette loss and the style loss [JAFF16].

We evaluate our method on the ShapeNetCore dataset. In Figure 1, we show results of pose editing and texture editing. By explicitly using a 3D mesh representation, our method can generate novel views with consistent geometry as shown in Figure 1 **(a)**. In addition, our method supports the editing of textures and shapes by replacing the texture code with that of another image. Results in Figure 1 **(b)** demonstrate our method can achieve an disentanglement of textures and shapes even without any supervision.

## 3. Conclusion

In this work, we present a novel approach to learn 3D representation from a single image. Our method can disentangle different factors of the 3D representation, which enables interesting applications, like texture remapping, pose editing and object replacement in an image. In the future, we will explore more challenging scenarios, like novel image synthesis based on product images in shopping websites and room re-furnishing based on a user image.

## 4. Acknowledgments

This work is supported by the National Natural Science Foundation of China (U2003109) and the Fundamental Research Funds for the Central Universities.

## References

[CCOCM20] Chen X., Cohen-Or D., Chen B., Mitra N. J.: Neural graphics pipeline for controllable image generation. *arXiv preprint arXiv:2006.10569* (2020). 2

[JAFF16] Johnson J., Alahi A., Fei-Fei L.: Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision* (2016), Springer, pp. 694–711. 2

[KLA19] Karras T., Laine S., Aila T.: A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 4401–4410. 1

[KTEM18] Kanazawa A., Tulsiani S., Efros A. A., Malik J.: Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 371–386. 2

[LLCL19] Liu S., Li T., Chen W., Li H.: Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 7708–7717. 2

[ZZZ*18] Zhu J.-Y., Zhang Z., Zhang C., Wu J., Torralba A., Tenenbaum J. B., Freeman W. T.: Visual object networks: Image generation with disentangled 3d representation. *arXiv preprint arXiv:1812.02725* (2018). 2