Harmony: A Generic Unsupervised Approach for Disentangling Semantic Content from Parameterized Transformations

Mostofa Rafid Uddin[1] Gregory Howe[2] Xiangrui Zeng[1] Min Xu[1,*]
[1]Computational Biology Department, [2]Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213, USA.
mru@andrew.cmu.edu gregory.s.howe@gmail.com xiangruz@andrew.cmu.edu mxu1@cs.cmu.edu

Abstract

In many real-life image analysis applications, particularly in biomedical research domains, the objects of interest undergo multiple transformations that alters their visual properties while keeping the semantic content unchanged. Disentangling images into semantic content factors and transformations can provide significant benefits into many domain-specific image analysis tasks. To this end, we propose a generic unsupervised framework, Harmony, that simultaneously and explicitly disentangles semantic content from multiple parameterized transformations. Harmony leverages a simple cross-contrastive learning framework with multiple explicitly parameterized latent representations to disentangle content from transformations. To demonstrate the efficacy of Harmony, we apply it to disentangle image semantic content from several parameterized transformations (rotation, translation, scaling, and contrast). Harmony achieves significantly improved disentanglement over the baseline models on several image datasets of diverse domains. With such disentanglement, Harmony is demonstrated to incentivize bioimage analysis research by modeling structural heterogeneity of macromolecules from cryo-ET images and learning transformation-invariant representations of protein particles from single-particle cryo-EM images. Harmony also performs very well in disentangling content from 3D transformations and can perform coarse and fast alignment of 3D cryo-ET subtomograms. Therefore, Harmony is generalizable to many other imaging domains and can potentially be extended to domains beyond imaging as well.

1. Introduction

In many real-life image analysis applications, particularly in biomedical research domains (e.g., electron microscopy, tomography, nanobody images, etc.), the appearance of the objects of interests are affected by a sequence of transformations. For instance, in single-particle cryo-electron microscopy (cryo-EM) images, the shapes, translations, rotations, and projections of protein particles are confounded [1]. In cryo-electron tomography (cryo-ET), macromolecules are present in different orientations and shifts in different subtomograms (3D cryo-ET subimages each containing a macromolecule) [27,29]. A Magnetic resonance image (MRI) of brain can be differently scaled and illuminated due to different imaging settings [23]. Consequently, in these image analysis domains, images can be encoded into and generated from a semantic content factor that is specific to the shape of the object of interest and a sequence of parameterized transformations that are unspecific to the shape of the object. This is referred to as disentangling content and transformations in the highly non-linear latent space of the images. Such content-transformation disentangling can provide insights into the shape space and transformation distributions inherent in the images [1] and further facilitate several downstream analysis tasks like image classification, image alignment, image translation, and image extrapolation for the corresponding imaging domains.

Disentangling semantic content from parameterized transformations has not been widely studied until very recently. Traditional disentangled representation learning (DRL) methods [3,13,18] decompose data into various generative latent factors (e.g., face, color, hair, etc.), and do not perform particularly well at content-transformation disentangling [1, 5]. To this end, some explicit DRL methods [1, 5, 8, 17] have been proposed with satisfactory performance. These explicit DRL methods constrain some latent factors explicitly to represent generative factors that are known to be inherent in the data beforehand. In the aforementioned image analysis applications, the types of transformations are usually known beforehand and can be esti-

*Corresponding author

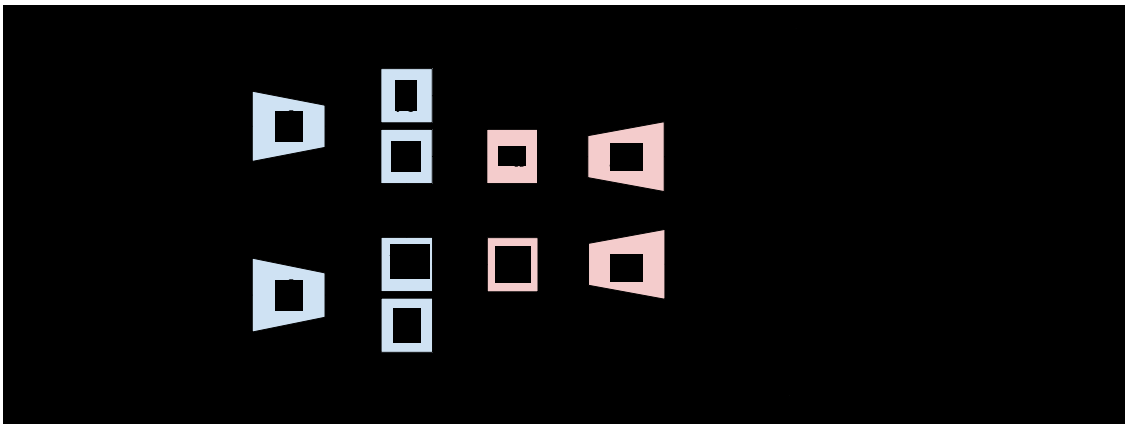

Figure 1. The full workflow of our proposed method 'Harmony'. The encoder f_θ infers the transformation factors k and semantic latent distribution Φ_x for a input datum x. k is then used to transform x with a differentiable transformer and matched with the decoded output generated from semantic latent factor z_x drawn from a distribution P_{Φ_x} proposed by decoder g_ϕ. A similar mechanism is adopted for a randomly transformed datum x. Then corresponding transformed instances by learned transformation factors k and k' are matched with each other. In addition, the corresponding distributions Φ_x and $\Phi_{x'}$ are contrasted with a KL loss. The transformation factors k, k', and the distributions Φ_x and $\Phi_{x'}$ are all updated with gradient-descent while training.

mated using few parameters. In such cases, it is particularly advantageous to use explicit methods that take into account the prior knowledge when designing content-transformation disentangling methods.

