# SEMANTICS-GUIDED REPRESENTATION LEARNING WITH APPLICATIONS


National Taiwan University

JIA-WEI YAN, CI-SIANG LIN, FU-EN YANG, YU-JHE LI, YU-CHIANG FRANK WANG

NATIONAL TAIWAN UNIVERSITY



25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021


## HIGHLIGHTS

- We are among the first to explore desirable semantic distribution of latent representations, resulting in interpretable and interpolatable representations
- We propose an Angular Triplet-Neighbor Loss (ATNL), followed by spherical semantic interpolation, which utilizes task-oriented semantic information for representation learning.
- We further extend our learning strategy as a data hallucination technique, which is successfully applied for few-shot image classification

## INTRODUCTION

- We aim at learning latent representations which properly describe the associated semantic information
- To manipulate the latent representations which semantically match the images of interest (e.g., numerical order)

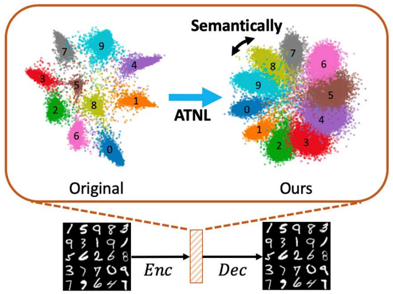


**Figure 1:** Take MNIST as examples, while VAE learns latent representation distributions which correlate with visual appearances of images (left), ours (right) follows the numerical order.

## RELATED WORK


- **Auto-Encoding Variational Bayes** [*Kingma et al., arXiv, 2013*]
- **Understanding and Improving Interpolation in Autoencoders via an Adversarial Regularizer** [*Berthelot et al., ICLR, 2019*]
- **Facenet: A Unified Embedding for Face Recognition and Clustering** [*Schroff et al., CVPR, 2015*]
- **Angular Triplet-Center Loss for Multi-View 3D Shape Retrieval** [*Li et al., AAAI, 2019*]
- **SphereFace: Deep Hypersphere Embedding for Face Recognition** [*Liu et al., CVPR, 2017*]


## ANGULAR TRIPLET-NEIGHBOR LOSS (ATNL)

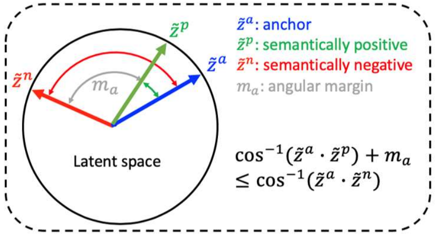


**Figure 2:** Our Angular Triplet-Neighbor Loss (ATNL) is developed to enforce the positive normalized feature $\tilde{z}^p$ become close to the anchor feature $\tilde{z}^a$, while the negative feature $\tilde{z}^n$ would kept away from the positive one by an angular margin $m_a$.

## COMPARISONS

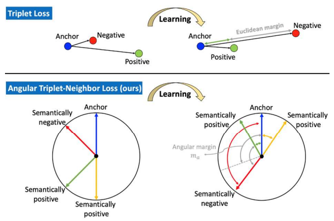


**Figure 3:** Comparisons between the standard triplet loss and our ATNL. The triplet loss minimizes the Euclidean distance between the anchor and the positive instance, while maximizing the distance between the anchor and a negative one. The difference of the two distances are enforced to kept a pre-defined margin. Our ATNL utilizes angular distances and semantically-positive/negative samples, which preserve data discriminativity with a normalized margin.

## SEMANTICS-GUIDED IMAGE GENERATION

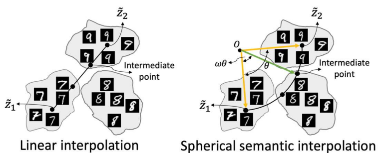


**Figure 4:** Difference between linear and spherical semantic interpolation on the latent spaces. The former interpolates along a straight line from $\tilde{z}_1$ to $\tilde{z}_2$, while the latter performs along a curve on a unit sphere from $\tilde{z}_1$ to $\tilde{z}_2$. Note that $\theta$ is the angle between $\tilde{z}_1$ and $\tilde{z}_2$ and $\omega$ is a controllable parameter, ranging from 0 to 1.

## VISUALIZATION FOR MNIST

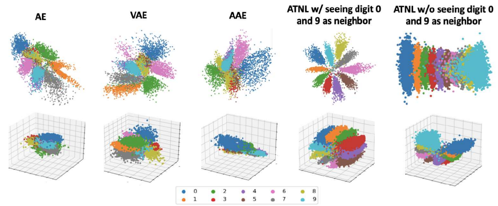


**Figure 5:** t-SNE visualization on Multi-PIE. The first and second rows show 2D and 3D visualizations of $z$ produced by AE, VAE, AAE and our ATNL, respectively.

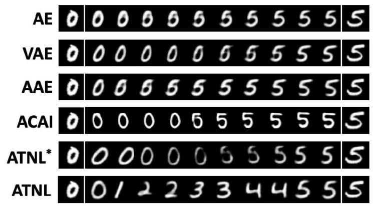


**Figure 6:** Image manipulation via linear or spherical semantic interpolation on MNIST. Given two input images in the first and last columns, we present the intermediate images of AE, VAE, AAE, ACAI, ATNL* and ATNL. Note that the first five models perform linear interpolation, while our ATNL performs the spherical one.

## QUANTITATIVE EVALUATION

**Table 1:** Image manipulation via linear or spherical semantic interpolation on MNIST. Note that the first five models perform linear interpolation, while our ATNL performs the spherical one.

| Dataset | Unsup. | | Sup. | |
|---|---|---|---|---|
| | AE | VAE | VAE-TL | **Ours** |
| MNIST | 94.63 | 96.42 | 98.46 | **99.14** |
| CMU Multi-PIE | 84.80 | 86.26 | 91.36 | **92.41** |

**Table 2:** Few-shot classification on MNIST with different data hallucination techniques. Note that the baseline denotes uses of traditional image variants, while ours applies our semantically interpolated images as hallucinated data

| Method | Classifiers | |
|---|---|---|
| | KNN | CNN |
| w/o data hallucination | 61.5 | 86.3 |
| w/ data augmentation (baseline) | 62.4 | 86.8 |
| w/ ATNL data augmentation (ours) | **78.7** | **94.4** |