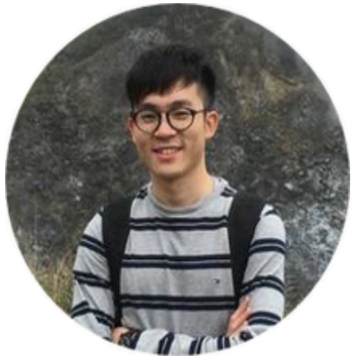
Semantics-Guided Representation Learning with Applications to Visual Synthesis

ICPR 2020


Jia-Wei Yan


Ci-Siang Lin


Fu-En Yang


Yu-Jhe Li

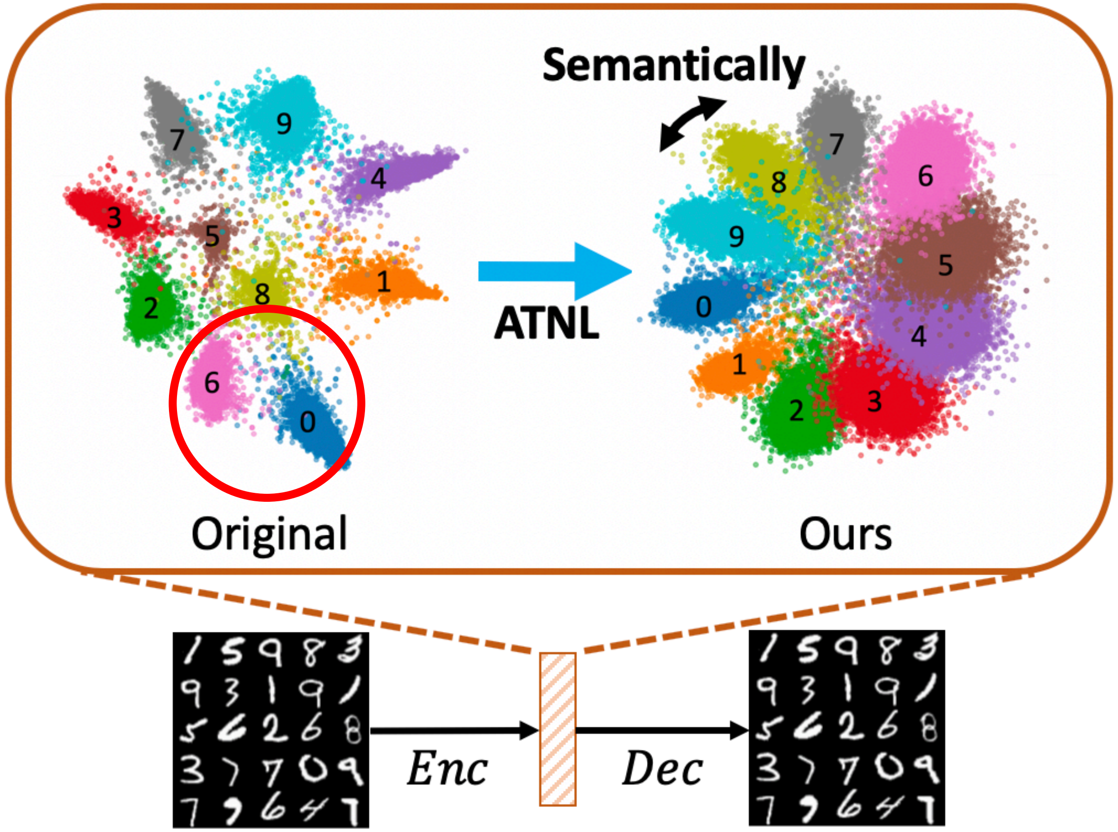

Yu-Chiang Frank Wang

Outline

- Introduction
- Semantics-Guided Representation Learning
  - VAE for representation learning
  - Angular Triplet-Neighbor Loss (ATNL)
  - Semantics-guided image generation
- Experiments
  - Visualization via t-SNE projection
  - Image generation
  - Quantitative evaluation
  - Analysis of ATNL
- Conclusion

Motivation

- To manipulate the latent representations which semantically match the images of interest (e.g., numerical order).