Recently, some supervised and semi-supervised explicit methods [8, 17] have been developed that disentangle content and parameterized transformation latent factors. But these methods require labeled data, which makes them undesirable in the aforementioned image analysis domains where labeled data is hard to obtain. The only notable unsupervised method that explicitly disentangles contents from parameterized transformations is SpatialVAE [1], which uses a specialized decoder architecture to disentangle content from rotation and translation in 2D images. SpatialVAE and its variants [16] have been applied to different imaging domains, such as, astronomical images, electron microscopy, nanorod images, and etc and has been remarkably successful. However, these methods [1, 16] do not provide a generic framework as they can not disentangle semantic content from other types of transformations (e.g., scaling, contrast, etc.). Moreover, efficacy of these methods for 3D images has been unexplored. Developing an unsupervised generic framework for disentangling semantics and transformations remains an open problem.

In this work, we develop a generic unsupervised framework, *Harmony* (Figure 1), to explicitly disentangle semantic content from multiple parameterized transformations. Harmony takes a set of images as inputs without any label information and learns disentangled latent representations where one latent factor corresponds only to shape-specific semantic contents of objects and the others correspond to the different parameterized transformations of the objects. To this end, like other explicit DRL methods [1, 5, 16], Harmony only uses types of transformations that are known to be present beforehand. To perform DRL, Harmony explicitly constrains one latent factor to correspond to the semantic content and others to represent transformation parameters of known transformation types. However, only using this constraint often results in trivial parameterization of transformations and consequently poor disentanglement. To avoid such trivial parameterization, Harmony incorporates cross-contrastive learning with data augmentations. It creates an augmented version of the input image with the known transformation types and enforces the decoded images to be similar to both the input and augmented images. For both images, it models the semantic latent factors as multivariate Gaussian distributions and enforces the two distributions to be similar to each other. Harmony is the first method to leverage cross-contrastive learning for unsupervised content-transformation disentangling and achieve remarkable success.

We experimented with Harmony to disentangle content from multiple geometric transformations in two real single-particle cryo-EM datasets and several simulated and real 3D cryo-ET subtomogram datasets. To assess the generalization ability of Harmony's method, we used it to disentangle contents from transformations in a randomly rotated, translated, and scaled version of MNIST digit dataset and disentangle content from lighting condition transformation in a variant of celebA facial image dataset. In all of our experiments, Harmony demonstrated significantly improved results over baseline methods in both qualitative and quantitative evaluations. In an ideal disentanglement setting, changing the semantic content factor would only alter the shapes

of the objects that are specific to image class, whereas having no effect on the transformations. Our experiments show that, Harmony performs very close to the ideal setting. As Harmony does not make any assumptions on input domain, it can be used for many other image analysis applications (astronomical images, nanobody image, etc.) and may be leveraged for domains beyond imaging (e.g.,voice, speech, etc.) as well.

Our primary contributions are as follows:

(i) We introduce a generic framework, Harmony, to disentangle semantic content from multiple parameterized transformations without requiring any image associated labels. We, for the first time, used cross-contrastive learning to accurately disentangle semantic content from transformations.

(ii) As an application of Harmony, we disentangle semantic content from multiple geometric and lighting condition transformations in various imaging datasets with significant improvement over baseline methods.

(iii) By disentangling content from transformations with Harmony, we resolved transformation-invariant representations of proteins from 2D single-particle cryo-EM images. We learned more accurate representation than previous methods with improved efficiency.

(iv) We, for the first time, disentangled transformation parameters from 3D images and applied it to model structural heterogeneity of extremely noisy real and simulated 3D cryo-ET subtomograms. Harmony can also perform coarse and fast unsupervised groupwise image alignment of cryo-ET subtomograms.

2. Related Works

Disentangled Representation Learning: Learning disentangled representation factors or independent factors of data is a well studied problem in data science [4, 14, 15, 21, 24, 26]. Recently, several variational-autoencoder (VAE) based deep generative methods, e.g., β-VAE [13], Factor-VAE [18], TC-β-VAE [3], DIP-VAE [19], etc., have shown promising results in disentangling factors in the highly non-linear latent space of data thanks to the ability of deep models to tackle non-linearity. These methods do not use any prior knowledge on the generative factors, but rather interpret them using latent traversals after learning the model. As contents and transformations are both generative factors, in principal, these methods can be used to implicitly disentangle semantic contents and transformations. However, it has been shown that implicitly disentangling content and transformation in such way gives very poor disentanglement in practice [1, 5].

