Unsupervised Identification of Protein Compositions and Conformations via Implicit Content-Transformation Disentanglement

Mostofa Rafid Uddin Jana Armouti Min Xu*

Carnegie Mellon University
Pittsburgh, PA 15213, USA.

mru@andrew.cmu.edu jarmouti@andrew.cmu.edu mxu1@cs.cmu.edu

1. Appendix

A. Appendix

Overview

- Appendix A.1 contains additional discussion on related works.
- Appendix A.2 contains an additional mathematical explanation of DualContrast and Baseline methods.
- Appendix A.3 contains additional details on the experiments.
- Appendix A.4 contains additional results and description on the datasets.

A.1. Detailed Related Works

Disentangled Representation Learning: PCA [10] and ICA [16] can be regarded as very preliminary work in the domain of disentangled representation learning. However, these methods assume linear subspace and do not work well for non-linear high-dimensional datasets. Deep learning-based approaches like Info-GAN [6], β-VAE [13], and their variants [4, 19, 20, 22] have recently gained wide attention as generic approaches for learning disentangled representations. Most of these works manipulated the variational bottleneck to achieve disentanglement of the latent codes. However, these works do not aim toward disentangling any specific factor, *e.g.*, content, group, style, transformation, etc., from the latent codes. Instead, they generate a series of images by traversing through each dimension of the latent space while keeping the remaining dimensions fixed. Thus, they infer the semantic meaning of each dimension of the learned latent factor. Consequently, these methods do not perform well in disentangling any specific generative factor compared to those that aim to disentangle several (two in most cases) specific generative factors [3, 35, 38]. Unlike these methods, our method specifically disentangles the content and transformation factors of data samples, whereas the content and transformation are defined in Section 3.1.

Unsupervised Content-style disentanglement: Apart from [39], several methods [23, 27, 30, 33, 34, 40] exist that perform unsupervised content-style disentanglement, focusing on natural images. Unlike these methods, our work primarily focuses on disentangling content and transformation in shape focused image datasets. Moreover, we do not depend on any ImageNet pretrained models as our images of interest differ greatly from the natural images of ImageNet. The very recent work by [30] assumes access to the style factors to disentangle that style from content in feature outputs from pretrained models. Unlike this work, we do not assume access to the transformations beforehand for disentanglement.

cryo-EM Heterogeneous Reconstruction: There exists several works on single particle cryo-EM and cryo-ET reconstruction, *e.g.*, cryoDRGN2 [42], cryoFIRE [25], cryoAI [24], etc. that performs amortized inference of transformation ($SO(3) \times d^2$) and latent space representing content. However, these works mainly focus on 2D-to-3D reconstruction instead of content-transformation disentanglement. Our work, on the contrary, focuses on content-transformation disentanglement. Though we use reconstruction loss to maximize informativeness of content and transformation factors, our reconstruction is 2D-to-2D or 3D-to-3D, unlike the aforementioned works. Also, our transformation factor is implicit and not explicitly limited to $SO(3) \times d^2$.

Shape Analysis: Disentangling content and transformation latent factors have special significance in the domain of shape analysis. Consequently, shape representation learning, modeling, and analysis [7, 15, 29, 31, 36, 43] are closely related to our work. Even for shape analysis, PCA can be regarded as one of the primitive methods. Even now, PCA is widely used in the shape analysis of protein complexes [2]. Recently, Huang et al. [15] demonstrated that

*Corresponding author

PCA with two components on the latent factor learned by an auto-encoder corresponds to content (shape) and style (pose) in 3D human mesh datasets. Nevertheless, PCA is a linear method assuming linear subspaces, which often does not hold true. A line of shape analysis research [1, 7, 36, 43] has been performed for non-linear disentanglement of content and style factors in 3D mesh or point cloud datasets. The goal of these works is to reduce per-vertex reconstruction loss of 3D meshes or point clouds for content-style-specific generation. These works claim unsupervised disentanglement as they do not require ground truth factors. However, they use the identity information of meshes, which is directly associated with content code. In contrast, our method does not require identity information apriori to learn latent codes specific to shape and code. Moreover, the mentioned works specifically investigate mesh-specific geodesic losses to achieve minimal per-vertex mesh reconstruction. On the other hand, we do not specifically aim to design mesh-specific losses in this work, rather, we propose a generic content-transformation disentanglement approach that can be applied to 3D mesh datasets with necessary modification in the model architecture.

A.2. Method

A.2.1. Evidence Lower Bound (ELBO)

Evidence lower bound (ELBO) is the objective that variational autoencoder (VAE)s optimize during training. The goal of a VAE is to approximate the true data distribution $p(x)$ using a latent variable model with a prior $p(z)$ and a likelihood $p_\theta(\mathbf{x}|\mathbf{z})$. However, computing the exact log-likelihood is intractable:

$$\log p(\mathbf{x}) = \log \int p_\theta(\mathbf{x}|\mathbf{z}) p(\mathbf{z})\, d\mathbf{z}$$

To address this, a lower bound on $\log p(\mathbf{x})$ is optimized, referred to as the ELBO. Let $q_\psi(\mathbf{z}|\mathbf{x})$ be the approximate posterior (learned by the encoder). The ELBO is defined as:

$$\text{ELBO}(\mathbf{x}) = \mathbb{E}_{q_\psi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - D_{\text{KL}}(q_\psi(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z}))$$

Since, we have two independent latent factors- content factor $\mathbf{c}$ and transformation factor $\mathbf{z}$, the objective can be written as:

$$\begin{aligned} p(\mathbf{x}) \geq & \mathbb{E}_{q_\psi(\mathbf{c},\mathbf{z}|\mathbf{x})}[\log \frac{p_\theta(\mathbf{x}|\mathbf{c},\mathbf{z})}{q_\psi(\mathbf{c},\mathbf{z}|\mathbf{x})}] \\ & = \mathbb{E}_{q_\psi(\mathbf{c},\mathbf{z}|\mathbf{x})} \log p_\theta(\mathbf{x}|\mathbf{c},\mathbf{z}) - KL(q_\psi(\mathbf{c}|\mathbf{x})||p(\mathbf{c})) \\ & - KL(q_\psi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \end{aligned} \tag{1}$$

A.2.2. Content-Transformation disentanglement with variational autoencoders (VAE)

A standard Variational Autoencoder (VAE) presumes data $\mathbf{x}$ to be generated by latent variable $\mathbf{z}$, whereas a standard Gaussian prior is assumed for $\mathbf{z}$.

$$\begin{aligned} p(\mathbf{x}) &= \int p(\mathbf{x}|\mathbf{z}) p(\mathbf{z}) d\mathbf{z} \\ p(\mathbf{z}) &= \mathcal{N}(0, \mathbf{I}_d) \end{aligned}$$

We extended the standard VAE to a two-latent variable setting. We assume latent variables $\mathbf{z}$ and $\mathbf{c}$ to generate the data $\mathbf{x}$.

$$p(\mathbf{x}) = \iint p(\mathbf{x}|\mathbf{z},\mathbf{c}) p(\mathbf{z}) p(\mathbf{c}) d\mathbf{c}$$

This setting is similar to VITAE [35], SpatialVAE [3], and Harmony [38]. However, in SpatialVAE [3] and Harmony [38], latent factor $\mathbf{z}$ is restricted as rotation and parameterized transformations, respectively. In SpatialVAE,

$$p(\mathbf{z}) = \text{Unif}(a, b) \tag{2}$$

$$\theta \sim p(\mathbf{z}|\mathbf{x}) \tag{3}$$

$$\mathbf{x}_{\text{cord}} = \mathcal{R}([-1, 1]^{d \times d}; \theta) \tag{4}$$

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{c}, \mathbf{x}_{\text{cord}}) p(\mathbf{c}) d\mathbf{c} \tag{5}$$

where a and b are specified constants, θ are transformation (2D rotation and translation) parameters, $\mathcal{R}$ is the corresponding transformation operator.

On the other hand, in Harmony,

$$\begin{aligned} \theta &= \mathbb{I}(\mathbf{z}|\mathbf{x}) \\ \mathbf{x}' &= \mathcal{T}(x; \theta) \\ p(\mathbf{x}') &= \int p(\mathbf{x}'|\mathbf{c}) p(\mathbf{c}) d\mathbf{c} \end{aligned}$$

where $\mathbb{I}$ is an identity function, θ are transformation parameters and $\mathcal{T}$ is the corresponding transformation operator.