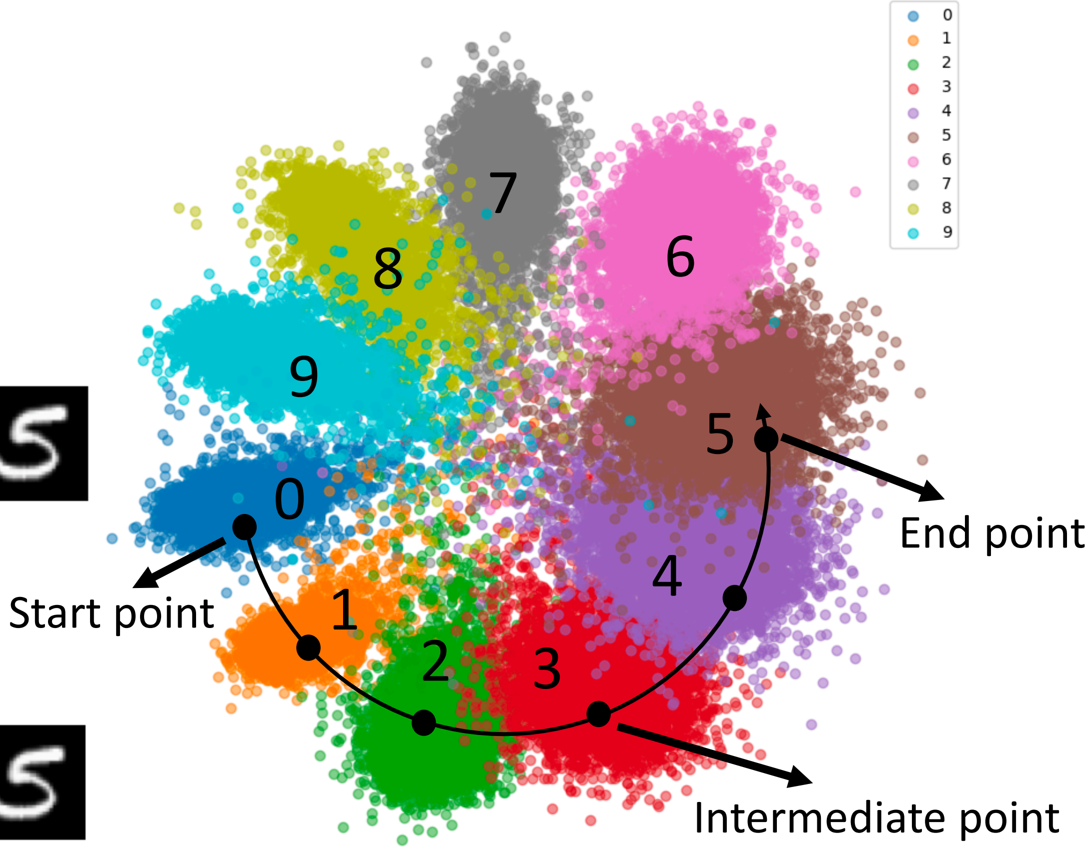


Application to Visual Synthesis

- Generate images which semantically match numerical order via **semantics-guided interpolation**.
  - 0 →1 → 2 → 3 → 4 → 5

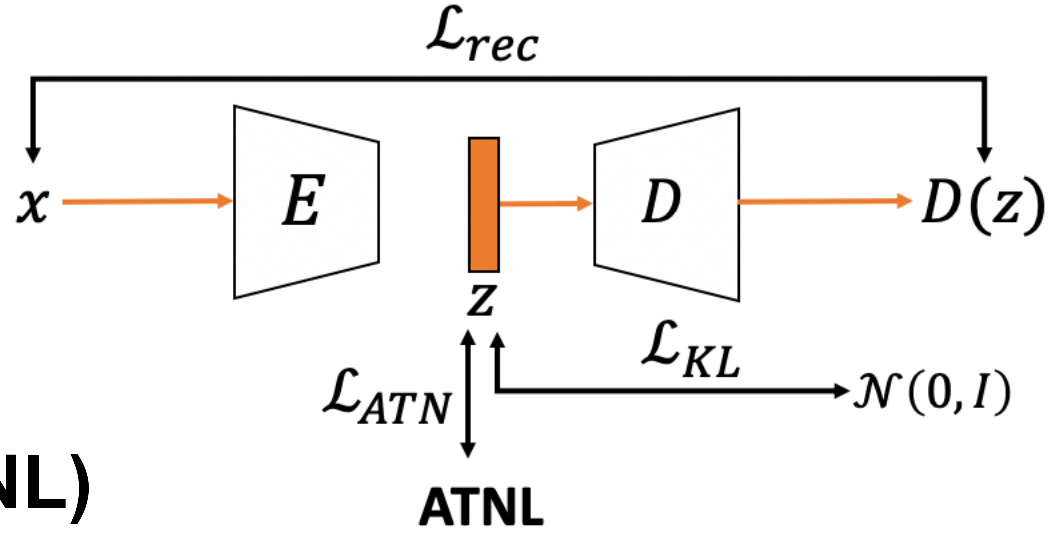
Linear interpolation

Semantics-guided interpolation



MNIST t-SNE

VAE for Representation Learning



- Reconstruction loss
  - $L_{rec} = |D(z) - x|$
- Kullback-Leibler (KL) divergence loss
  - $L_{KL} = E[KL(P(z) \parallel N(0,1))]$
- **Angular Triplet-Neighbor Loss (ATNL)**
  - $L_{ATN} = \sum_{i=1}^{N} \max(\cos^{-1}(\tilde{z}_i^a \cdot \tilde{z}_i^p) - \cos^{-1}(\tilde{z}_i^a \cdot \tilde{z}_i^n) + m_a, 0)$
- Total loss
  - $L_{total} = \lambda_1 \cdot L_{rec} + \lambda_2 \cdot L_{KL} + \lambda_3 \cdot L_{ATN}$

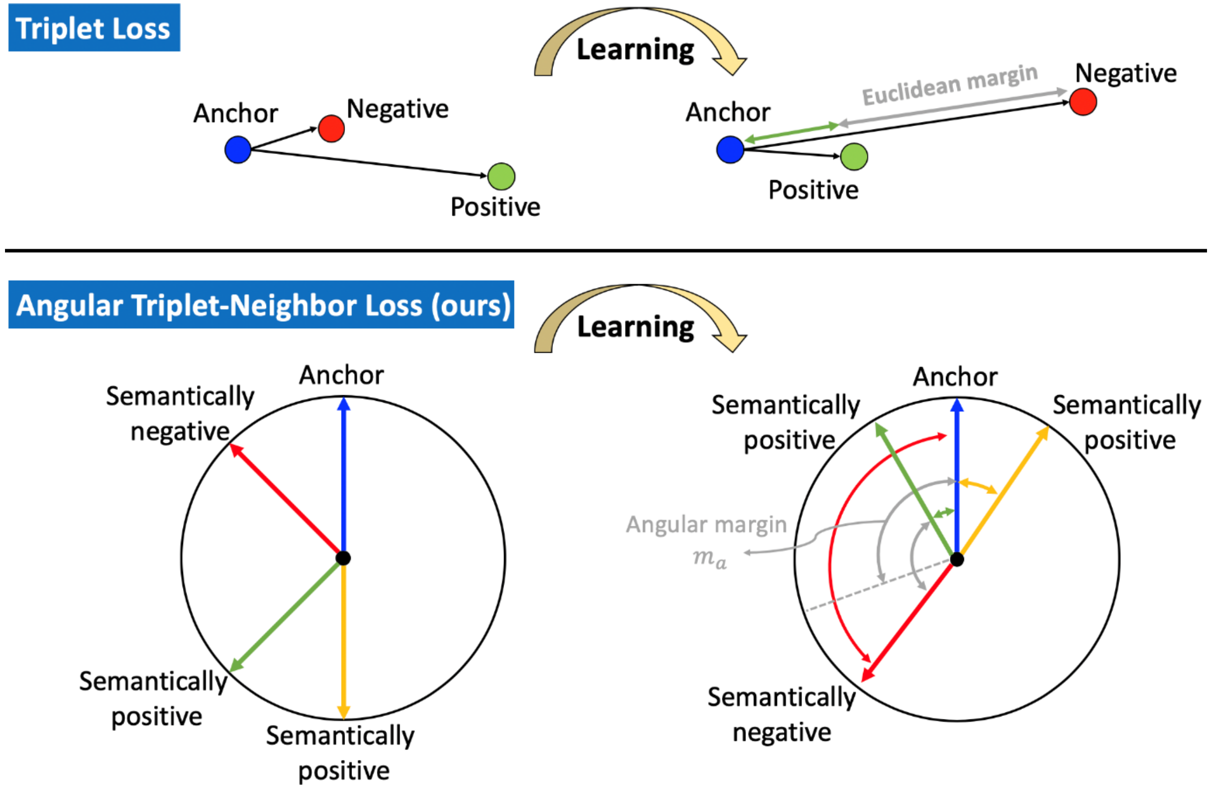
[Ref: Kingma et al. "Auto-Encoding Variational Bayes" *Arxiv*'13.]