Content-style disentanglement: A specific version of the DRL problem related to our work is content-style disentanglement (CSD), where images are decomposed into content specific and style specific factors. Though there exists a large number of methods for CSD, the most relevant method to our work is Deforming Autoencoder [25] and U-VITAE [5] which explicitly disentangles appearance (content) and perspective (style) using two different latent spaces in an unsupervised way. They experimentally disentangled content from 2D translations and rotations by considering them as implicit parts of a style factor. Despite achieving some success, the disentanglement of transformations with respect to contents remain implicit and is hard to interpret. In contrast, Harmony explicitly disentangles content and transformations in the same latent space which is easily interpretable.

Disentangling content and transformations: For unsupervised explicit disentangling of content from transformations, a very recent yet pioneering work is SpatialVAE [1], that explicitly disentangles content from two parameterized transformations (rotation and translation) in 2D images. SpatialVAE exploits the fact that rotation and translation are present as generative factors in many real-life 2D image datasets, but does not require any labeled data. It is a VAE-based architecture with rotation and translation specific prior constraints on the rotation and translation latent factors respectively. SpatialVAE parameterizes the pixel intensity distribution at a spatial coordinate explicitly as a function of the Euclidean coordinates and thus makes the image reconstruction term differentiable with respect to the latent rotation and translation parameters. Afterwards, Kalinin et. al developed a rotation-invariant VAE architecture, rVAE, to disentangle rotational dynamics from nanorod images [16]. SpatialVAE and its variants has achieved remarkable success in disentangling 2D rotation and translation from single particle cryo-EM, astronomical, and nano-particle images. Nevertheless, these frameworks are not directly applicable by design to other kinds of transformations e.g., scaling, contrast, etc. Moreover, these methods depend on euclidean geometries that often do not work for 3D image objects.

Contrastive Learning: One of the building blocks of Harmony is the use of contrastive learning. Contrastive learning is a technique for learning feature representations using similarity and dissimilarities present in an unlabeled dataset. Contrastive learning techniques involve maximizing a similarity score between semantically similar images, 'positive pairs', while simultaneously minimizing a similarity score between all semantically dissimilar images, 'negative pairs'. Recent works [2, 6, 28] have achieved state of the art performance on ImageNet without the need for negative examples. Similarly, Harmony performs contrastive learning without negative examples to avoid trivial parameterization of transformations while disentangling.

3. Method

3.1. Disentangled Representation Learning

DRL aims to learn a map $h : X \to Z$, where X is an input space and Z is a disentangled intermediary representation space. In content-transformation disentangling, transformations are treated as a group action $\mathcal{T}$ acting on X. The transformation parameter space, K, is treated as a group with a corresponding group action $\mathcal{T}' : Z \times K \to Z$ to carry out the transformation [12]. Higgens *et al.* [12] define the intermediary representation, Z, to be disentangled with respect to the subgroup decomposition, $K = K_1 \times \cdots \times K_n$, if the following conditions hold:

- There is a decomposition $Z = Z_1 \times ... \times Z_n$ s.t. $k_i = k'_i \Rightarrow \mathcal{T}'(z, k)_i = \mathcal{T}'(z, k')_i$
- h is equivariant between actions $\mathcal{T}$ and $\mathcal{T}'$ performed on X and Z

Under this decomposition, for every $i \neq j$, the output $z'_i = \mathcal{T}'_i(z_i, k_i)$ is unaffected by changes to parameter k_j.

3.2. Method Overview and Notation

At a high level, our method relies on an autoencoder-like architecture (Figure 1). Specifically, consider a sample space X, a transformation parameter space K, a latent distribution parameter space Ψ, and a latent space Z. Harmony first applies an encoder network $f_\theta : X \to K \times \Psi$ on a sample $x \in X$ to encode transformation parameters $k_x \in K$ and $\psi_x \in \Psi$ where $(k_x, \psi_x) = f_\theta(x)$. These encoded parameters are used to construct two outputs $x_k, x_z \in X$. The first output is constructed by applying a transformation function $\mathcal{T} : X \times K \to X$ to produce $x_k = \mathcal{T}(x, k_x)$. The second output is constructed using a two step process. Latent parameters z_x are drawn from the encoded latent probability distribution P_{ψ_x}. These parameters are then decoded using a decoder network $g_\phi : Z \to X$ to produce $x_z = g_\phi(z_x)$.

Harmony is trained to produce instances x_k and x_z that are similar to each other. This is done through penalizing a reconstruction loss or dissimilarity score between x_k and x_z, but this loss alone does not encourage consistency across transformations and results in trivial transformation parameters. For example, assuming an identity transformation parameterized by $k^{(\text{id})} \in K$, the decoder f_θ could feasibly learn to always output $k^{(\text{id})}$. In effect, this would change the method into a standard autoencoder-like architecture, which would not disentangle semantic representation from transformations. Some VAE-based DRL methods [1, 5] have addressed this issue by applying transformation-specific priors on the latent space that worked as an inductive bias. On the contrary, we have used a concept of similarity to avoid the trivial paramaterization of transformations.