Unlike these two methods, in VITAE [35] and our proposed DualContrast, we use standard Gaussian priors for latent codes $\mathbf{z}$ and $\mathbf{c}$.

$$\begin{aligned} p(\mathbf{z}) &= \mathcal{N}(0, \mathbf{I}_d) \\ p(\mathbf{c}) &= \mathcal{N}(0, \mathbf{I}_d) \end{aligned}$$

However, in VITAE [35], $\mathbf{z}$ is used to explicitly sample continuous piecewise affine velocity (CPAB) transformation parameter θ, and $\mathbf{c}$ is used to sample appearance sam-

ples $\mathbf{x}'$.

$$\theta \sim p(\mathbf{x}|\mathbf{z})$$
$$\mathbf{x}' \sim p(\mathbf{x}|\mathbf{c})$$
$$\mathbf{x} = \mathcal{T}(\mathbf{x}'; \theta)$$

where $\mathcal{T}$ is the transformation operator for CPAB transformation. CPAB transformation parameter is highly expressive compared to affine transformation parameters used in spatialVAE.

Contrary to VITAE [35], we do not use $\mathbf{z}$ to sample any transformation parameters explicitly; rather use both $\mathbf{z}$ and $\mathbf{c}$ to generate $\mathbf{x}$. To this end, we use a contrastive learning strategy that is described in Section 3 of the main paper. Without explicitly sampling any transformation parameter, we improve the expressiveness of our transformation latent factor $\mathbf{z}$ even more than the CPAB transformation used in VITAE [35].

A.2.3. Feature Suppression of SimCLR and MoCo contrastive losses:

The contrastive losses used in popular self-supervised learning methods SimCLR [5] or MoCo [12] as did not help much in disentangling content and transformation in our experiments. It has been demonstrated that these methods often learn nuisance image features or noise to obtain a shortcut solution to the contrastive objective [18]. This phenomenon is referred to as *feature suppression* of contrastive objectives. We found that using reconstructive loss was necessary to prevent the feature suppression problem.

A.2.4. Choice of Transformation to create Contrastive pairs

We leveraged different transformation functions $T(x)$ to create contrastive pairs in DualContrast for LineMod RGB object dataset. We used rotation, translation, scaling, blur, saturation, and hue as $T(x)$. We performed both qualitative (Table 1) and quantitative analysis on the effect of different $T(x)$ for content-transformation disentanglement in DualContrast. We observe that using Scale or Blur makes the transformation factor $\mathbf{z}$ uninformative of the data and it does not capture anything at all. Consequently, changing this $\mathbf{z}$ factor while generating images does not change the image at all for these two codes (Figure 1). On the other hand, using translation shows small negligible differences in the $\mathbf{c}$-$\mathbf{z}$ transfer-based image generation. Color-based transformations like saturation and Hue only change the color of the generated image, instead of affecting its shape-based transformation. Only rotation provides generalization of $\mathbf{z}$ and enables $\mathbf{z}$ to capture viewpoint transformations present in the dataset.

Table 1. Transformation Factors and Corresponding $D_{\text{score}}(c|z)$ and $D_{\text{score}}(c|c)$ values.

Transformation Factor	$D_{\text{score}}(c\|z)(\downarrow)$	$D_{\text{score}}(c\|c)(\uparrow)$
Rotation	0.48	0.95
Translation	0.65	0.98
Scale	0.52	0.91
Contrast	0.51	0.93
Saturation	0.71	0.86
Hue	0.61	0.88
Blur	0.47	0.92

A.3. Experiments

A.3.1. Implementation Details

We implemented our model in Pytorch (version 1.9.0). We used a convolutional neural network (CNN) (3 convolutional layers for MNIST, 4 for others) to implement the encoder and a fully connected network (FCN) (5 layers) to implement the decoder. For subtomograms, we used a 3D convolutional network for the encoder. We do not use any pooling layers in our networks.

While training the models, we use a batch size of 100 and an Adam optimizer with a learning rate of 0.0001. We used a linear learning rate scheduler that decays the learning rate of each parameter group by 0.1 every 50 epochs. We trained our models for 200 epochs. We used the same setting for our models and the baseline models. We used NVIDIA RTX A500 and AMD Radeon GPUs to train the models.

Choice of latent dimension: For Harmony [38] and SpatialVAE [3], the transformation latent factor is restricted to dimension 3. For VITAE [35], SimCLR [39], and our DualContrast, there is no such restriction on the transformation latent factor dimension and same dimension was used as the content factor. For all the methods, the dimension of the content factor was set as 10 for MNIST, 50 for subtomogram dataset. For hyperparameters $\gamma_{\mathbf{c}}$ and $\gamma_{\mathbf{z}}$, we set a very small value ($\rightarrow 0$) in our experiments. The values determine how strictly the content factor and the transformation factor should mimic the prior standard multivariate gaussian distribution.

Evaluation metrics in protein subtomograms: Apart from measuring disentanglement in the latent space, we also quantitatively assessed the latent space clustering performance and the quality of the structures obtained by coarsely refining each cluster of subtomograms with RELION. To measure the clustering performance, we used Adjusted Rand Index (ARI). The ARI is commonly used to measure the similarity between two data clusterings, correcting for chance. Given a contingency table where:

- n_{ij} is the number of objects in both cluster i of the ground

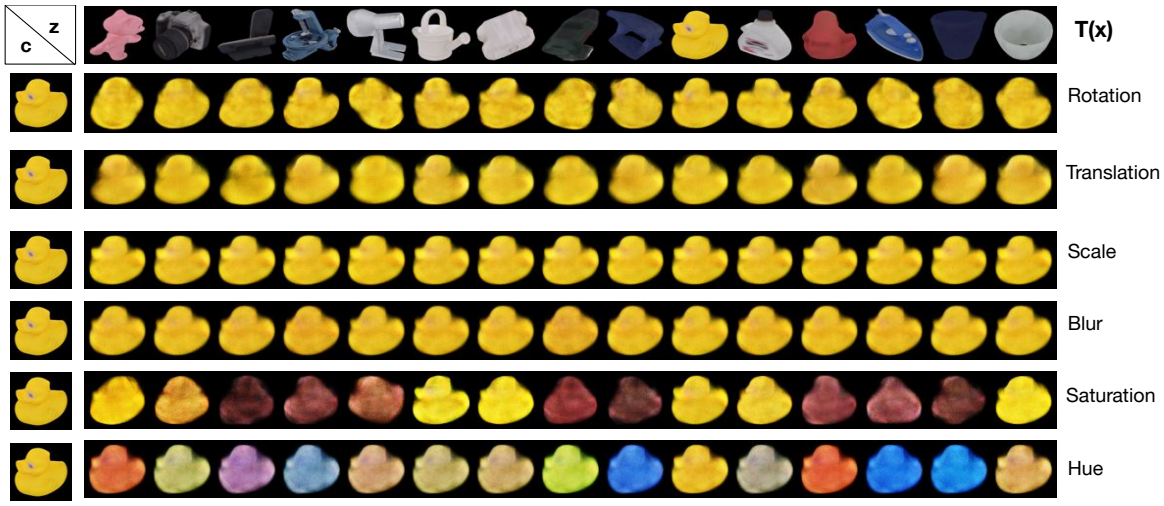


Figure 1. Content-transformation transfer based image generation results using different transformation functions ($T(x)$) to create contrastive pairs. Only rotation ensures transformation factor z to capture object viewpoints- the transformation factor present in the original dataset.