Why Use Angular-based Triplet Loss Instead of Traditional Triplet Loss?

- Angular margin provides geometric interpretation as it corresponds to the angular distance on the unit-sphere, with a fixed range of $[0, \pi]$.

Euclidean-based

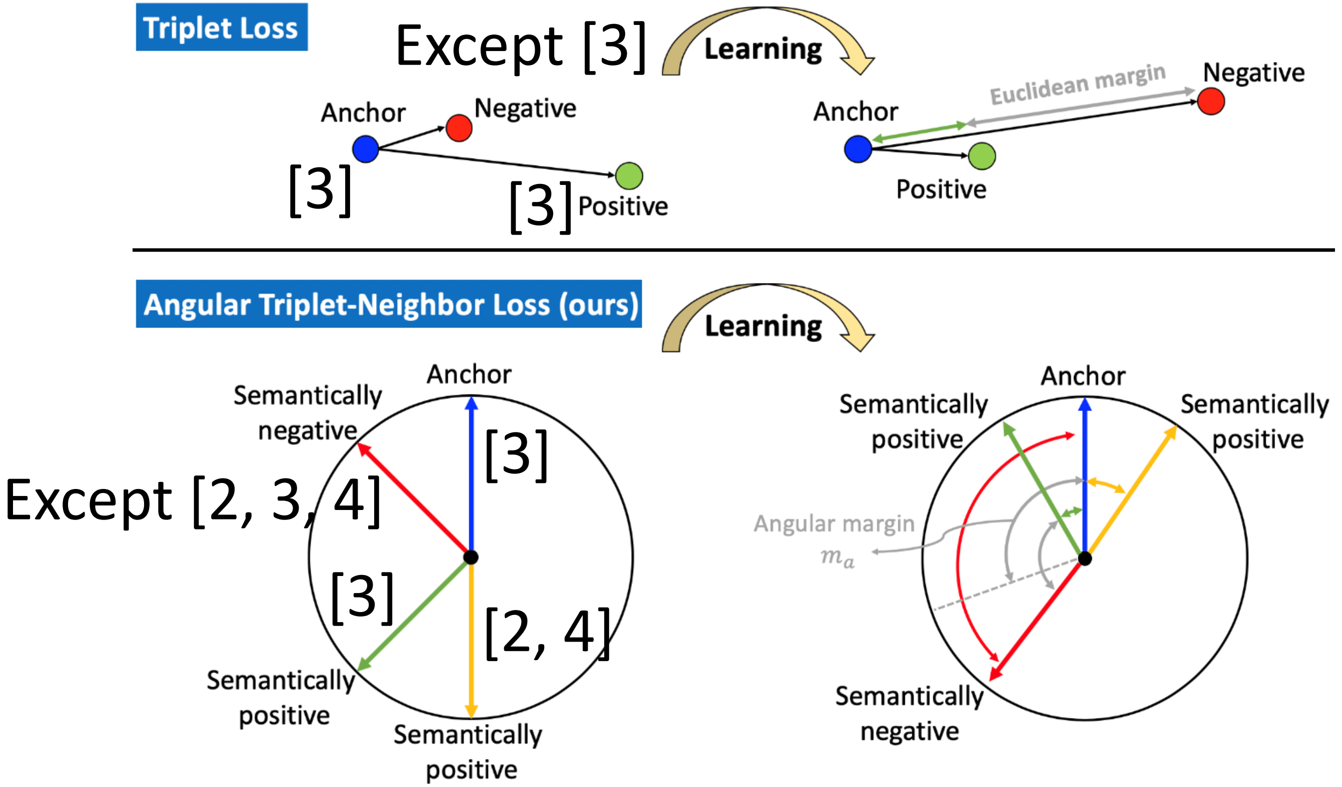
Angular-based



[Ref: Shoemake, Ken. "Animating rotation with quaternion curves." *ACM SIGGRAPH computer graphics*. Vol. 19. No. 3. ACM, 1985.]

Angular Triplet-Neighbor Loss (ATNL)

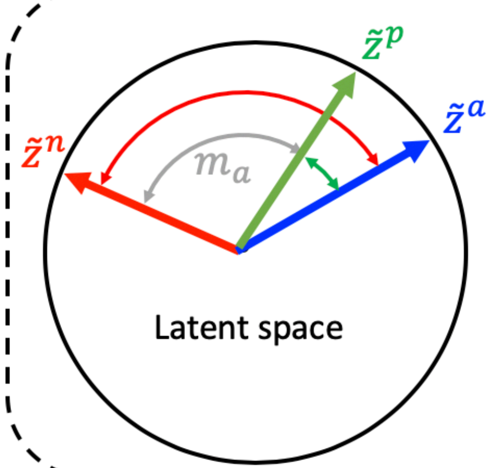
- Comparison between the definition of the original triplet loss and our developed ATNL



[Ref: Schroff et al. "FaceNet: A Unified Embedding for Face Recognition and Clustering." CVPR'15.]

