Supplementary of 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

Contents

*Corresponding author

S1 Harmony with Sum of Squared Errors

Let x and x' be single dimensional real valued random variables. These random variables are i.i.d. transformed by transforming $x^{(\text{og})}$ by randomly drawn transformations from K. Further, let $\bar{x}_z$ and $\bar{x}_k$ denote $\mathbb{E}_x[x_z]$ and $\mathbb{E}_x[x_k]$ respectively - note that x_z and x_k are functions of x. Additionally, we choose $\mathcal{D}^{(1)}, \mathcal{D}^{(2)}$, and $\mathcal{D}^{(3)}$ to be squared errors.

Lemma S1.1. $\mathbb{E}[(x_k - x'_{k'})^2] = 2 \cdot \mathbb{V}[x_k]$

Proof.

$$\begin{aligned}
\mathbb{E}[(x_k - x'_{k'})^2] &= \mathbb{E}[((x_k - \bar{x}_k) + (\bar{x}_k - x'_{k'}))^2] \\
&= 2 \cdot \mathbb{E}[(x_k - \bar{x}_k)^2] + 2 \cdot \mathbb{E}[(x_k - \bar{x}_k)(\bar{x}_k - x'_{k'})] \\
&= 2 \cdot \mathbb{V}[x_k] - 2 \cdot \mathbb{E}[(x_k - \bar{x}_k)]^2 && \text{[i.i.d.]} \\
&= 2 \cdot \mathbb{V}[x_k] - 2 \cdot 0 \\
&= 2 \cdot \mathbb{V}[x_k]
\end{aligned}$$

■

Lemma S1.2. $\mathbb{E}[(x_z - x_k)^2] = \mathbb{V}[x_z] - 2 \cdot Cov(x_z, x_k) + \mathbb{V}[x_k] + (\bar{x}_z - \bar{x}_k)^2$

Proof.

$$\begin{aligned}
\mathbb{E}[(x_z - x_k)^2] &= \mathbb{E}[((x_z - \bar{x}_z) + (\bar{x}_z - \bar{x}_k) + (\bar{x}_k - x_k))^2] \\
&= \mathbb{E}[(x_z - \bar{x}_z)^2] + \mathbb{E}[(\bar{x}_z - \bar{x}_k)^2] + \mathbb{E}[(\bar{x}_k - x_k)^2] \\
&\quad + 2 \cdot \mathbb{E}[(x_z - \bar{x}_z)(\bar{x}_z - \bar{x}_k)] + 2 \cdot \mathbb{E}[((x_z - \bar{x}_z)(\bar{x}_k - x_k)] \\
&\quad + 2 \cdot \mathbb{E}[(\bar{x}_z - \bar{x}_k)(\bar{x}_k - x_k)] \\
&= \mathbb{V}[x_z] + (\bar{x}_z - \bar{x}_k)^2 + \mathbb{V}[x_k] + 2 \cdot 0 - 2 \cdot \text{Cov}(x_z, x_k) + 2 \cdot 0 \\
&= \mathbb{V}[x_z] - 2 \cdot \text{Cov}(x_z, x_k) + \mathbb{V}[x_k] + (\bar{x}_z - \bar{x}_k)^2
\end{aligned}$$

■

Proposition S1.3. $\mathbb{E}[\mathcal{L}(x, x')] \propto \mathbb{V}[x_z] - 2 \cdot Cov(x_z, x_k) + (1 + \lambda) \cdot \mathbb{V}[x_k] + (\bar{x}_z - \bar{x}_k)^2$

Proof.

$$\begin{aligned}
\mathbb{E}[\mathcal{L}(x, x')] &= 2 \cdot \mathbb{E}[(x_z - x_k)^2] + \lambda \cdot \mathbb{E}[(x_k - x'_{k'})^2] \\
&= 2 \cdot \mathbb{E}[(x_z - x_k)^2] + 2\lambda \cdot \mathbb{V}[x_k] && \text{[by S1.1]} \\
&\propto \mathbb{E}[(x_z - x_k)^2] + \lambda \cdot \mathbb{V}[x_k] \\
&= \mathbb{V}[x_z] - 2 \cdot \text{Cov}(x_z, x_k) + \mathbb{V}[x_k] + (\bar{x}_z - \bar{x}_k)^2 + \lambda \cdot \mathbb{V}[x_k] && \text{[by S1.2]} \\
&= \mathbb{V}[x_z] - 2 \cdot \text{Cov}(x_z, x_k) + (1 + \lambda) \cdot \mathbb{V}[x_k] + (\bar{x}_z - \bar{x}_k)^2
\end{aligned}$$

■

In the general case when x and x' are d dimensional and $\mathcal{D}^{(1)}, \mathcal{D}^{(2)}, \mathcal{D}^{(3)}$ are the Sum of Squared Errors (SSE). The covariance and variances described above become index wise. Further, due to linearity of expectation, in expectation, the loss becomes a sum of the expression in Proposition S1.3 over the dimensions, d.