3.3. Avoiding Trivial Parameterization of Transformations

We call two samples x and x' *semantically similar* if there exist $k \in K$ and $k' \in K$ such that $x' \approx \mathcal{T}(x, k)$ and $x \approx \mathcal{T}(x', k')$. To achieve a setting where contents are disentangled from transformations, Harmony incentives x_z to be a transformation-invariant sample, $x^{(\text{p})}$ that is consistent across different transformations from K. To accomplish this, Harmony creates a semantically similar sample x' for each x, by transforming x with a random $k \in K$. Then it uses a siamese like branch to create $x'_{z'}$ and $x'_{k'}$ from x' with the same encoder and decoder. Finally, it uses a new dissimilarity score between x_k and $x'_{k'}$ to make them close to a transformation invariant sample $x^{(\text{p})}$. With the other two dissimilarity scores (between (x_z, x_k) and ($x'_{z'}$, $x'_{k'}$)), the model eventually learns to achieve the disentanglement goal where $x_z = x^{(\text{p})}$. This mechanism helps avoid trivial solutions and works to create an inductive bias in the transformation space.

We show in supplementary section S1 that for a fixed sample x when $\mathcal{D}^{(2)}$ is chosen to be the sum of squared errors in expectation this is equivalent to minimizing $2 \cdot \mathbf{V}_x [x_k]$ where x_k is a transformed instance of x. Intuitively, to minimize this variance, for every transformed instance of x, the encoder must learn to propose transformations k such that for every x, x_k is 'close' to some prototypical sample, $x^{(\text{p})}$. Furthermore, all transformations $k \in K$ are transformations, thus we have every output x_k is both 'close' to $x^{(\text{p})}$ and retains the same semantic meaning as the original x. So, in effect, the encoder, f_θ, must learn to transform any already transformed instance of x, x_k, to be 'close' to $x^{(\text{p})}$, thus disentangling x from the transformations in K.

3.4. Encouraging proximity of semantic latent distributions of homogeneous classes

When x and x' are two *similar* input samples, we want to enforce proximity in the semantic latent space. By assuming there is an $x^{(\text{p})}$ for x and x' and imposing the assumption with the $\mathcal{D}^{(2)}$ loss between the transformed samples obtained from each of x and x', we indirectly encourage the proximity of the semantic latent factors z_k and $z_{k'}$ of x and x'. To further increase proximity in the latent space, we introduce a per dimension KL divergence $D_{KL}\left(P_\Psi || P_{\Psi'}\right)$ between the encoded semantic latent distributions of x and its *similar* instance x' to our objective function. In our experiments, P_Ψ and $P_{\Psi'}$ are both multivariate gaussians, which allows for efficient computation of the per-dimension KL divergence between P_Ψ and $P_{\Psi'}$.

3.5. Objective function

Given a data sample x, we compute the distance between the instance reconstructed from the semantic content,

x_z, and the instance transformed by the proposed transformation parameters, x_k. This loss enforces that the reconstructed instance x_z is similar to the transformed instance x_k, which serves two main purposes. First, it incentivizes x_z to be similar to a transformed instance x_k that is in the same semantic group as x. So, in effect, this forces the network's reconstruction to be both semantically meaningful and semantically similar to the original x. Second, making x_z similar to x_k indirectly makes x_z similar to anything that x_k is similar too. Further we include a KL divergence between P_Ψ and $P_{\Psi'}$, the semantic latent distributions of x and x' respectively. This loss encourages proximity of the semantic distributions of homogeneous objects in the latent space. Combining the losses, yields Harmony's objective loss as follows:

$$\begin{aligned}\mathcal{L}(x, x') =& \gamma[\mathcal{D}^{(1)}(x_z, x_k) + \mathcal{D}^{(2)}(x_k, x'_{k'}) + \mathcal{D}^{(3)}(x'_{k'}, x'_{z'})] \\ &+ D_{KL}\left(P_\Psi || P_{\Psi'}\right)\end{aligned}$$

Here, the hyperparameter γ helps avoid training instability that may be caused by the per dimension KL divergence loss. While training, γ is usually set proportional to $\frac{M}{N}$, where N is the batch size and M is the number of training data points. The experimental analysis on the effect of γ on disentanglement is provided in supplementary section S3.3. Our entire framework is depicted in figure 1.

We note that our loss only indirectly penalizes representations of x_z and $x'_{z'}$ that are far apart. We don't explicitly include a loss term, $\mathcal{D}^{(4)}(x_z, x'_{z'})$ or $\mathcal{D}^{(4)}(z_x, z'_{x'})$, because we experimentally observed that this loss introduces difficulty in escaping local minima.

3.6. Regularisation of Transformations

Harmony is designed to disentangle any parameterized transformation that does not alter the semantic meaning of the datum. Therefore, while training, we restricted the domain of transformations that can alter the semantic meaning of datum (e.g., scale, shift) by using appropriate activation functions. Moreover, our implementation of $\mathcal{T}$ includes a grid generator with a bilinear kernel. The kernel weights are shared across two branches of Harmony, which causes implicit regularization of transformation parameters.

4. Experiments & Results

The experimental setup, implementation and training details are provided in supplementary section S3.1 and S3.2.