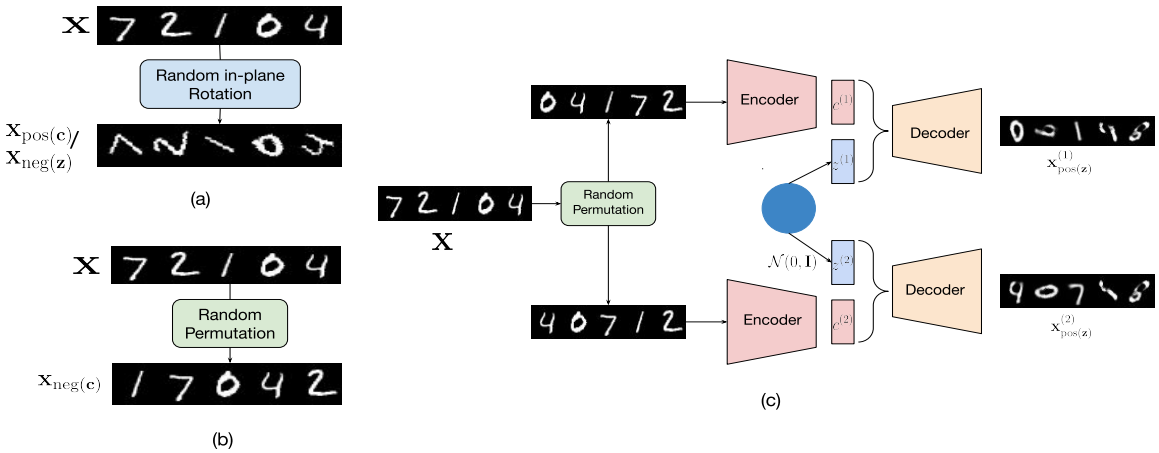


Figure 2. **Visualization of the creation of contrastive pairs for MNIST**. (a) Creating positive pair with respect to (w.r.t) content factor and negative pair w.r.t. transformation. (b) Creating negative pair w.r.t. content. (c) Creating positive pair w.r.t. transformation. We show the processes for a batch of MNIST digits with a batch size of 5.

truth and cluster j of the predicted labels,

- $a_i = \sum_j n_{ij}$ is the sum over row i,
- $b_j = \sum_i n_{ij}$ is the sum over column j,
- n is the total number of data points.

The ARI is defined as:

$$\text{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \frac{\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \frac{\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}}{\binom{n}{2}}}$$

To calculate ARI, we used the ADJUSTED_RAND_SCORE function from SKLEARN.METRICS.

To evaluate the quality of structures obtained by coarse refinement, we used AUC-FSC (Area Under the Curve of Fourier Shell Correlation), which has been adopted by the community to measure structure recovery performance in heterogeneous cryo-EM/ET datasets [17]. It is a scalar metric used to summarize the overall agreement between two 3D volumes in frequency space.

The **Fourier Shell Correlation (FSC)** measures the correlation between two 3D volumes in Fourier space as a function of spatial frequency s. It is defined as:

$$\text{FSC}(s) = \frac{\sum_{i \in s} F_1(i) \cdot F_2^*(i)}{\sqrt{\sum_{i \in s} |F_1(i)|^2 \cdot \sum_{i \in s} |F_2(i)|^2}}$$

where:

- $F_1(i)$ and $F_2(i)$ are the complex Fourier coefficients of the two volumes,
- $F_2^*(i)$ is the complex conjugate of $F_2(i)$,
- $i \in s$ denotes the voxels in the shell corresponding to spatial frequency s.

The **AUC-FSC** summarizes the FSC curve over the full frequency range $[0, 1]$ and is defined as:

$$\text{AUC-FSC} = \int_0^1 \text{FSC}(s)\, ds$$

In our case, we obtain a coarsely refined structure per each subtomogram class. For protein compositions, we determine 4 structure classes and for compositions, we obtain 6 structures. We calculate the AUC-FSC between the refined structure and its best matching structure template we used for simulating the dataset. Before calculating AUC-FSC, we manually align the template and coarsely-refined structures with ChimeraX [28]. We report the arithmetic average and minimum of the AUC-FSCs we obtain for all the structural classes.

A.4. Additional Results

A.4.1. MNIST

We use the commonly used MNIST dataset to initialize our experiments. MNIST is a dataset in the public domain that the research community has extensively used. It contains grayscale images of handwritten digits. Each image is of size 28×28. The training set contains 60,000 images, whereas the test set contains 10,000. We use the same train test split for our experiments.

Our quantitative results (Table 2) demonstrate the superior performance of DualContrast compared to the baselines. It also shows that the **z** factor in DualContrast contains information other than rotation (low $S(z)$), but is also separated from content (low $D(c|z)$ and high $D(c|c)$.

In qualitative evaluation of image generation with varying **c** and **z** codes in MNIST (Figure 3), we observed that-Harmony, SpatialVAE, and C-VITAE generated many images with erroneous content and transformations. However, DualContrast did not make such mistakes. Moreover, DualContrast visibly disentangled several conformation-like transformations. Harmony and SpatialVAE simply rotated the image, not capturing any other information about the transformation source. This is consistent with their high $S(z)$ score. VITAE somewhat represented the conformation with its explicit modeling of piecewise-linear transformation. Nevertheless, DualContrast most appropriately captured the transformation of the source digits. We also include tSNE embedding of the content codes inferred by the models on the MNIST test dataset associated with class labels (Figure 4). On the embedding space, DualContrast clearly shows superior clustering performance.

Table 2. Disentanglement metrics on MNIST

Method	$D(c \mid c)(\uparrow)$	$D(c \mid z)(\downarrow), S(z)(\downarrow)$	$SAP(c)(\uparrow)$
SpatialVAE	0.81	0.28,0.98	0.53
Harmony	0.82	0.31,1.00	0.51
SimCLR (Discriminative)	0.58	0.60,0.63	0.02
SimCLR (Generative)	0.53	0.67,0.61	0.14
VAE with 2 latent space	0.63	0.63,0.69	0.00
VITAE	0.77	0.32,0.72	0.45
DualContrast (w/o $L_{\text{con}(c)}$)	0.87	**0.21**,1.00	**0.66**
DualContrast (w/o $L_{\text{con}(z)}$)	0.79	**0.85**,0.60	0.06
DualContrast	**0.89**	0.31,0.75	0.58

A.4.2. LineMod

LineMod [14] dataset is originally designed for object recognition and 6D pose estimation. It contains 15 unique objects: 'ape', 'bench vise', 'bowl', 'cam', 'can', 'cat', 'cup', 'driller', 'duck', 'eggbox', 'glue', 'hole puncher', 'iron', 'lamp' and 'phone', photographed in a highly cluttered environment. We use a synthetic version of the dataset [41], which has the same objects rendered under different viewpoints. The dataset is publicly available at this url. The dataset is publicly available under MIT License.

This dataset has $1,313$ images per object category. We used $1,000$ images per category for training and used the remaining for testing. For many objects, the object region covers only a tiny part of the original image. To this end, we cropped the object region from the original image using the segmentation masks provided with the original dataset. After cropping the object regions, we padded 8 pixels to each side of the cropped image and then reshaped the padded image to the size of $(64, 64, 3)$. Thus, we prepared the training and testing datasets for content-transformation disentanglement in LineMod. We used the same dataset and train-test splits for our model and the baseline models. The associated processing codes are provided in the supplementary material.

We trained our proposed DualContrast, VITAE [35], SpatialVAE [3], and Harmony [38] on the LineMod dataset. We provide qualitative results of image generation with content-transformation transfer in Fig. 5 obtained with each method. It is noticeable that both Harmony [38] and SpatialVAE [3] have shown good performance when it comes to reconstruction. However, these two methods can only perform rotation and translation of the objects with explicitly defined transformations and can not capture complex transformations, e.g., projection, viewpoint change, etc. Compared to SpatialVAE and Harmony, VITAE [35] can perform better transformation transfer but performs poor re-