Angular Triplet-Neighbor Loss (ATNL)

- Equation for satisfying ATNL
  - $\cos^{-1}(\tilde{z}^a \cdot \tilde{z}^p) + m_a \leq \cos^{-1}(\tilde{z}^a \cdot \tilde{z}^n)$
    $\forall(\tilde{z}^a, \tilde{z}^p, \tilde{z}^n) \in T_p$
- ATNL
  - $L_{ATN} = \sum_{i=1}^{N} \max(\cos^{-1}(\tilde{z}_i^a \cdot \tilde{z}_i^p) - \cos^{-1}(\tilde{z}_i^a \cdot \tilde{z}_i^n) + m_a, 0)$



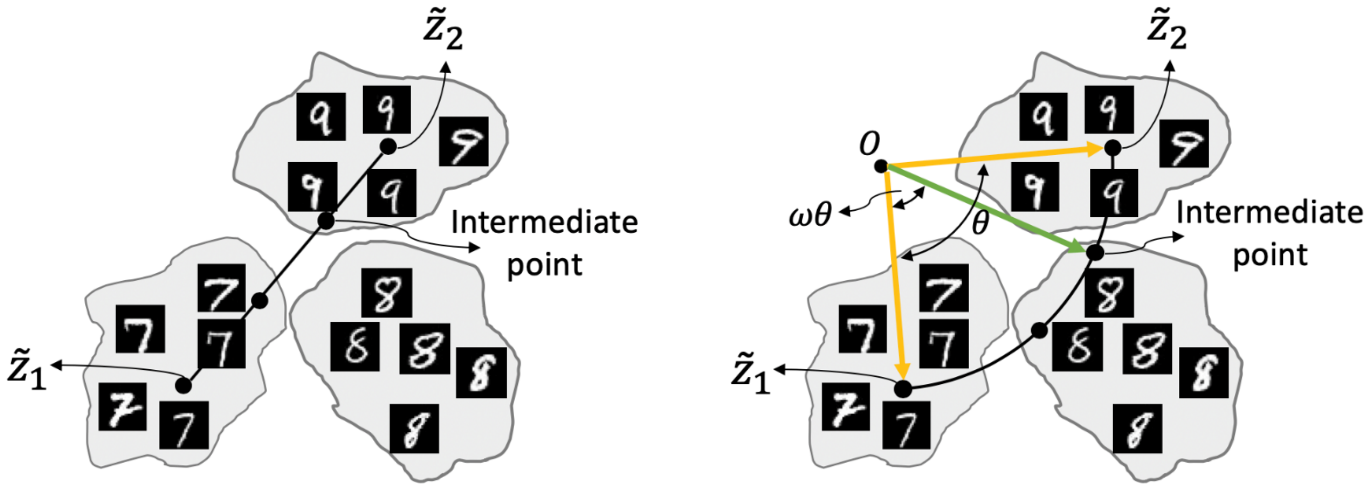
$\tilde{z}^a$: anchor [3]
$\tilde{z}^p$: semantically positive [2, 3, 4]
$\tilde{z}^n$: semantically negative Except [2, 3, 4]
m_a: angular margin

$\cos^{-1}(\tilde{z}^a \cdot \tilde{z}^p) + m_a$
$\leq \cos^{-1}(\tilde{z}^a \cdot \tilde{z}^n)$

[Ref: Schroff et al. "FaceNet: A Unified Embedding for Face Recognition and Clustering." CVPR'15.]

Semantics-Guided Image Generation

- Linear interpolation
  - $\tilde{z}_l(\tilde{z}_1, \tilde{z}_2, \omega) = (1-\omega)\tilde{z}_1 + \omega\tilde{z}_2$
- Spherical semantic interpolation
  - $\tilde{z}_s(\tilde{z}_1, \tilde{z}_2, \omega) = \frac{\sin(1-\omega)\theta}{\sin\theta}\tilde{z}_1 + \frac{\sin\omega\theta}{\sin\theta}\tilde{z}_2$



Linear interpolation   Spherical semantic interpolation

[Ref: Shoemake, Ken. "Animating rotation with quaternion curves." *ACM SIGGRAPH computer graphics*. Vol. 19. No. 3. ACM, 1985.]

Datasets

- **MNIST**
  - 60,000/10,000 training/testing handwritten digit images of 10 classes.
  - All images are resized from 28x28 to 32x32 in our experiments.

Examples of MNIST

[Ref: Lecun et al. ”Gradient-based learning applied to document recognition." IEEE, 1998.]

Datasets

- **CMU Multi-PIE**
  - Face images with viewpoint, illumination and expression variations.
  - Use a subset of CMU Multi-PIE with viewpoint variants (24,402 images with 7 viewpoints from -90° to 90° per 30°).
  - All images are resized from 128x128 to 64x64 in our experiments

