Learning Task-Agnostic 3D Representations of Objects by De-Rendering

Tzofi Klinghoffer[*] [1], Kushagra Tiwary[*] [1], Arkadiusz Balata[1], Vivek Sharma[1,2], and Ramesh Raskar[1]

[1]Massachusetts Institute of Technology, [2]Harvard Medical School
{tzofi,ktiwary,arkadius,vvsharma,raskar}@mit.edu
[*] Equal contribution

Abstract. State-of-the-art unsupervised representation learning methods typically do not exploit the physical properties of objects, such as geometry, albedo, lighting, and camera view, and, when they do, multi-view images are often needed for training. We show that de-rendering, a way to reverse the rendering process to recover these properties from single images without supervision, can also be used to learn task-agnostic representations, which we dub *physically disentangled representations* (PDRs). While de-renderers predict distinct physical properties, the features learned in the process may not be disentangled. To ensure meaningful features are encoded by de-rendering and thus prevent overreliance on decoders, we propose a novel Leave-One-Out, Cycle Contrastive loss (LOOCC) to improve feature disentanglement w.r.t. physical properties, *which leads to higher downstream accuracy.* We evaluate PDRs on downstream clustering tasks, including car classification and face identification. We perform a comparison of our method with other generative representation learning methods for these tasks and find PDRs consistently yield higher accuracy, outperforming evaluated baselines by as much as 18%. Code is available here.

1 Introduction

Unsupervised representation learning (RL) is a long-standing goal in the computer vision community, but many methods do not consider the 3D nature of objects. Generative methods, such as as variational autoencoders (VAEs) [9] and generative adversarial networks (GANs) [6], have been shown to learn semantically disentangled features [1,2], but using a renderer to capture and disentangle the features of physical properties, such as geometry, albedo, lighting, and camera view, is less studied. We propose a way to learn and disentangle the features of physical properties of objects, such as cars and faces, from single-view images with de-rendering and show the use of these features for downstream 2D tasks.

Whereas rendering yields images from scene properties, called scene parameters, de-rendering, a.k.a inverse rendering, enables scene parameters to be predicted from images. By leveraging available priors, synthetic data, or implicit cues, such as symmetry, de-rendering can be learned without supervision, and thus we show can be used for RL. Information on related work is in the supp.

In this paper, we adopt existing de-rendering methods to learn *physically disentangled representations* (PDRs) of two categories of objects - faces and cars.

We show our learned representations, while task-agnostic, have utility across a wide-range of downstream tasks, suggesting physically-meaningful features are useful for semantic tasks. We first introduce a general framework for RL using de-rendering. While de-renderers disentangle scene parameters, there is no guarantee that the feature space is also disentangled, or that the encoders, and not the decoders, have learned physical properties. Thus, we then propose a loss term called Leave-One-Out, Cycle Contrastive loss (LOOCC) to improve disentanglement in the feature space. LOOCC applies contrastive learning to the features of images and their physically augmented counterparts, generated by a renderer. By generating augmented images with only a single changed scene parameter, we can then enforce that features of other scene parameters remain unchanged. Finally, we evaluate our method on clustering, which highlights the ***utility of our raw features without additional, task-specific learning*** and is frequently used for representation learning [11,12,14,18,19]. We show improved performance over baselines on car classification and face identification. In summary, we make the following contributions:

- A novel objective called Leave-One-Out, Cycle Contrastive loss (LOOCC) that improves encoding of de-renderers by disentangling the features of physical scene parameters such that they are more useful for downstream tasks.
- Empirical results showing the utility of our learned features for clustering on car classification and face identification.

2 Proposed Method

2.1 Learning Representations by De-Rendering

We use de-rendering as a mechanism to learn feature representations that are disentangled with regard to the physical scene parameters modeled by the renderer. We leverage the work of Wu *et al.* [17] for de-rendering. The de-renderer is composed of a set of encoders $\mathcal{E}_\theta$, which each take an image $\mathbf{x}$ to a feature, geometry, albedo, light, or camera. The output of each encoder is used to train a decoder $\mathcal{D}_\theta$ to predict the corresponding scene parameter. Predicted light values (1×4) embed ambient and diffuse intensity, pitch, and yaw, while camera values embed camera rotation and translation in x, y, and z (1×6). Both the encoders and decoders are parameterized by neural networks. While $\mathcal{E}_\theta$ predicts physically disentangled features, their explicit counterparts, scene parameters, are predicted by each decoder $\mathcal{D}_\theta$. The scene parameters are then fed into a differentiable renderer (NMR) [8]. The NMR is responsible for constraining the encoder-decoders by reconstructing the input image $\mathbf{x}$.