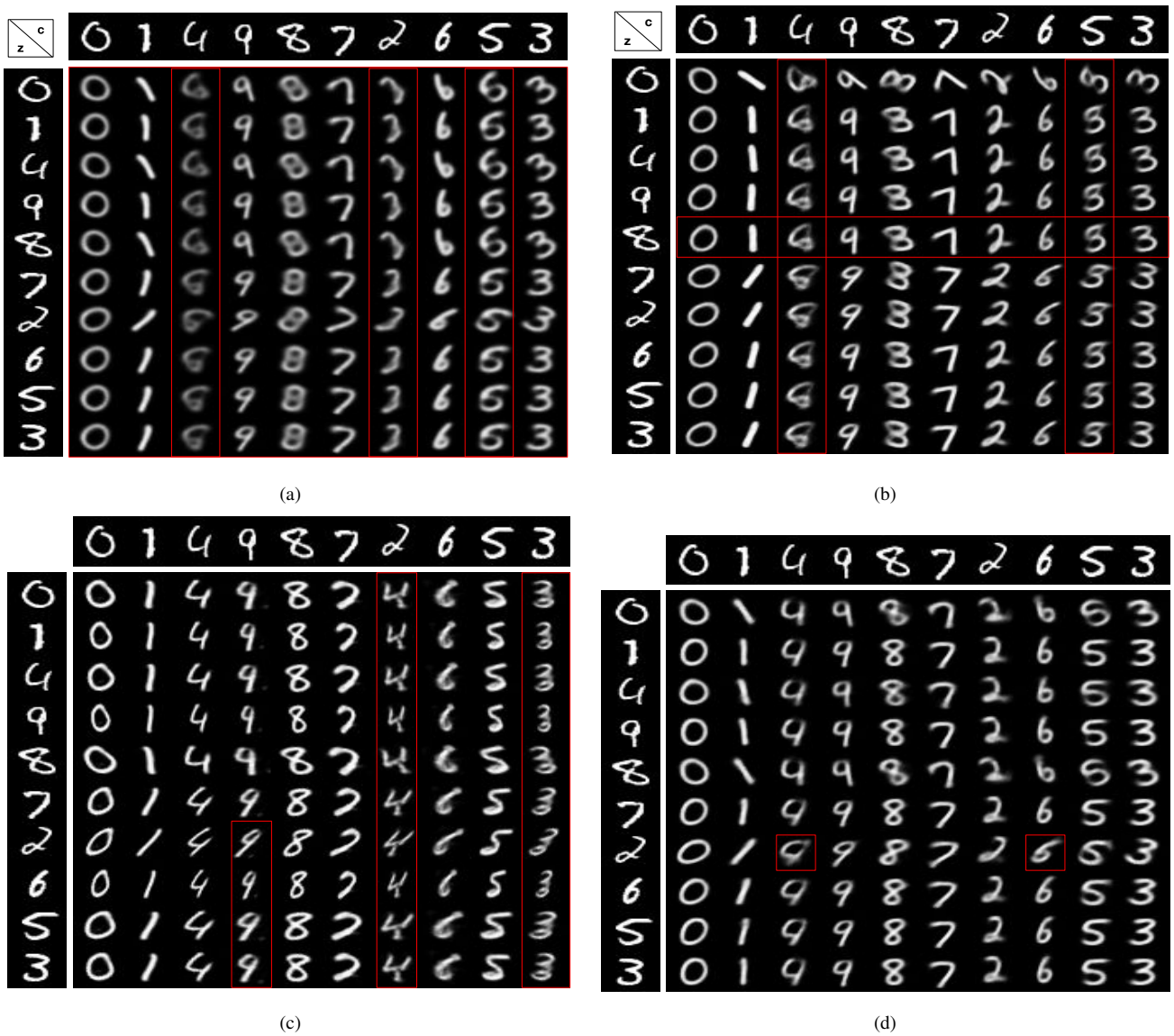


Figure 3. Content-transformation transfer results from ablation analysis (a), (b), (c), and (d) shows the results with Harmony [38], Spatial-VAE [3], VITAE [35], and DualContrast respectively. When generating image grids, the transformation factor is uniform across rows, and the content factor is uniform across columns. Erroneous generations (both in terms of content and transformation) are marked within red boxes.

construction. Nevertheless, DualContrast stands out for its superior ability to perform transformation transfer while ensuring optimal performance in reconstruction.

A.4.3. Protein Subtomogram Dataset

We created a realistic simulated cryo-ET subtomogram dataset of 18,000 subtomograms of size 32^3. The dataset consists of 4 protein classes of similar sizes- Nucleosomes, pre-fusion Sars-Cov-2 spike protein, post-fusion Sars-Cov-2 spike protein, and Fatty Acid Synthase Unit. These proteins are significantly different in terms of their composition, which determines their different identities. Moreover, structures for all of these three types of proteins have been resolved in cellular cryo-ET [8, 11, 21], which makes it feasible to use them for our study. Moreover, cellular cryo-ET is the primary method to capture these proteins inside the cells in their native state [9].

For each protein class (except post-fusion spike protein for which different conformations are not available), we collected different protein structures from the RCSB PDB website [32]. RCSB PDB is a web server containing the structure of millions of proteins. For nucleosomes, we collected PDB IDs '2pyo', '7kbe', '7pex', '7pey', '7xzy', and '7y00'. All of these are different conformations of nucleosomes that slightly vary in the spatial arrangement

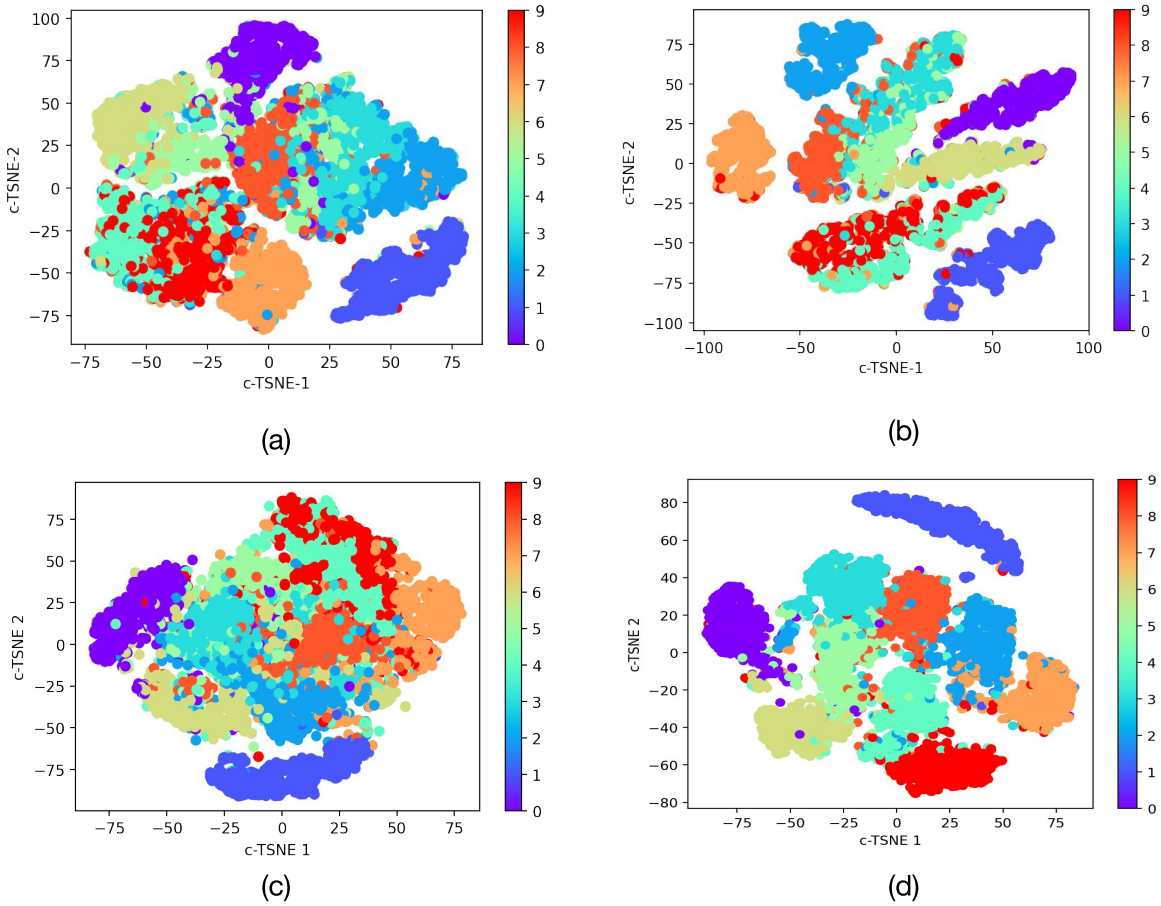
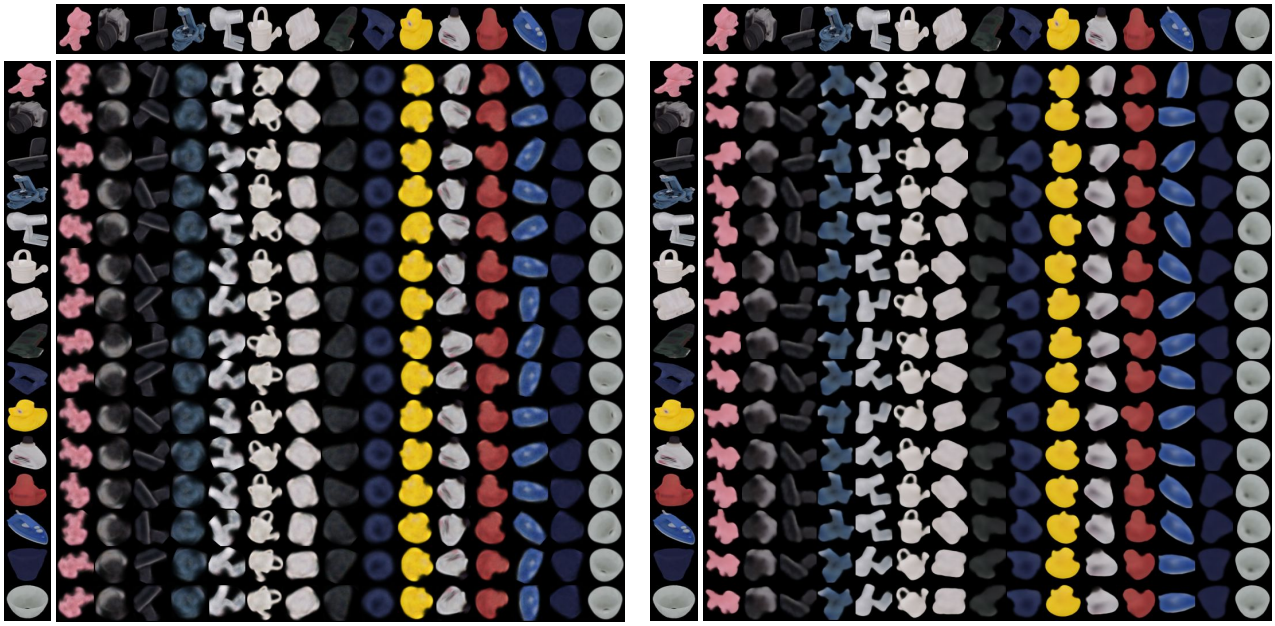