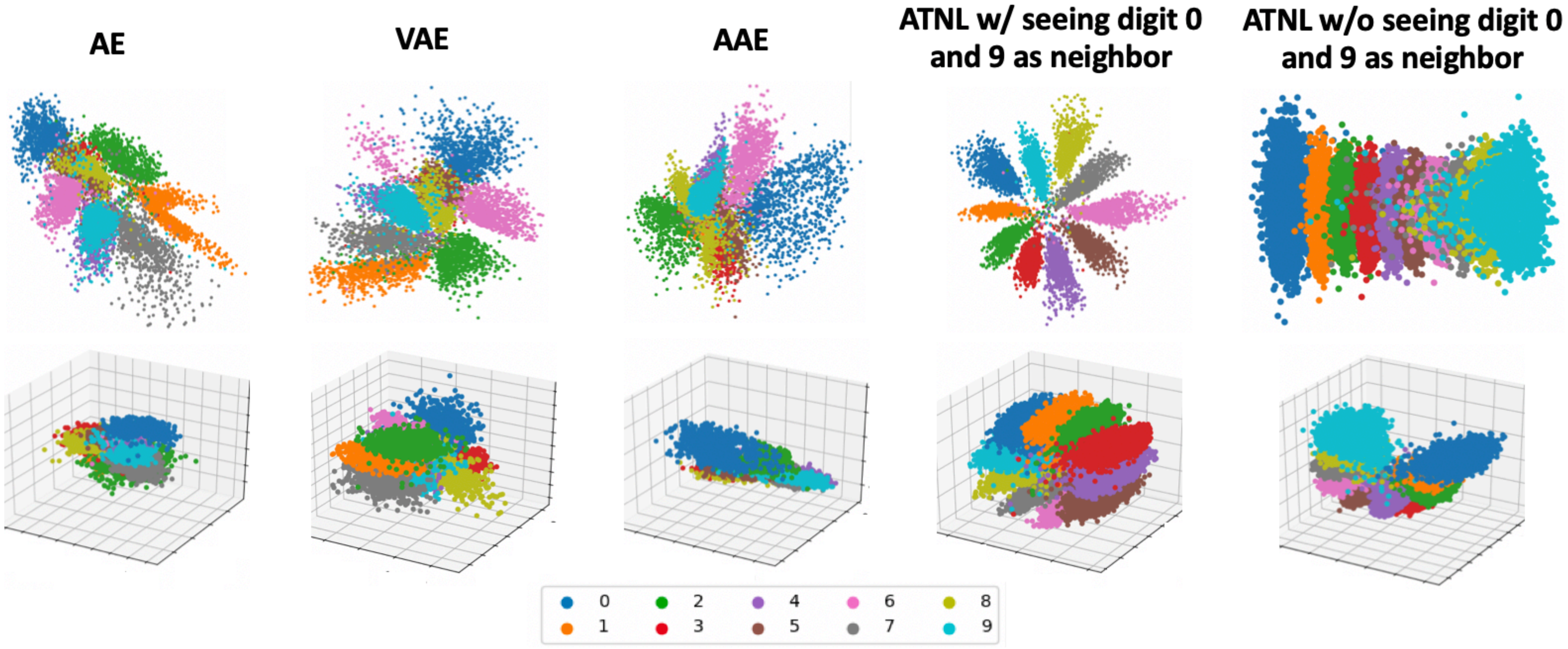
Examples of CMU Multi-PIE

[Ref: Gross et al. ”Multi-pie." Image and Vision Computing, 2010.]

Implementation Details

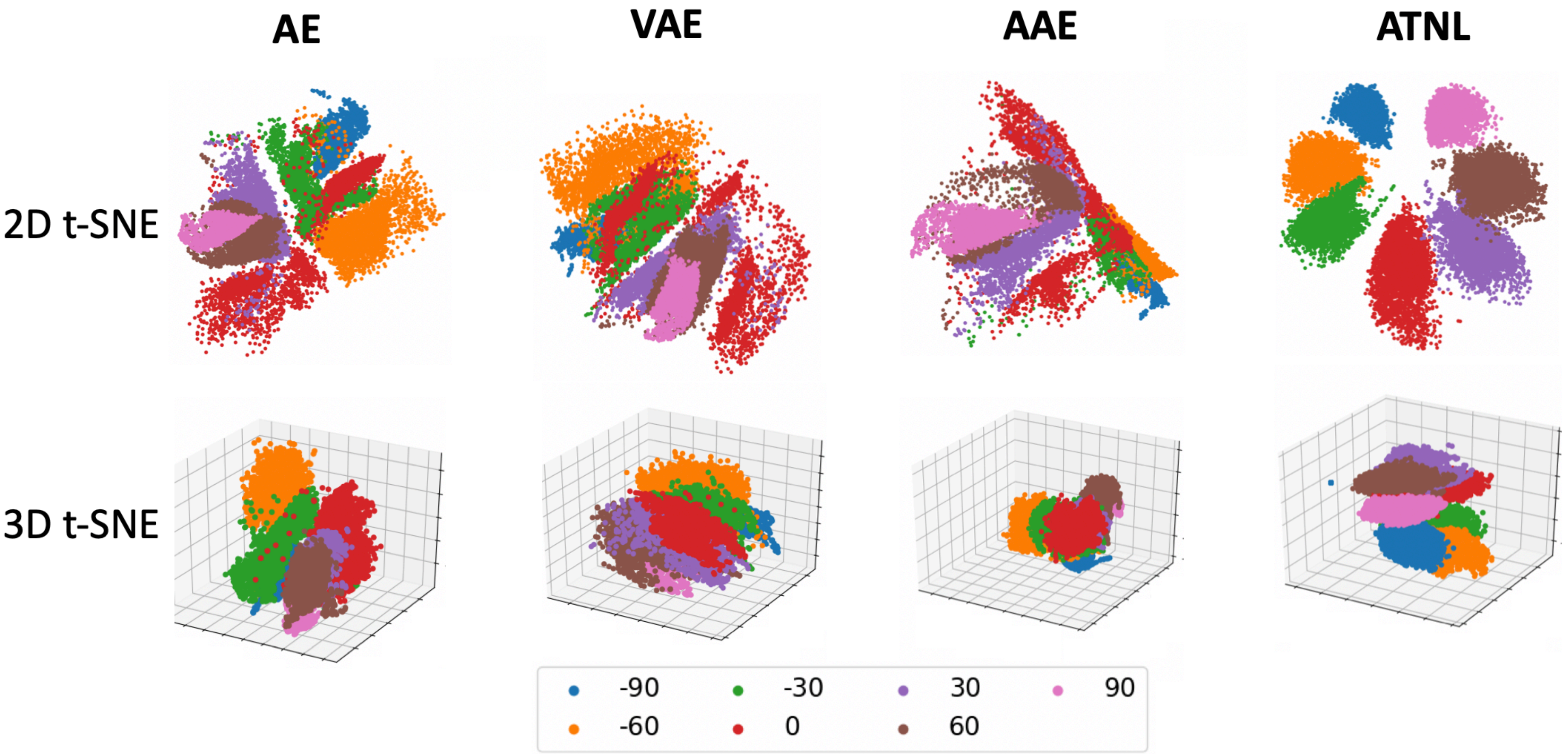
- **Architecture**
  - Encoder: 3 convolution layers followed by 3 fully connected layers
  - Decoder: 3 fully connected layers followed by 3 transpose convolution layers
- **Initialization**: all randomly initialized.
- **Optimizer**: Adam, the learning rate: 0.001
- **Batch size**: 256 for MNIST, 16 for CMU Multi-PIE
- The **margin** m_a for ATNL
  - 1.2($\approx 72°$) for MNIST
  - 0.9 ($\approx 50°$) for CMU Multi-PIE.
- **Hyper-parameters**: $\lambda_1(L_{rec}) = 10, \lambda_2(L_{KL}) = 1e-4, \lambda_3(L_{ATN}) = 1$
- **Run time**: Take about 3 hours on a single NVIDIA GeForce GTX 1080Ti GPU with 11 GB memory.