Define a transformation function $\mathcal{T} : X \times K \to X$ to be reversible if and only if for any $x \in X$ and $k \in K$ there exist parameters k_+ such that $\mathcal{T}(\mathcal{T}(x, k), k_+) = x$ – many commonly used transformation functions such as rotating, scaling, and shifting are reversible.

Lemma S1.4. *If $\mathcal{T}$ is reversible, for any similar x and x' there exist parameters k_+, k'_+ such that $\mathcal{T}(x, k_+) = \mathcal{T}(x', k'_+)$*

Proof.

$$\begin{aligned} \exists k, k' \in K : x = \mathcal{T}(x^{(\mathrm{og})}, k) \wedge x' = \mathcal{T}(x^{(\mathrm{og})}, k') \qquad & [\text{similarity}] \Rightarrow \\ \exists k, k' \in K \exists k_+, k'_+ \in K : x^{(\mathrm{og})} = \mathcal{T}(\mathcal{T}(x^{(\mathrm{og})}, k), k_+) & \\ \wedge\, x^{(\mathrm{og})} = \mathcal{T}(\mathcal{T}(x^{(\mathrm{og})}, k'), k'_+) \qquad & [\text{reversibility}] \Rightarrow \\ \exists k_+, k'_+ \in K : \mathcal{T}(x, k_+) = \mathcal{T}(x', k'_+) & \end{aligned}$$

■

Observe, when $\mathcal{D}^{(2)}$ is the sum of squared errors, $\mathbb{E}\left[\mathcal{D}^{(2)}(x_k, x'_{k'})\right] = 0$, which is it's minimum, if and only if for any $\epsilon > 0$ $P\left(x_k =_\epsilon x'_{k'}\right) = 1$. As a consequence, $\mathbb{E}\left[\mathcal{D}^{(2)}(x_k, x'_{k'})\right]$ achieves its minimum if and only if there exists some $x^{(\mathrm{p})}$ such that $P\left(x_k =_\epsilon x^{(\mathrm{p})}\right) = 1$.

Proposition S1.5. *There exists f such that $\mathbb{E}\left[\mathcal{D}^{(2)}(x_k, x'_{k'})\right]$ is minimized*

Proof. Consider $f^* : X \to K \times \Psi$ such that for any transformed instance of $x^{(\mathrm{og})}$, x, for f^*'s decoded parameters $(k_x, \psi_x) = f^*(x)$, we have $\mathcal{T}(x, k_x) = x^{(\mathrm{og})}$ – such an f^* exists by S1.4. Then, setting $x^{(\mathrm{p})} = x^{(\mathrm{og})}$ gives $P\left(x_k =_\epsilon x^{(\mathrm{p})}\right) = 1$, so $\mathbb{E}\left[\mathcal{D}^{(2)}(x_k, x'_{k'})\right] = 0$, which is its minimum. ■

As a consequence, for a sufficiently expressive parameter space Θ, if the transformation function $\mathcal{T}$ is reversible, Harmony's encoder f_θ can minimize the dissimilarity loss $\mathcal{D}^{(2)}$ (once such example of a minimizer is $f_\theta = f^*$), and this minimum is achieved if and only if for every original sample $x^{(\mathrm{og})}$ and all of its randomly transformed instances x every proposed re-transformed instance x_k is arbitrarily close to $x^{(\mathrm{p})}$ (except for a set of 0 measure).

S2 Datasets

S2.1 2D single-particle cryo-EM datasets

Cryo-EM is an imaging technique that images multiple 2D projections of protein paticles after isolating and purifying them. It recovers high-resolution structure of known type of protein particles. The primary use of cryo-EM imaging is to determine electron density of the protein complexes from these set of 2D projection images. Since the purified proteins lay in random poses and shifts on the grid, the 2D view of the proteins are also confounded with a variety of orientations and shifts. Decoupling the orientations and shifts from the images and aligning them to recover better resolution structure is a crucial task in structural bioinformatics.

In our experiments to this end, we tested Harmony against two negative stain cryogenic electron microscopy (cryo-EM) image dataset. Among the two datasets, one contains the StrepMAB-Classic antibody and the other contains the CODH/ACS protein complex. Each image is 40 by 40 pixels and has a very low signal to noise ratio. To use a setting similar to spatial-VAE [1], the antibody dataset was split into 10,821 training and 2,705 testing images and the CODH/ACS dataset was split into 11,473 training and 2,868 testing images.