Figure 4. tSNE embedding plots of content latent factor learned by the unsupervised content-transformation disentanglement methods. (a), (b), (c), and (d) shows the results for Harmony [38], Spatial-VAE [3], C-VITAE [35], and DualContrast respectively. Overall, DualContrast shows superior performance.

of its DNA arms around the histone core (Figure 7). For sars-cov-2 spike proteins, we collected PDB IDs '6vxx', '6vyb', '6xr8', '6xra', '6zox', and '6zp0'. Among them, '6xra' is the post-fusion spike protein which is compositionally different from the others. The remaining PDB structures are highly similar and almost indistinguishable in 10 Åresolution. For Fatty Acid Synthase (FAS) Unit, we used PDB IDs '8prv', '8ps1', '8ps9', '8psj', '8psm', '8psp'. They also have highly similar structure and almost indistinguishable in 10 Åresolution.

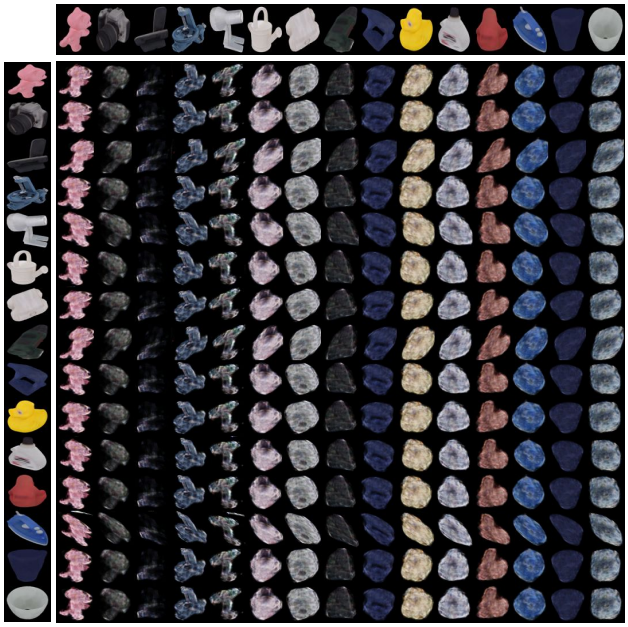
After collecting these 18 PDB structures as PDB files, we used EMAN PDB2MRC [37] to create density maps (as MRC file extension) from the PDB files. We create density maps of size 32^3 with 1 nm resolution. We then randomly rotate and translate each density map and create 1000 such copies. We then convolve the density maps with CTF with CTF parameters common in experimental datasets (Defocus -5 μm, Spherical Abberation 1.7, Voltage 300 kV). Afterward, we add noise to the convolved density maps so that the SNR is close to 0.1. Thus, we prepare $18,000$ realistic subtomograms with 3 different protein identities, each with 6 different conformations. We uploaded the entire dataset anonymously at https://zenodo.org/records/11244440 under CC-BY-SA license. Sample subtomograms for nucleosomes, spike proteins, and FAS units are provided in Figure 6, Figure 8, and Figure 9 respectively. The figures show 3D slice-by-slice visualization for each conformation of the corresponding protein.

We could not train VITAE [35] on subtomogram datasets since it did not define any transformation for 3D data. Designing CPAB transformation for 3D data by ourselves was challenging. However, we trained SpatialVAE and Harmony as baselines against our subtomogram dataset. Between these two, spatialVAE could not distinguish the protein identities with high heterogeneity at all, which is evident by its embedding space UMAP (Figure **??**). Only Harmony and DualContrast showed plausible result, where DualContrast showing much superior disentanglement than

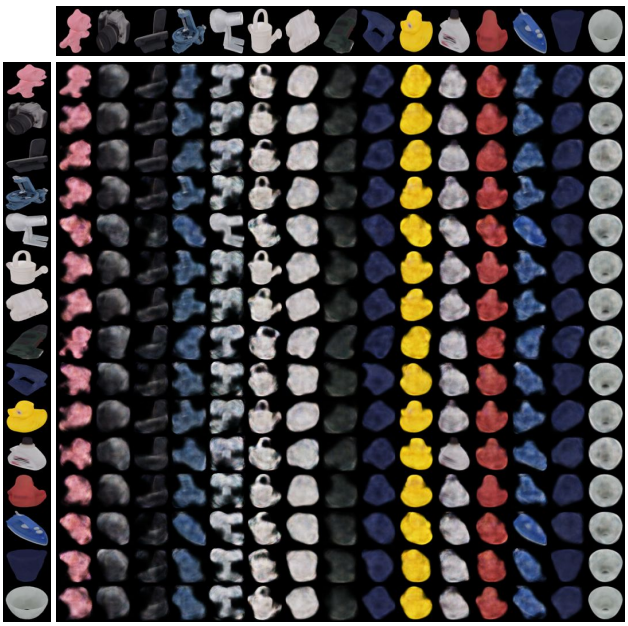


(a)

(b)



(c)



(d)

Figure 5. Qualitative results of image generation with content-transformation transfer obtained by (a) Harmony [38], (b) SpatialVAE [3], (c) VITAE [35], and (d) DualContrast respectively. Harmony and SpatialVAE perform well in reconstruction. but can only perform rotation and translation with its explicitly defined transformations. On the other hand, VITAE can comparatively perform better disentanglement with very poor reconstruction results, distorting the images. On the other hand, DualContrast provides superior content-transformation transfer with optimal performance in reconstruction.

Harmony (Figure **??**).

Hydra Training:

For training Hydra [26], we require 2D cryo-EM images as inputs. To make equivalent inputs from our 3D subtomograms, we took projection sum across depth (z-axis). This resulted in 18,000 2D grayscale images of size 32×32. We

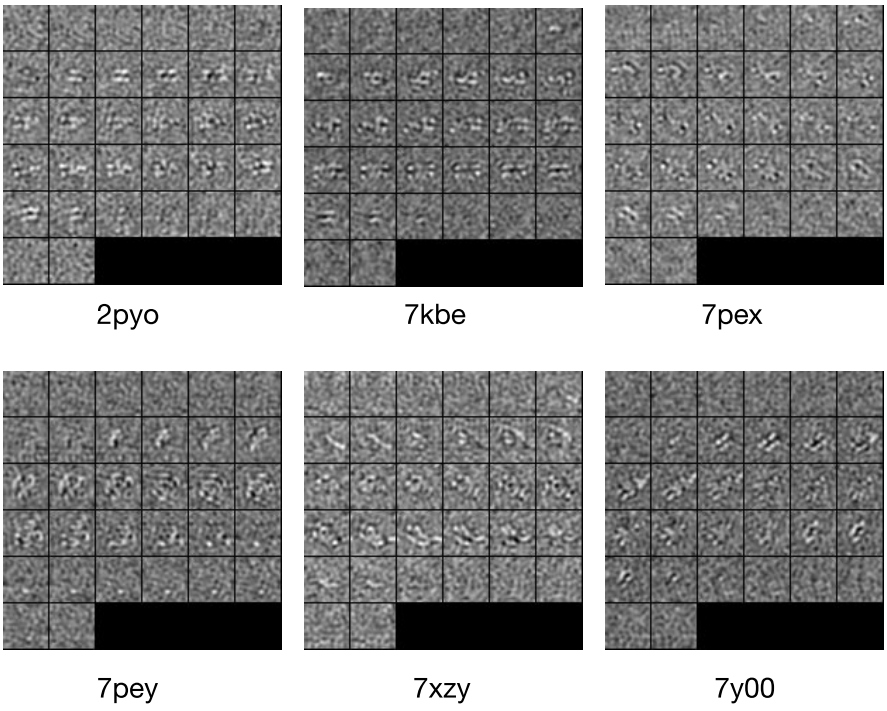


Figure 6. 3D slice-by-slice visualization of Nucleosome subtomograms. Each subtomogram is associated with the PDB ID of the original structure.