Visualization via t-SNE Projection (MNIST)



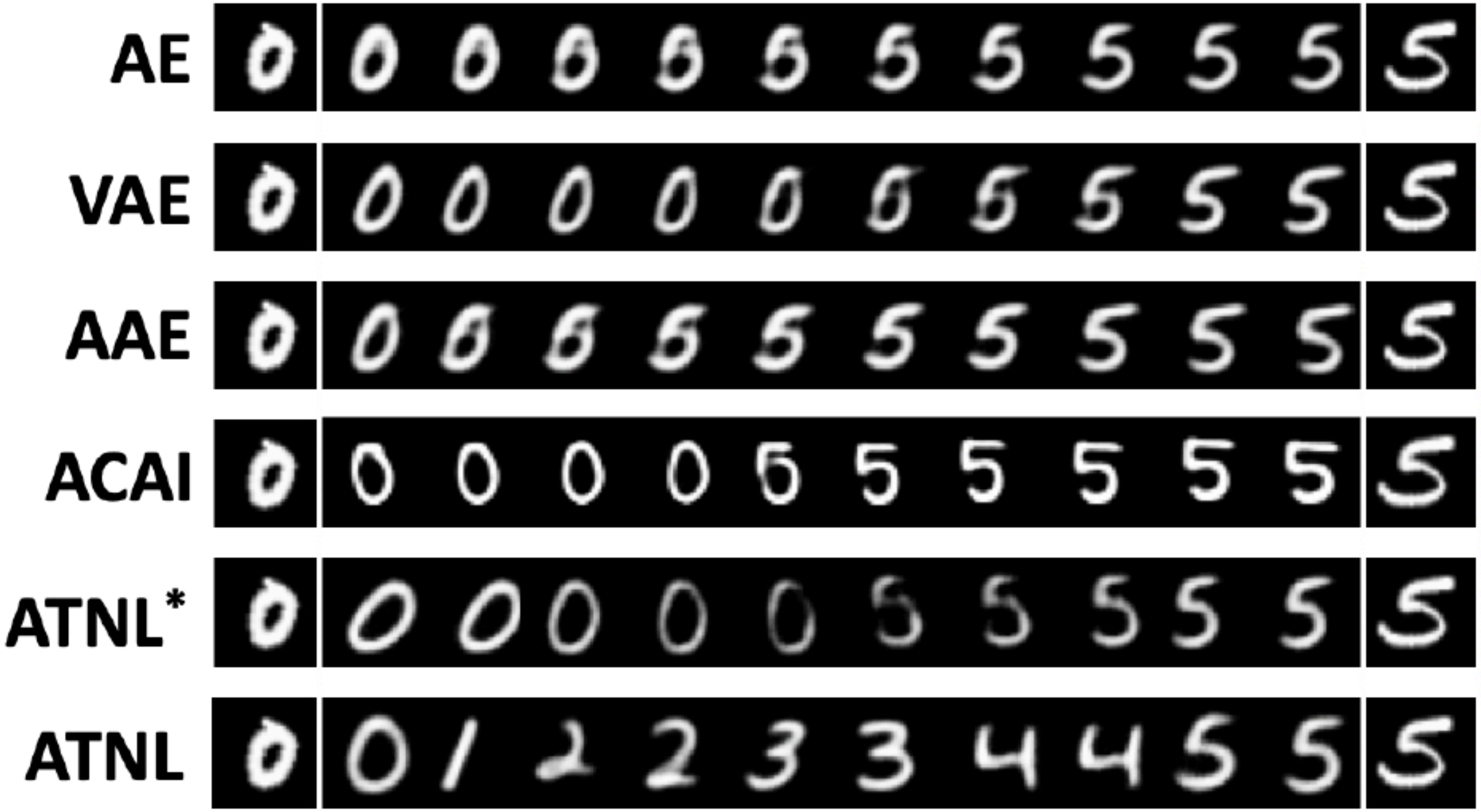
[Ref: Lecun et al. "Gradient-based learning applied to document recognition." IEEE, 1998.]
[Ref: Kingma et al. "Auto-Encoding Variational Bayes" *Arxiv'*13.]
[Ref: Makhzani et al. "Adversarial Autoencoders" *Arxiv'*15.]
[Ref: Berthelot et al. "Understanding and Improving Interpolation in Autoencoders via An Adversarial Regularizer" *ICLR'19*.]

Visualization via t-SNE projection (CMU Multi-PIE)



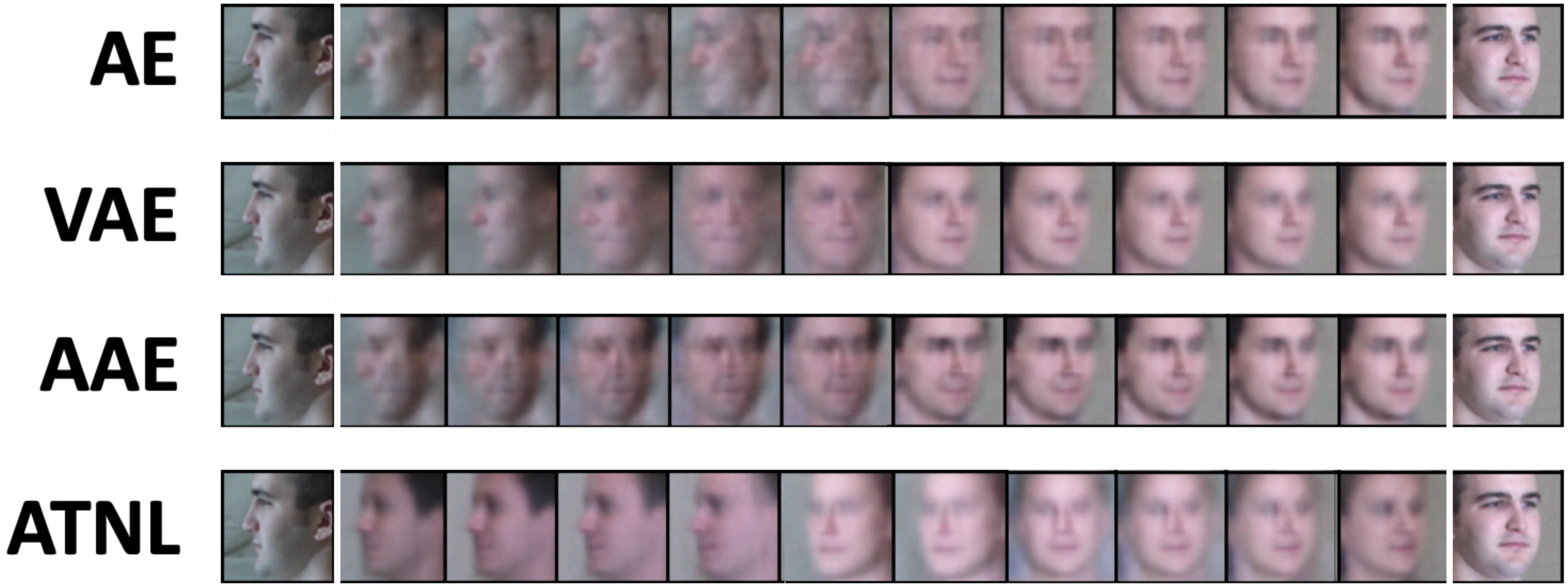
[Ref: Gross et al. ”Multi-pie." Image and Vision Computing, 2010.]
[Ref: Kingma et al. ”Auto-Encoding Variational Bayes" *Arxiv’*13.]
[Ref: Makhzani et al. ”Adversarial Autoencoders" *Arxiv’*15.]