4.1. Evaluation Metrics

The evaluation of DRL methods is still mostly qualitative and no quantitiative metric is universally despite some are commonly used. Nevertheless, Locatello *et al.* [22] showed that all the evaluation metrics are highly correlated. As we have access to the discrete class labels as ground truth factors for evaluation purpose, a feasible approach is to evaluate how well our semantic latent factor can predict the ground truth class labels. In such scenario, SAP score is the one of the most acceptable metrics by the community [20]. SAP score simply denotes the difference between the top two predictivity scores for ground truth factor by individual latent factors. SAP score for content identity disentanglement can be defined as follows:

$$\text{SAP}_{\text{score}} = \arg\max_{i \in [N]} P(c|z^{(i)}) - \arg\max_{j \neq i \in [N]} P(c|z^{(j)})$$

Here, c is content identity, z is the N dimensional latent factor, and $P(c|z^{(i)})$ is the predictivity of content identity by i^{th} dimension of the latent factor. For measuring predictivity we used a non-linear K-Nearest Neighbor (KNN) algorithm, similar to U-VITAE [5]. We report SAP score along with maximum predictivity of content from semantic latent factor $P(\mathbf{c}|\mathbf{z})$ in our quantitative evaluations.

A. 2D single-particle cryo-EM images

First, we tested Harmony against two negative stain noisy single-particle cryo-EM image datasets to learn transformation-invariant structures of the corresponding protein particles. Among the two cryo-EM image datasets used in our experiments, one contains the StrepMAB-Classic antibody and the other contains the CODH/ACS protein complex. The model was trained in a setting similar to the setting used by SpatialVAE [1]. The details of the dataset and training are provided in the supplementary section S2.1 and S5.2 respectively.

From the latent manifold generated by Harmony (Figure 2), it is evident that Harmony learns a transformation invariant representation of the protein particles and simultaneously generates better resolution pose normalized particle images from the set of noisy cryo-EM images. Figure 2 shows the interpolated protein image generated from the content latent manifold by Spatial-VAE and Harmony along with sample input images for both the CODH/ACS and antibody datasets. Spatial-VAE occasionally captures structure in the image background and sometimes learns inconsistent representations for the antibody dataset. But for Harmony implementation, we did not face such issues. (Figure 2).

Decoupling the orientations and shifts from the images and inferring the transformation-invariant shape of a structure is a crucially important task in single-particle shape analysis in cryo-EM. Harmony serve as a fast and useful tool to this end.

B. 3D cryo-electron subtomogram images

Next, we validated Harmony for shape analysis of macromolecules in 3D cryo-ET images. Unlike cryo-EM, cryo-ET enables 3D visualization of a single cell in its near na-

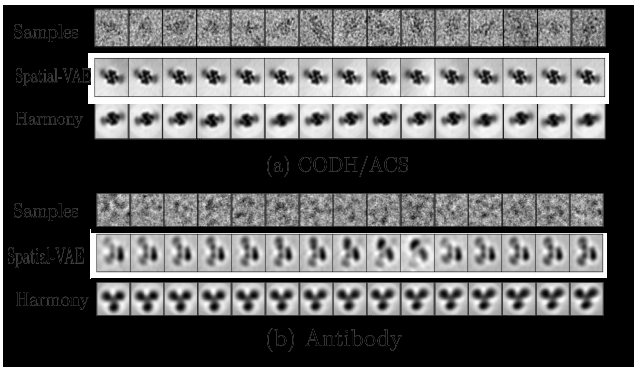

Figure 2. Learning Transformation Invariant Representation of proteins from cryo-EM images using Harmony. (a) shows exemplary cryo-EM images of CODH/ACS protein dataset, interpolated protein conformations from semantic latent manifold learned by Spatial-VAE and Harmony respectively. (b) shows the similar corresponding images from the antibody protein dataset.

tive state. To this end, a 3D cryo-ET image (called a tomogram) contains visualization of all subcellular particles inside a cell simultaneously whereas cryo-EM single-particle 2D images only image a single particle. On the other hand, due to direct imaging in crowded cytoplasmic native environment and spatial anisotropy, cryo-ET images are noisier than cryo-EM.

To perform analysis on the macromolecular structures, cubic sub-volumes each containing a single macromolecule are extracted from the whole tomogram. These sub-volumes are called subtomograms. Some important subtomogram-level analysis tasks are grouping semantically similar subtomograms, aligning, and averaging them to get better resolution structures of the macromolecules. These tasks are challenging because the 3D subtomograms are extremely noisy and the macromolecules are minuscule structures residing in random orientations and shifts inside the noisy subtomograms. Disentangling the orientation and shifts from the shape of macromolecules in 3D subtomograms has the potential to significantly improve the shape analysis of particles and performing the aforementioned downstream analysis tasks in an unsupervised manner.

However, a well performing method for disentangling contents and transformations is notably missing for 3D images, despite there existing a few promising methods [1, 5] for corresponding tasks in 2D images. For a quantitative comparison with Harmony, we extended Spatial-VAE [1] for 3D images. However, such an extension was non-trivial. As spatial-VAE generates images by conditioning them on 2D spatial coordinates, we used 3D spatial coordinates while extending it to 3D images. But doing that alone gives very poor performance on cryo-ET subtomograms. We incorporated a convolution-only encoder architecture in our 3D-Spatial-VAE extension to make it work for subtomograms. However, since 3D space is much larger than 2D, conditioning on 3D co-ordinates makes 3D-Spatial-VAE very slow. Despite using the same encoder architecture in Harmony, the training remains fast due to dependence on fewer parameters. Therefore, Harmony framework is easily extendable to 3D transformation disentangling.