set the CTF parameters according to the ground truth during the simulation. For instance, we used Defocus -5 μm, Defocus Angle 90 degrees, Voltage 300kV, Spherical Abberation 1.7, Amplitude 0.1, and Phase Shift 0 degrees.

Since there are 4 ground truth compositions, we trained the Hydra model with K=4. We trained it for 130 epochs, where the first 30 epochs are used for exhaustive pose search. It took around 15 hours to train on a single NVIDIA A5000 GPU.

A.4.4. Ablation Study

To evaluate the individual contribution of the contrastive losses, we conduct both quantitative and qualitative ablation analyses of DualContrast. We trained (1) DualContrast without any contrastive loss, which is basically a VAE with two latent spaces, (2) DualContrast with only $L = L_{\text{VAE}} + L_{\text{con (z)}}$, and (3) $L = L_{\text{VAE}} + L_{\text{con (c)}}$. We qualitatively and quantitatively evaluated each model.

For (1), the D_{score} is almost the same for both **c** and **z** codes, indicating equal predictivity of the digit classes by both codes. This is obvious given that the model has no inductive bias to make different codes capture different information. In model (2), using contrastive loss w.r.t. only **z** factor makes it uninformative of the data. Thus, it provides a small $D(c|z)$ score as desired, but the changing **z** does not affect the image generation (Fig. 10). On the other hand, in the model (3), using contrastive loss w.r.t. only **c** gives a less informative **c** factor, a lower $D(c|c)$ score, and a higher $D(c|z)$ score, contrary to what is desired. These results indicate that contrastive loss w.r.t to both codes is crucial for the desired disentanglement.

Furthermore, we investigated whether using only positive pairs or negative pairs for both codes is sufficient for disentanglement. Nevertheless, we found that both leads to suboptimal disentanglement. If only negative pairs are used, only rotation is disentangled. If only positive pairs are used, then the transformation code becomes uninformative of the data, similar to the degenerate solution.

We provide further quantitative results on the ablation study in Figure 10. The image grids show decoder-generated images where the content factor is used from the corresponding topmost row, and the transformation factor is used from the corresponding leftmost column image.

References

[1] Tristan Aumentado-Armstrong, Stavros Tsogkas, Allan Jepson, and Sven Dickinson. Geometric disentanglement for generative latent shape models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8181–8190, 2019. 2

[2] Ahmet Bakan, Lidio M Meireles, and Ivet Bahar. Prody: pro-

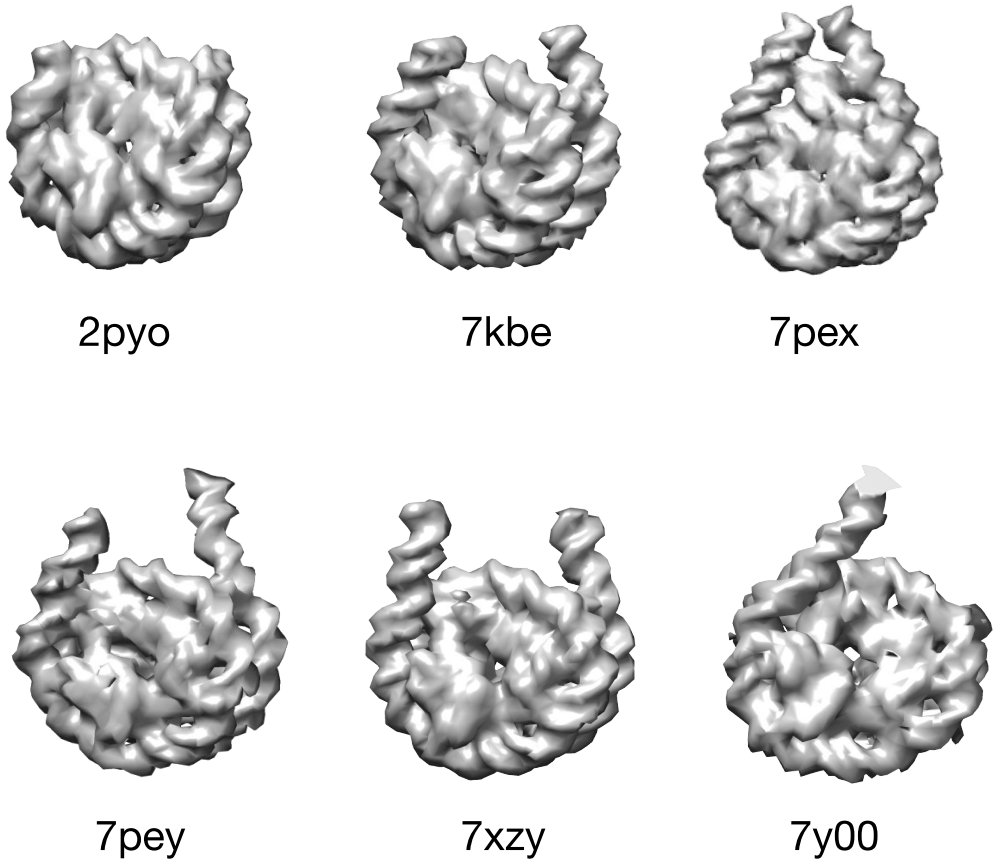


Figure 7. Isosurface visualization of Nucleosome Density Maps. Each density map slightly varies in terms of conformation. They are associated with the PDB IDs in the figure.

tein dynamics inferred from theory and experiments. *Bioinformatics*, 27(11):1575–1577, 2011. 1

[3] Tristan Bepler, Ellen Zhong, Kotaro Kelley, Edward Brignole, and Bonnie Berger. Explicitly disentangling image content from translation and rotation with spatial-vae. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2, 3, 5, 6, 7, 8

[4] Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018. 1

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[6] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016. 1

[7] Luca Cosmo, Antonio Norelli, Oshri Halimi, Ron Kimmel, and Emanuele Rodola. Limp: Learning latent shape representations with metric preservation priors. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 19–35. Springer, 2020. 1, 2

[8] Irene de Teresa-Trueba, Sara K Goetz, Alexander Mattausch, Frosina Stojanovska, Christian E Zimmerli, Mauricio Toro-Nahuelpan, Dorothy WC Cheng, Fergus Tollervey, Constantin Pape, Martin Beck, et al. Convolutional networks for supervised mining of molecular patterns within cellular context. *Nature Methods*, 20(2):284–294, 2023. 6

[9] Allison Doerr. Cryo-electron tomography. *Nature Methods*, 14(1):34–34, 2017. 6