Image Generation via Spherical Semantic/Linear Interpolation (MNIST)



[Ref: Lecun et al. ”Gradient-based learning applied to document recognition." IEEE, 1998.]
[Ref: Kingma et al. ”Auto-Encoding Variational Bayes" *Arxiv’*13.]
[Ref: Makhzani et al. ”Adversarial Autoencoders" *Arxiv’*15.]
[Ref: Berthelot et al. ”Understanding and Improving Interpolation in Autoencoders via An Adversarial Regularizer" *ICLR’19*.]

Image Generation via Spherical Semantic/Linear Interpolation (CMU Multi-PIE)

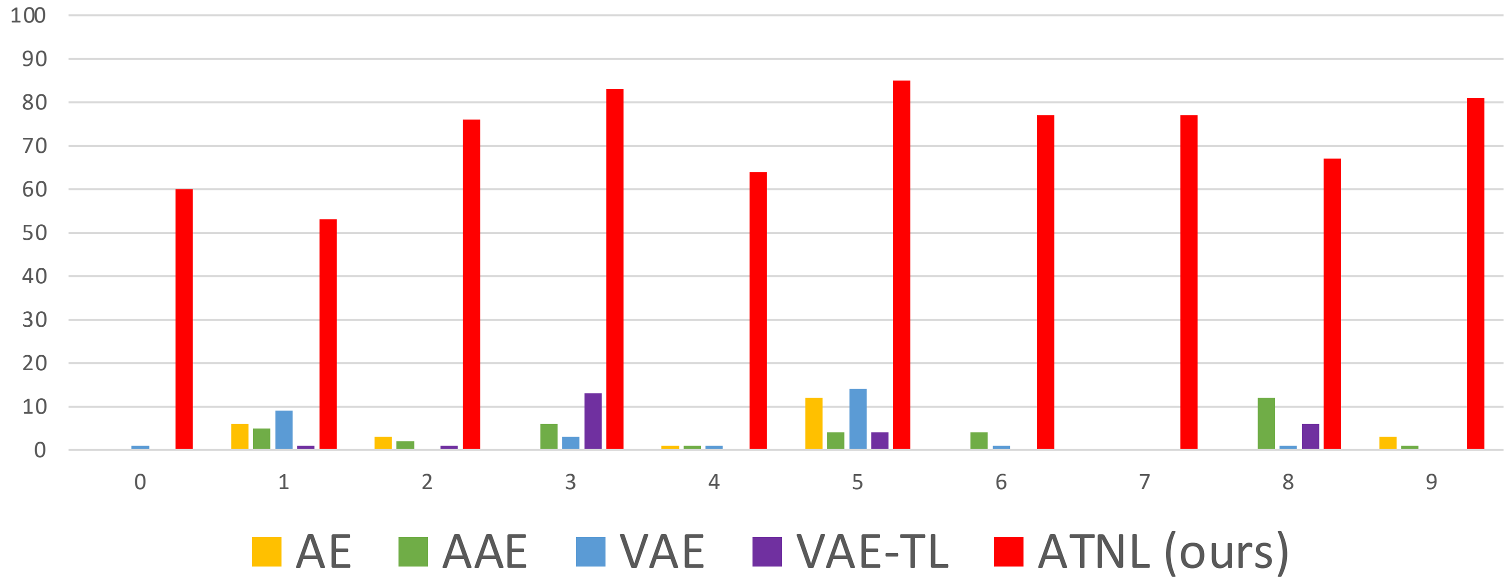


[Ref: Gross et al. ”Multi-pie." Image and Vision Computing, 2010.]
[Ref: Kingma et al. ”Auto-Encoding Variational Bayes" *Arxiv’*13.]
[Ref: Makhzani et al. ”Adversarial Autoencoders" *Arxiv’*15.]

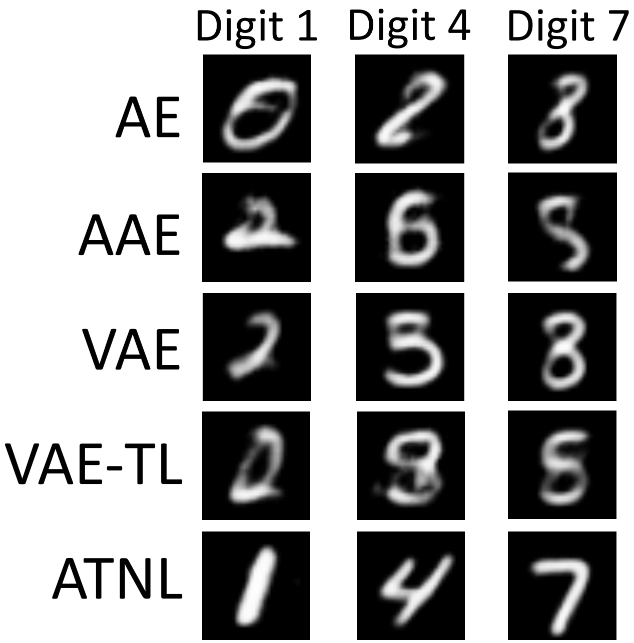
Quantitative Evaluation

- Classification performances of interpolated images on MNIST using different models.