Dataset	SNR	Harmony		3D Spatial-VAE	
		SAP	$P(\mathbf{c}\|\mathbf{z})$	SAP	$P(\mathbf{c}\|\mathbf{z})$
Simulated [30]	100	**0.494**	0.996	0.409	**0.997**
	0.1	**0.384**	**0.861**	0.382	0.858
	0.05	0.169	0.63	**0.286**	**0.65**
	0.03	**0.064**	**0.527**	0.013	0.478
	0.01	0.003	**0.47**	**0.009**	0.46
Rat Neuron [7]	0.01	**0.268**	**0.999**	0.011	0.69

Table 1. Quantitative results on disentangling semantic identity of macro-molecules from parameterized 3D affine transformations in cryo-electron subtomograms.

We tested 3D Harmony and our implementation of 3D-Spatial-VAE against five realistically simulated benchmark datasets with varying signal to noise ratios (SNR) [30] and a real cryo-ET subtomogram dataset [7]. Details on the datasets are provided in the supplementary section S2.2.

We tested Harmony and our 3D-Spatial-VAE against these datasets. For the real dataset and simulated datasets (with SNR > 0.03), Harmony achieved remarkable disentanglement performance (Table 1) with high predictivity of macromolecular identity from semantic latent factor. For simulated datasets, Harmony and our implemented 3D-Spatial-VAE achieve similar results, while both perform very poor in extremely low SNR (≤ 0.03). Though the two models performed similar in simulated datasets, their performance noticeably differed in the real dataset. From the latent manifold learned by our implementation of 3D-Spatial-VAE, the heterogeneity of ribosome and proteasome subtomograms are limited for real dataset (Figure 4). But, in the latent space manifold learned by Harmony (Figure 4), the semantic difference between ribosome and proteasome subtomograms is clearly evident along its only one dimension of semantic latent factor. Thus, Harmony can model the transformation-free structural variability of a set of real subtomograms without requiring any templates or estimations on the number distinct macromolecules beforehand. By disentangling 3D transformations from subtomograms, Harmony can also perform coarse and fast subtomogram alignment (Figure 3).

C. MNIST digit images

To validate the generalizability of Harmony to datasets of other imaging domains and other types of transformations, we tested Harmony against a randomly rotated, translated, and scaled version of the MNIST dataset. For better qualita-

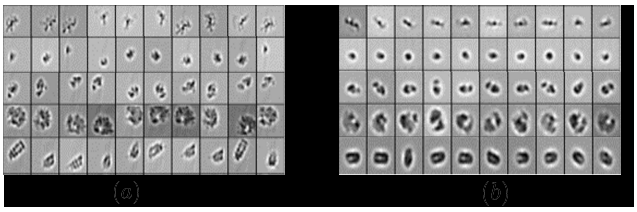

Figure 3. Unsupervised coarse and fast subtomogram alignment with Harmony (in simulated SNR 100 dataset) (a) shows the 2D central slice representation of sample input 3D subtomograms. (b) shows the corresponding representations of the decoded subtomograms. The subtomograms are ordered by class, for better visualization. The order and labels were not used during training.

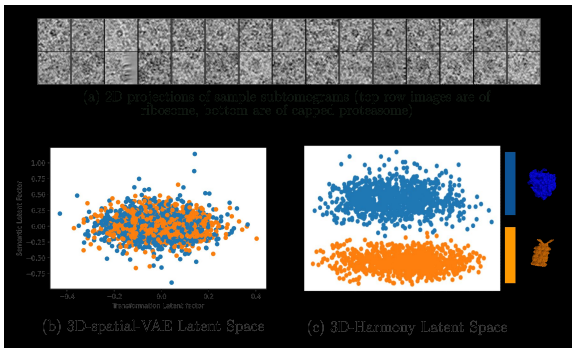

Figure 4. Mapping structural heterogeneity of macromolecules in cryo-ET subtomograms with Harmony (c) and 3D-Spatial-VAE (b). (a) shows 2D central slice projections of some sample subtomograms.

tive visualization, we randomly picked one image per each digit class and created 3000 randomly rotated, translated, and scaled instances of each image. From Figure 5, it is evident that Harmony disentangles full affine matrix (rotation, translation, scaling) from the semantic content and automatically align the semantically similar images (images of same digit) in the dataset. These results demonstrate Harmony's ability to 'harmonize' semantically similar objects in the images out of the multiple parameterized transformations in a generic image analysis datasets.

We reported quantitative scores for disentangling content from affine transformations using Harmony and other related models in Table 2. Except Deforming Autoencoder [25] and U-VITAE [5], all the methods tend to disentangle latent factors in the same latent space. Deforming Autoencoder and U-VITAE uses separate latent spaces to separate content and diffeomorphic transformations. So, for calculating SAP for these two methods, the difference between predictivity for two latent spaces were calculated. For all other methods, including Harmony, the difference between predictivity for different latent factors in the same latent space was calculated. Harmony achieves the highest predictivity and second highest SAP score. The results

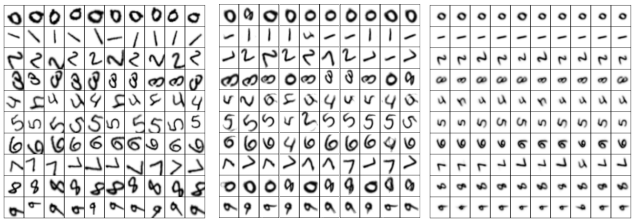

Figure 5. Unsupervised Groupwise Image Alignment using Harmony (a) Exemplary input images of randomly rotated, translated, and scaled versions of mnist digits (b) Corresponding decoded images generated by Deforming Autoencoder [25]. (b) Corresponding decoded images generated by Harmony (with one dimensional semantic latent factor). The ordering of class digits is done for visualization and such information is not used in training.