[10] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms

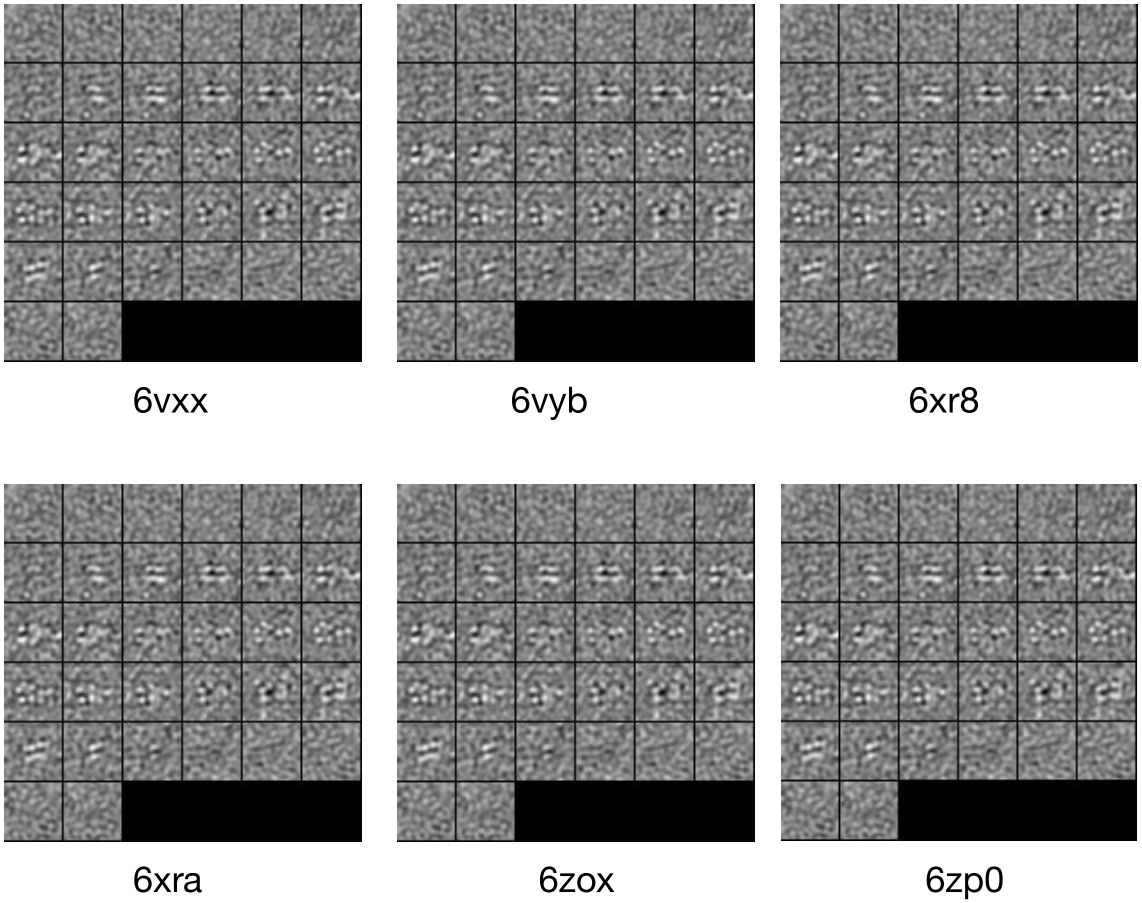


Figure 8. 3D slice-by-slice visualization of Spike Protein Subtomograms. Each subtomogram is associated with the PDB ID of the original structure.

for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011. 1

[11] Mohamad Harastani, Mikhail Eltsov, Amélie Leforestier, and Slavica Jonic. Tomoflow: Analysis of continuous conformational variability of macromolecules in cryogenic subtomograms based on 3d dense optical flow. *Journal of molecular biology*, 434(2):167381, 2022. 6

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3

[13] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2016. 1

[14] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Computer Vision–ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part I 11*, pages 548–562. Springer, 2013. 5

[15] Qixing Huang, Xiangru Huang, Bo Sun, Zaiwei Zhang, Junfeng Jiang, and Chandrajit Bajaj. Arapreg: An as-rigid-as possible regularization loss for learning deformable shape generators. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5815–5825, 2021. 1

[16] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13 (4-5):411–430, 2000. 1

[17] Minkyu Jeon, Rishwanth Raghu, Miro Astore, Geoffrey Woollard, J Feathers, Alkin Kaz, Sonya Hanson, Pilar Cossio, and Ellen Zhong. Cryobench: Diverse and challenging datasets for the heterogeneity problem in cryo-em. *Advances in Neural Information Processing Systems*, 37:89468–89512, 2024. 4

[18] Jonathan Kahana and Yedid Hoshen. A contrastive objective for learning disentangled representations. In *European Conference on Computer Vision*, pages 579–595. Springer, 2022. 3

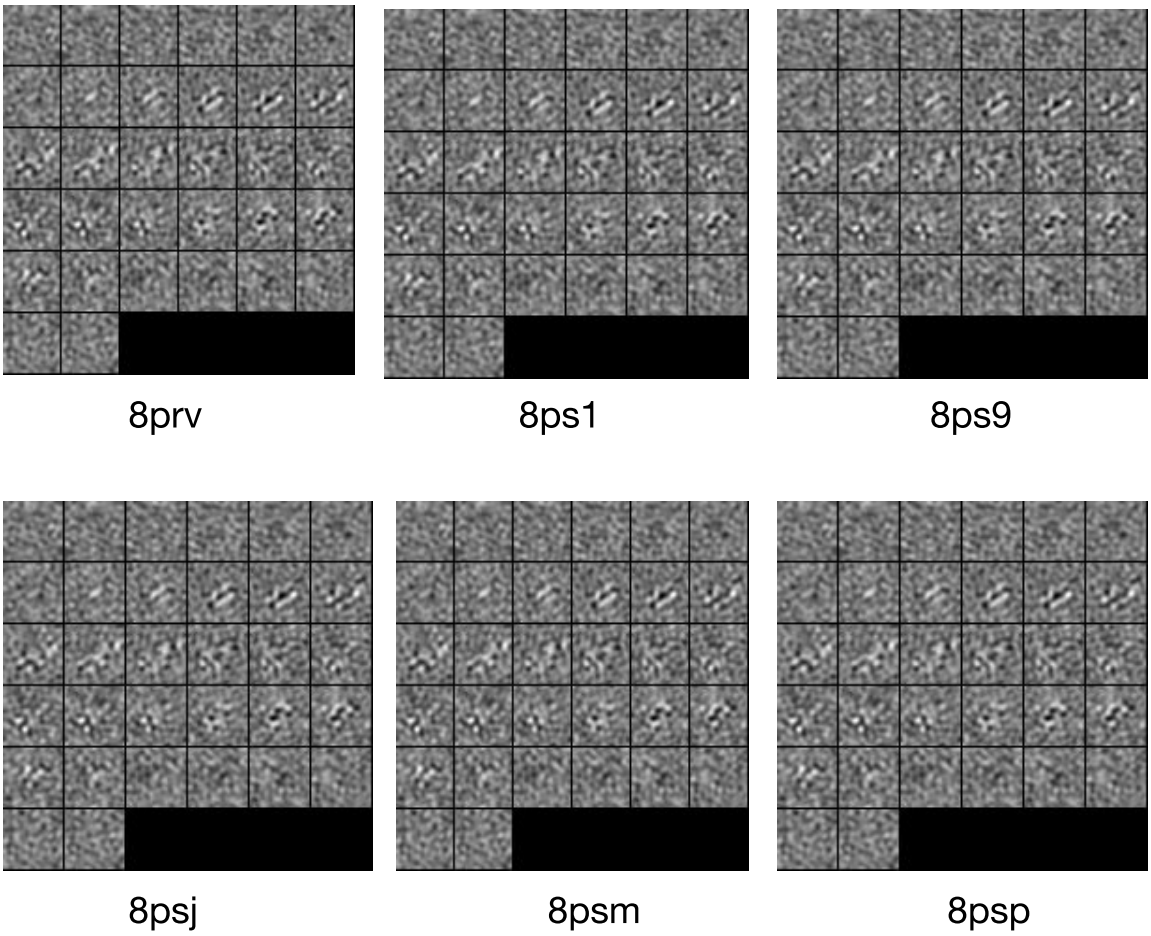


Figure 9. 3D slice-by-slice visualization of FAS subtomograms. Each subtomogram is associated with the PDB ID of the original structure.