Classification of interpolated images on MNIST



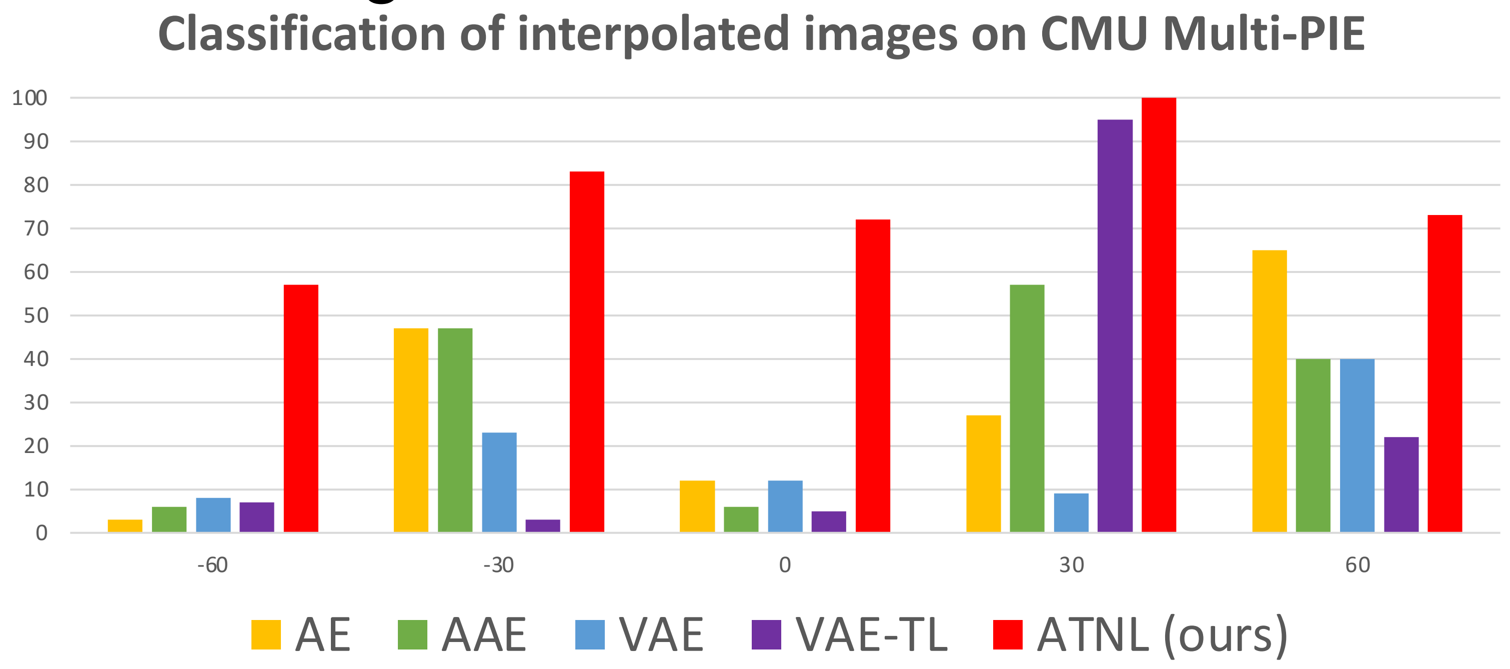
Examples of interpolated images

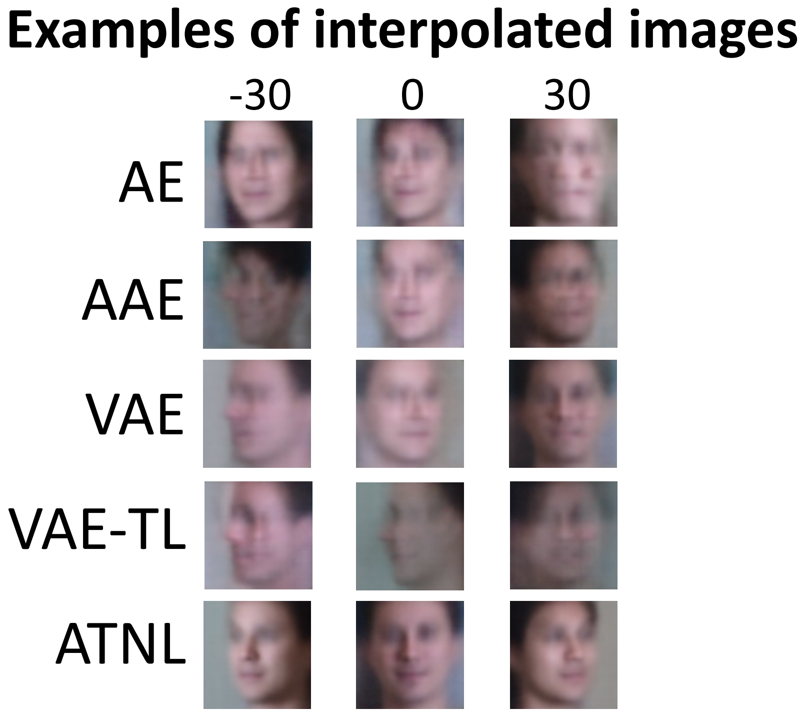


[Ref: Lecun et al. ”Gradient-based learning applied to document recognition." IEEE, 1998.]
[Ref: Gross et al. ”Multi-pie." Image and Vision Computing, 2010.]
[Ref: Kingma et al. ”Auto-Encoding Variational Bayes" *Arxiv'*13.]
[Ref: Schroff et al. ”FaceNet: A Unified Embedding for Face Recognition and Clustering." CVPR'15.]

Quantitative Evaluation

- Classification performances of interpolated images on CMU Multi-PIE using different models.





[Ref: Lecun et al. ”Gradient-based learning applied to document recognition." IEEE, 1998.]
[Ref: Gross et al. ”Multi-pie." Image and Vision Computing, 2010.]
[Ref: Kingma et al. ”Auto-Encoding Variational Bayes" *Arxiv’*13.]
[Ref: Schroff et al. ”FaceNet: A Unified Embedding for Face Recognition and Clustering." CVPR’15.]

Conclusion

- We are among the first to explore desirable semantic distribution of latent representations, based on the visual classification tasks of interest.
- We propose an Angular Triplet-Neighbor Loss (ATNL), which utilizes task-oriented semantic information for representation learning.
- With ATNL for semantics-guided representation learning, we are able to perform spherical semantic interpolation which produces desirable image outputs and allows satisfactory classification performances.