Method	SAP score	$P(\mathbf{c}\|\mathbf{z})$
PCA [9]	0.065	0.486
FastICA [14]	0.099	0.500
β-VAE [13]	0.017	0.46
β-TC-VAE [3]	0.001	0.356
Spatial-VAE [1]	0.002	0.359
Deforming Autoencoder [25]	0.25	0.92
U-VITAE [5]	**0.731**	0.828
Harmony	*0.55*	**0.944**

Table 2. Quantitative results of disentangling content from affine transformations (rotation, translation, scaling). Except Deforming Autoencoder [25] and U-VITAE [5], all methods use one latent space and SAP scores were calculated for latent factors on that space. Those two methods use two latent spaces and scores was calculated between two latent spaces.

demonstrate the efficacy of Harmony in disentangling content from parameterized affine transformations. In addition, we used Harmony to disentangle rotation and translation in the whole MNIST dataset and provided the results in supplementary section S5.1. We observed that interpolated digits obtained from semantic latent space of Harmony is disentangled from translations and rotation.

D. CelebA RGB Facial Images

We further demonstrate that Harmony can also disentangle lighting condition transformations from semantic contents. To this end, we created a specialized version of the CelebA dataset with 10,000 images of ten distinct facial identity affected by different contrast factors. The details of the dataset creation is provided in the supplementary section 2.4. We provided the latent place scatter-plot obtained for the contrast disentangling dataset using Harmony and other baseline models in Figure 6 with a visualization of some exemplary input images. From the latent plots (Figure 6),

the disentangling of facial identity and contrast is evident for Harmony, where traversing along one latent factor (plotted on y-axis) only affects the content and the other latent factor (plotted on x-axis) affect the contrast. None of the latent space plot for other baseline methods show such disentanglement. It is to be mentioned that Spatial-VAE [1] or U-VITAE [5], the well-performing methods for affine transformation disentangling, can not be used for contrast disentangling by design. These results justify Harmony's wide applicability and efficacy as a generic method.

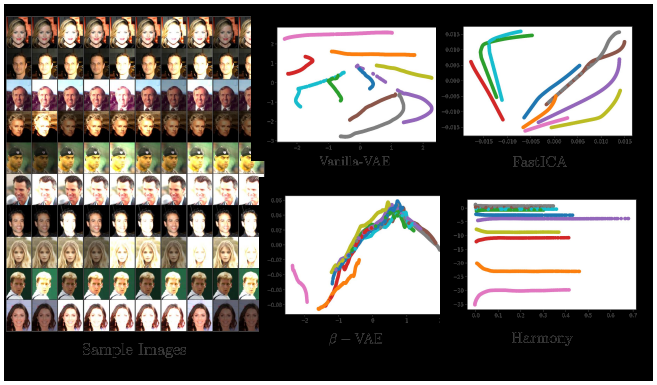

Figure 6. Disentangling Facial Identity from contrast (lighting condition symmetry transformation) from facial image dataset. The latent space plot for Harmony, FastICA, vanilla-VAE, and β-VAE are provided along with sample input images of the facial contrast dataset. Each color in the latent space plot corresponds to a distinct facial identity. In Harmony latent space, traversing along the y-axis (semantic latent factor) only changes the facial identity, while being unaffected by contrast. This disentanglement is not observed in latent space of the other methods.

We have also performed ablation experiments to assess the contribution of the cross-contrastive module and the KL loss component in objective function. The results (provided in the supplementary section S4) indicate their importance in disentangling content from transformations.

5. Discussion & Limitations

Despite Harmony showing pioneering results in disentangling 3D cryo-ET data, the disentanglement performance becomes poor in extremely low SNR (≤ 0.03) conditions. This may be caused by the sum of squared errors loss used between images. In future work, noise-robust networks and loss functions can be incorporated with Harmony to solve this problem. Moreover, despite Harmony can capture discrete conformational variability of macromolecules when the conformations have sufficient structural difference, their ability to capture subtle or continuous conformational differences is still not as good as template-based supervised methods HEMNMA-3D [10] or TomoFlow [11].

From our experiments we observe that the performance of Harmony is sensitive to the choice of encoder-decoder architecture in some specific datasets. This isn't surprising given the variety of datasets we used in our experiments. Furthermore, although Harmony separates semantic content and transformations along different latent factors very well, there exists some cross-semantic class overlap. In future work, this issue can be resolved by incorporating contrast with semantically dissimilar samples with a reasonable strategy.