[19] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018. 1

[20] Minyoung Kim, Yuting Wang, Pritish Sahu, and Vladimir Pavlovic. Bayes-factor-vae: Hierarchical bayesian deep auto-encoder models for factor disentanglement. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2979–2987, 2019. 1

[21] Steffen Klein, Mirko Cortese, Sophie L Winter, Moritz Wachsmuth-Melm, Christopher J Neufeldt, Berati Cerikan, Megan L Stanifer, Steeve Boulant, Ralf Bartenschlager, and Petr Chlanda. Sars-cov-2 structure and replication characterized by in situ cryo-electron tomography. *Nature communications*, 11(1):5885, 2020. 6

[22] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations*, 2018. 1

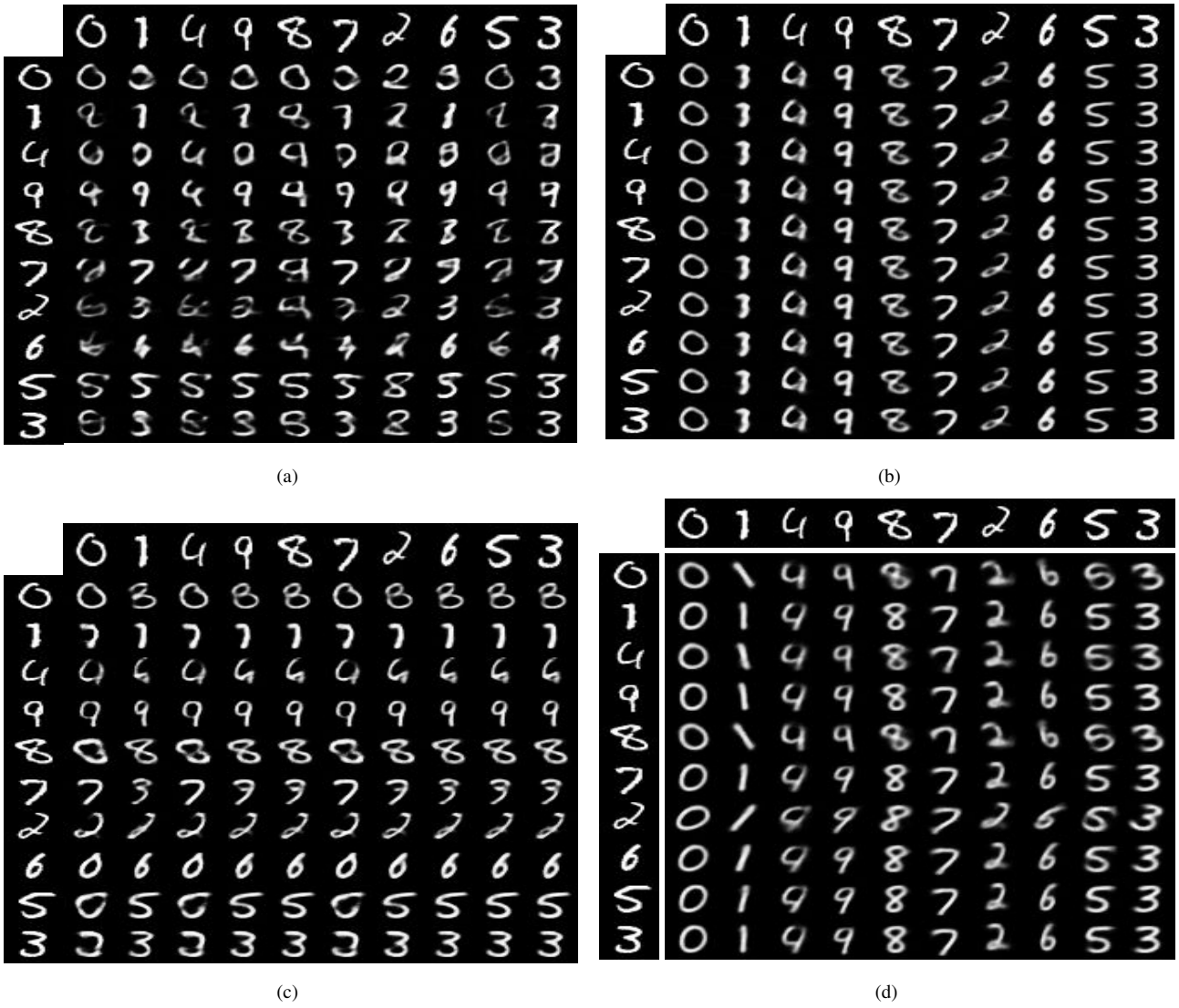
[23] Gihyun Kwon and Jong Chul Ye. Diagonal attention and style-based gan for content-style disentanglement in image generation and translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13980–13989, 2021. 1

[24] Axel Levy, Frédéric Poitevin, Julien Martel, Youssef Nashed, Ariana Peck, Nina Miolane, Daniel Ratner, Mike Dunne, and Gordon Wetzstein. Cryoai: Amortized inference of poses for ab initio reconstruction of 3d molecular volumes from real cryo-em images. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022. 1

[25] Axel Levy, Gordon Wetzstein, Julien NP Martel, Frederic Poitevin, and Ellen Zhong. Amortized inference for heterogeneous reconstruction in cryo-em. *Advances in neural information processing systems*, 35:13038–13049, 2022. 1

[26] Axel Levy, Rishwanth Raghu, David Shustin, Adele Peng, Huan Li, Oliver Clarke, Gordon Wetzstein, and Ellen Zhong. Mixture of neural fields for heterogeneous reconstruction in cryo-em. *Advances in Neural Information Processing Systems*, 37:56988–57017, 2025. 8

[27] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. Unsupervised part segmentation through disentangling appearance and shape. In *Proceedings of the IEEE/CVF Con-*



(a) (b) (c) (d)

Figure 10. Content-transformation transfer results from ablation analysis. (a), (b), (c), and (d) show the results when the model is trained with L_{VAE}, $L_{\text{VAE}} + L_{\text{con (z)}}$, $L_{\text{VAE}} + L_{\text{con (c)}}$ and full objective respectively.

ference on Computer Vision and Pattern Recognition, pages 8355–8364, 2021. 1

[28] Elaine C Meng, Thomas D Goddard, Eric F Pettersen, Greg S Couch, Zach J Pearson, John H Morris, and Thomas E Ferrin. Ucsf chimerax: Tools for structure building and analysis. *Protein Science*, 32(11):e4792, 2023. 5

[29] Federico Monti, Davide Boscaini, Jonathan Masci, Emanuele Rodola, Jan Svoboda, and Michael M Bronstein. Geometric deep learning on graphs and manifolds using mixture model cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5115–5124, 2017. 1

[30] Lilian Ngweta, Subha Maity, Alex Gittens, Yuekai Sun, and Mikhail Yurochkin. Simple disentanglement of style and content in visual representations. In *International Conference on Machine Learning*, pages 26063–26086. PMLR, 2023. 1

[31] Pablo Palafox, Aljaž Božič, Justus Thies, Matthias Nießner, and Angela Dai. Npms: Neural parametric models for 3d deformable shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12695–12705, 2021. 1

[32] PDB RCSB. Rcsb pdb, 2000. 6

[33] Xuanchi Ren, Tao Yang, Yuwang Wang, and Wenjun Zeng. Learning disentangled representation by exploiting pretrained generative models: A contrastive learning view. In *International Conference on Learning Representations*, 2021. 1

[34] Xuanchi Ren, Tao Yang, Yuwang Wang, and Wenjun Zeng. Rethinking content and style: exploring bias for unsuper-

vised disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1823–1832, 2021. 1

[35] Nicki Skafte and Søren Hauberg. Explicit disentanglement of appearance and perspective in generative models. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2, 3, 5, 6, 7, 8

[36] Qingyang Tan, Lin Gao, Yu-Kun Lai, and Shihong Xia. Variational autoencoders for deforming 3d mesh models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5841–5850, 2018. 1, 2

[37] Guang Tang, Liwei Peng, Philip R Baldwin, Deepinder S Mann, Wen Jiang, Ian Rees, and Steven J Ludtke. Eman2: an extensible image processing suite for electron microscopy. *Journal of structural biology*, 157(1):38–46, 2007. 7

[38] Mostofa Rafid Uddin, Gregory Howe, Xiangrui Zeng, and Min Xu. Harmony: A generic unsupervised approach for disentangling semantic content from parameterized transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20646–20655, 2022. 1, 2, 3, 5, 6, 7, 8

[39] Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467, 2021. 1, 3

[40] Zhizhong Wang, Lei Zhao, and Wei Xing. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7677–7689, 2023. 1

[41] Paul Wohlhart and Vincent Lepetit. Learning descriptors for object recognition and 3d pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3109–3118, 2015. 5

[42] Ellen D Zhong, Adam Lerer, Joseph H Davis, and Bonnie Berger. Cryodrgn2: Ab initio neural reconstruction of 3d protein structures from real cryo-em images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4066–4075, 2021. 1

[43] Keyang Zhou, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Unsupervised shape and pose disentanglement for 3d meshes. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 341–357. Springer, 2020. 1, 2