We extract features from the last *conv* layer of each encoder for use on downstream tasks. In this work, we only utilize geometry and albedo features for downstream tasks because they capture the most information about the scene.

2.2 Leave-One-Out, Cycle Contrastive Loss

Disentangling scene parameters without supervision is a challenge in de-rendering. We propose a Leave-One-Out, Cycle Contrastive loss (LOOCC) to improve dis-

entanglement as illustrated in Figure S.1 of the supp. material. Our method consists of physical augmentation, cyclic encoding, and contrastive learning.

Physical Augmentation: In addition to reconstructing $\mathbf{x}$, our method generates an augmented image of the scene, $\mathbf{x}_{aug}^{recon}$, by randomly perturbing a predicted scene parameter, $S_{p_{\mathbf{x}_{aug}}}$. Since the light and camera parameters are represented as four and six dimensional vectors, respectively, they can be perturbed by sampling a uniform distribution bounded by the desired range of each value. We randomly perturb light or camera while keeping other parameters the same, and use NMR to render the augmented image.

Cyclic Encoding: We leverage the observation that if we reconstruct an augmented image $\mathbf{x}_{aug}^{recon}$ from $S_{p_{\mathbf{x}_{aug}}}$, it should differ from $\mathbf{x}$ *only* by one scene parameter. The augmented image can then cycle back through $\mathcal{E}_\theta$, generating a set of augmented features that should be the same as the features of $\mathbf{x}$, except for the features of the perturbed param.

Leave-One-Out Contrastive Loss: We leave out the features of the single perturbed scene parameter and use the contrastive loss proposed by [4] to enforce that the rest of the features in $\mathbf{x}$ and $\mathbf{x}_{aug}^{recon}$ are similar, and thus that perturbing one scene parameter does not impact the features for the rest. Our intuition is that by leaving out one set of features, we allow these features in $\mathbf{x}$ and $\mathbf{x}_{aug}^{recon}$ to be pushed apart, while the contrastive loss pulls the rest of the features together. We arrive at the following equation for the contrastive loss. We denote $Z_{\mathbf{x}}^{U}$ and $Z_{\mathbf{x}_{aug}}^{U}$ to be all the features of the corresponding unchanged scene parameters from $\mathbf{x}$ and $\mathbf{x}_{aug}^{recon}$, respectively, leaving out the features that were perturbed.

$$L_{cont}(Z_{\mathbf{x}}^{U}, Z_{\mathbf{x}_{aug}}^{U}) = \frac{exp(sim(Z_{\mathbf{x}}^{U}, Z_{\mathbf{x}_{aug}}^{U}))}{\sum_{k=1}^{2N} exp(sim(Z_{\mathbf{x}}^{U}, Z_{\mathbf{x}}^{U^k})/\tau)} \tag{1}$$

In Eqn. 1, *sim* measures cosine similarity of two feature vectors, N is the mini-batch size of the input, and τ is the temperature parameter.

Total Loss: We define the total loss as a weighted sum of the reconstruction and LOOCC loss terms (Eqn. S.1 in supp.). In our work, we utilize the reconstruction loss proposed by Wu *et al.* [17], but any reconstruction loss can be used.

3 Experiments

In this section, we share our datasets and results. We compare the features learned by de-rendering with [17] (Unsup3D) with those learned with LOOCC. Other baselines are VQ-VAE (best of the VAEs we tested), StyleGAN2 (inversion done with [15]), and Retrieve in Style (RIS). Metrics are described in the supp.

3.1 Datasets

Training: We train two sets of models with our proposed method. Our face models are trained on the UTK Face dataset [20], containing 23,708 images, and we use an 80/10/10 split for train, validation, and test. Our car models are trained on the ShapeNet [3] cars dataset rendered by [17], consisting of 28k train images, 7k validation images, and 7k test images.

Method	Dataset	Cluster Accuracy	F1-Score
VQ-VAE (NeuRIPS'17) [16]	BBT	0.4168	0.2796
StyleGAN2 (CVPR'20) [7]	BBT	0.4261	0.3528
RIS (ICCV '21) [5]	BBT	0.4790	0.4346
Unsup3D (CVPR '20) [17]	BBT	0.5754	0.4572
Ours - Light Only	BBT	**0.6252** ↑	**0.6133** ↑
Ours - Light + View	BBT	0.6096	0.5669
VQ-VAE (NeuRIPS '17) [16]	ShapeNet	0.4915	0.3696
Unsup3D (CVPR '20)	ShapeNet	0.5100	0.3877
Ours - Light Only	ShapeNet	0.5270	0.4016
Ours - Light + View	ShapeNet	**0.5485** ↑	**0.4995** ↑

Table 1: Clustering results for face identification and car classification. Our method using Leave-One-Out, Cycle Contrastive loss (LOOCC) outperforms the baselines. Light indicates light augmentation and Light + View indicates light & view augmentation.