6. Conclusion

Disentangling images into a shape specific content factor and shape-unspecific parameterized transformations is a critical task in many biomedical image analysis domains. In this work, we present *Harmony*, a novel unsupervised generic framework that can disentangle semantic content from multiple parameterized transformations. It operates by using cross-contrastive learning and the explicit decomposition of latent space into content and transformation factors. We used Harmony to decouple protein shapes from orientations and shifts in real cryo-EM images and recover structural heterogeneity of macromolecular shapes in 3D cryo-ET subtomograms. By disentangling content from transformations, Harmony can perform coarse and fast unsupervised groupwise alignment of cryo-ET subtomograms. Harmony is further tested against datasets of more generic image analysis domains, e.g., MNIST and face images. Our experiments show that Harmony can successfully disentangle semantic content from affine transformations (rotation, translation, scaling) and a lighting condition transformation (contrast). These promising results demonstrate that, Harmony is not just an impactful contribution to bioimage analysis research, but also an important step toward for fully utilizing disentangled representation learning for separating content and transformations in image analysis domains and domains beyond imaging (e.g, speech, text) as well.

Acknowledgment

This work was supported in part by U.S. NIH grants R01GM134020 and P41GM103712, NSF grants DBI-1949629 and IIS-2007595, and Mark Foundation for Cancer Research 19-044-ASP. We thank the computational resources support from AMD COVID-19 HPC Fund. X.Z. was supported in part by a fellowship from CMU CMLH. We thank Dr. Qiang Guo for sharing with us experimental rat-neuron culture tomogram which we used as real cryo-ET dataset.

References

[1] Tristan W Bepler, Ellen Zhong, Kotaro Kelley, Edward Brignole, and Bonnie Berger. Explicitly disentangling image

content from translation and rotation with spatial-vae. 2019. 1, 2, 3, 4, 5, 6, 7, 8

[2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments, 2021. 3

[3] Ricky TQ Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in vaes. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 2615–2625, 2018. 1, 3, 7

[4] Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994. 3

[5] Nicki Skafte Detlefsen and Søren Hauberg. Explicit disentanglement of appearance and perspective in generative models. In *33rd Conference on Neural Information Processing Systems*, 2019. 1, 2, 3, 4, 5, 6, 7, 8

[6] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020. 3

[7] Qiang Guo, Carina Lehmer, Antonio Martínez-Sánchez, Till Rudack, Florian Beck, Hannelore Hartmann, Manuela Pérez-Berlanga, Frédéric Frottin, Mark S Hipp, F Ulrich Hartl, et al. In situ structure of neuronal c9orf72 poly-ga aggregates reveals proteasome recruitment. *Cell*, 172(4):696–705, 2018. 6

[8] Sina Hajimiri, Aryo Lotfi, and Mahdieh Soleymani Baghshah. Semi-supervised disentanglement of class-related and class-independent factors in vae. *arXiv preprint arXiv:2102.00892*, 2021. 1, 2

[9] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011. 7

[10] Mohamad Harastani, Mikhail Eltsov, Amélie Leforestier, and Slavica Jonic. Hemnma-3d: Cryo electron tomography method based on normal mode analysis to study continuous conformational variability of macromolecular complexes. *Frontiers in molecular biosciences*, 8, 2021. 8

[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. 8

[12] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018. 4

[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. 2016. 1, 3, 7

[14] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000. 3, 7

[15] Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, 2019. 3

[16] Sergei V Kalinin, Shuai Zhang, Mani Valleti, Harley Pyles, David Baker, James J De Yoreo, and Maxim Ziatdinov. Disentangling rotational dynamics and ordering transitions in a system of self-organizing protein nanorods via rotationally invariant latent representations. *ACS nano*, 15(4):6471–6480, 2021. 2, 3

[17] Bo-Kyeong Kim, Sungjin Park, Geonmin Kim, and Soo-Young Lee. Semi-supervised disentanglement with independent vector variational autoencoders. *arXiv preprint arXiv:2003.06581*, 2020. 1, 2

[18] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018. 1, 3

[19] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations*, 2018. 3

[20] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations*, 2018. 5

[21] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999. 3

[22] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019. 5

[23] Ronald Carl Petersen, PS Aisen, Laurel A Beckett, MC Donohue, AC Gamst, Danielle J Harvey, CR Jack, WJ Jagust, LM Shaw, AW Toga, et al. Alzheimer's disease neuroimaging initiative (adni): clinical characterization. *Neurology*, 74(3):201–209, 2010. 1

[24] Takeshi Shakunaga and Kazuma Shigenari. Decomposed eigenface for face recognition under various lighting conditions. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001. 3

[25] Zhixin Shu, Mihir Sahasrabudhe, Riza Alp Guler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *Proceedings of the European conference on computer vision (ECCV)*, pages 650–665, 2018. 3, 7

[26] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991. 3

[27] Min Xu, Jitin Singla, Elitza I Tocheva, Yi-Wei Chang, Raymond C Stevens, Grant J Jensen, and Frank Alber. De novo

structural pattern mining in cellular electron cryotomograms. *Structure*, 27(4):679–691, 2019. 1

[28] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021. 3

[29] Xiangrui Zeng, Anson Kahng, Liang Xue, Julia Mahamid, Yi-Wei Chang, and Min Xu. Disca: high-throughput cryo-et structural pattern mining by deep unsupervised clustering. *bioRxiv*, 2021. 1

[30] Xiangrui Zeng and Min Xu. Gum-net: Unsupervised geometric matching for fast and accurate 3d subtomogram image alignment and averaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4073–4084, 2020. 6