Testing: We use Big Bang Theory (BBT) [13] for identification and a subset of the ShapeNet cars test dataset mentioned above for car classification.

3.2 Face Identification

We demonstrate the utility of our learned representations on the challenging task of video face clustering. We use BBT season 1, episode 1 as prepared by [11]. As in prior work [10,11], clustering is done on a per-track basis by averaging the features of each frame in the track. Our BBT dataset contains 644 tracks and five identities. Our model outperforms all baselines for clustering accuracy and F1 on BBT. Incorporation of the LOOCC loss is especially helpful since it improves disentanglement and thus robustness to lighting and viewpoints (omitted features), which are challenging in this dataset.

3.3 ShapeNet Car Classification

We utilize ShapeNet car data rendered by [17] to evaluate car classification. For each test image, we extract the car name from the ShapeNet metadata and test on five classes: police car, ambulance, limousine, jeep, and Ferrari. The test set contains 1000 images, each rendered with random lighting and viewpoint. We compare our proposed method with VQ-VAE, but not StyleGAN2 or RIS due to lack of synthetic car pre-trained models. Despite the VQ-VAE model having a much larger latent space, our method yields higher accuracy and F1.

4 Discussion

We observe that our method improves *physical* disentanglement. The learned features are predictive of scene parameters, which themselves are disentangled and can be rendered to form an image. We compute Pearson's correlation coefficient (PCC) between each combination of the four learned physical features, both with and without our LOOCC loss. Not only do results indicate low correlation between features, but also that LOOCC further reduces correlation. Without LOOCC, the features have a mean PCC of 0.26 on the BBT dataset, whereas, with LOOCC, mean PCC is 0.18. In conclusion, we have presented a method to learn physically-meaningful features without multi-view images or ground truth.

4.1 Acknowledgements

This research was supported by the SMART Contract IARPA Grant #2021-20111000004. The authors would also like to thank Shangzhe Wu and Ayush Chopra for valuable conversations related to this research.

References

1. Bermano, A.H., Gal, R., Alaluf, Y., Mokady, R., Nitzan, Y., Tov, O., Patashnik, O., Cohen-Or, D.: State-of-the-art in the architecture, methods and applications of stylegan (2022)
2. Burgess, C.P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., Lerchner, A.: Understanding disentangling in β-vae. arXiv preprint arXiv:1804.03599 (2018)
3. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
5. Chong, M.J., Chu, W.S., Kumar, A., Forsyth, D.: Retrieve in style: Unsupervised facial feature transfer and retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3887–3896 (2021)
6. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
7. 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. pp. 4401–4410 (2019)
8. Kato, H., Ushiku, Y., Harada, T.: Neural 3d mesh renderer. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3907–3916 (2018)
9. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
10. Sharma, V., Sarfraz, M.S., Stiefelhagen, R.: A simple and effective technique for face clustering in tv series. In: CVPR: Brave New Motion Representations Workshop. IEEE (2017)
11. Sharma, V., Tapaswi, M., Sarfraz, M.S., Stiefelhagen, R.: Self-supervised learning of face representations for video face clustering. In: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). pp. 1–8. IEEE (2019)
12. Sharma, V., Tapaswi, M., Sarfraz, M.S., Stiefelhagen, R.: Clustering based contrastive learning for improving face representations. In: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). pp. 109–116. IEEE (2020)
13. Tapaswi, M., Bäuml, M., Stiefelhagen, R.: "Knock! Knock! Who is it?" Probabilistic Person Identification in TV-Series. In: CVPR (2012)
14. Tapaswi, M., Parkhi, O.M., Rahtu, E., Sommerlade, E., Stiefelhagen, R., Zisserman, A.: Total cluster: A person agnostic clustering method for broadcast videos. In: Proceedings of the 2014 Indian Conference on Computer Vision Graphics and Image Processing. pp. 1–8 (2014)
15. Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., Cohen-Or, D.: Designing an encoder for stylegan image manipulation. ACM Transactions on Graphics (TOG) **40**(4), 1–14 (2021)
16. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Advances in neural information processing systems **30** (2017)

17. Wu, S., Rupprecht, C., Vedaldi, A.: Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1–10 (2020)
18. Zhang, S., Gong, Y., Wang, J.: Deep metric learning with improved triplet loss for face clustering in videos. In: Pacific Rim Conference on Multimedia. pp. 497–508. Springer (2016)
19. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Joint face representation adaptation and clustering in videos. In: European conference on computer vision. pp. 236–251. Springer (2016)
20. Zhang, Z., Song, Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2017)