S2.2 3D cryo-ET subtomogram datasets

Cryo-Electron Tomography (cryo-ET) produces 3D views of cellular samples at nanometer resolution (< 4 nm) [13]. In cryo-ET, the cellular samples are not isolated and purified like cryo-EM. So the native organization of subcellular components remain intact. The cellular samples first vitrified at cryogenic temperature ($< -150°C$) and the sample is placed in a grid and thinned by cryo-focused-ion-beam milling [8]. Then the cellular sample is placed under a cryo-transmission electron microscope and electron beams are passed through it. The electrons are then detected by an electron detector. The detection results in a projection image. The projection image is formed as electrons are less likely to pass through a thick structural region. In cryo-ET, the cell sample is tilted through a series of angles, typically at 1° to 3° tilt step. At each angle view, a projection image is produced. From the projection images, whole 3D tomogram

is reconstructed. The reconstructed tomogram is of very large size, so the analysis is performed mainly on subtomogram level. Subtomograms are small subvolumes extracted from the whole 3D tomogram. Each subtomogram contains only one entire macromolecule. Subtomograms are extracted by biologists annotation or particle picking.

We tested 3D Harmony against five realistically simulated benchmark datasets by Zeng *et al.*[15] and a real cryo-ET subtomogram dataset by Qiang *et al.* [6]. The simulated dataset contains 5000 subtomograms of five representative macromolecular complexes: spliceosome (PDB ID: 5LQW), RNApolymerase-rifampicin complex (1I6V), RNA polymerase II elongation complex (6A5L), ribosome (5T2C), and capped proteasome (5MPA). There are 1000 subtomograms for each representative macromolecular classes in diverse orientations and shifts. Among the five simulated datasets, one is relatively clean (SNR 100) and four are with SNR close to the experimental conditions (0.1, 0.05, 0.03, and 0.01). Each subtomogram is of size 32^3 with voxel size 1.2 nm.

We further validated Harmony's efficacy against real cryo-ET subtomograms extracted from rat neuron culture [6]. In total 1095 ribosome subtomograms and 1527 double capped proteasome subtomograms were extracted by template matching [2] and biology expert annotation. Each subtomogram is of size 32^3 with voxel size 1.368 nm and 30° missing wedge (tilt angle range $-50°$ to $70°$).

S2.3 MNIST images

We tested Harmony against full MNIST dataset, and also a rotated, translated, and scaled version of MNIST digits. We created the later to assess whether Harmony can disentangle full affine matrix. To this end, we randomly picked a image for each digit class in MNIST and created randomly rotated (by angle uniformly sampled between $[-\pi, \pi]$), translated (by random shifts uniformly sampled between $[-4, 4]$), and scaled (by scale factors uniformly sampled between $[0.5, 1.5]$) copies for each digit. This resulted in a total of 30,000 digits.

S2.4 CelebA RGB Facial Images

For lighting condition (contrast) disentangling, we created a variant of celebA RGB image dataset. To this end, we randomly picked 10 images of 10 different facial identities in the celebA dataset. The images are 218 by 178 by 3 pixels (RGB). We down sampled the images to 128×128 for faster training with minimal affect on image quality. Then we changed contrast of the images with contrast factors uniformly sampled between $[0.5, 3]$. A contrast factor of 0 generates a complete blank image and a contrast factor higher than 3 tends to destroy the facial identity of the image. Consequently, $[0.5, 3]$ was a reasonable range. We generated $10,000$ images in total where each facial identity has an equal amount of transformed images.

S3 Implementation Details

We applied Harmony to disentangle parameterized transformations from two real cryo-EM datasets, several simulated and real 3D cryo-ET subtomogram dataset, MNIST and its variant datasets, and a variant of celebA facial image dataset. We used Convolutional Neural Networks (CNN) or Multilayer Perceptrons (MLP) to implement the encoder f_θ and decoder g_ϕ in our experiments with image data (Figure S1. We used kornia [12], an open source computer vision library in PyTorch, to implement the differentiable transformation operations, e.g., translation, rotation, scale, contrast, etc. Kornia's implementation uses spatial transformer [10] with bilinear interpolation for warping images to the applied transformations. We have used sum of squared errors (SSE) to measure image dissimilarity in Harmony's objective. We used β-VAE [7], Annealed-β-VAE [7], β-TC-VAE [4], and U-VIATE [5] as baseline models. As SpatialVAE disentangles content from translation and rotation explicitly for cryo-EM datasets, we compared our results with Spatial-VAE [1] and a 3D extension of Spatial-VAE [1] for cryo-EM and cryo-ET datasets respectively. All models were implemented in PyTorch. We trained the models using ADAM optimizer [11] with a learning rate of 0.0001 and a minibatch size of 100. Only for 3D implementation of spatial-VAE, we had to use minibatch size of 20 as larger batch size than that was causing memory issues. We used 3 NVIDIA RTX 3090 GPUs to train our models.

S3.1 Encoder and Decoder Architectures

The architectures for Harmony's encoder and decoder used in our experiments with imaging data are provided in Fig.S1. However, because Harmony is a general framework, different architectures are also suitable and our current architecture does not necessarily generate the best performance among all possible model architectures.

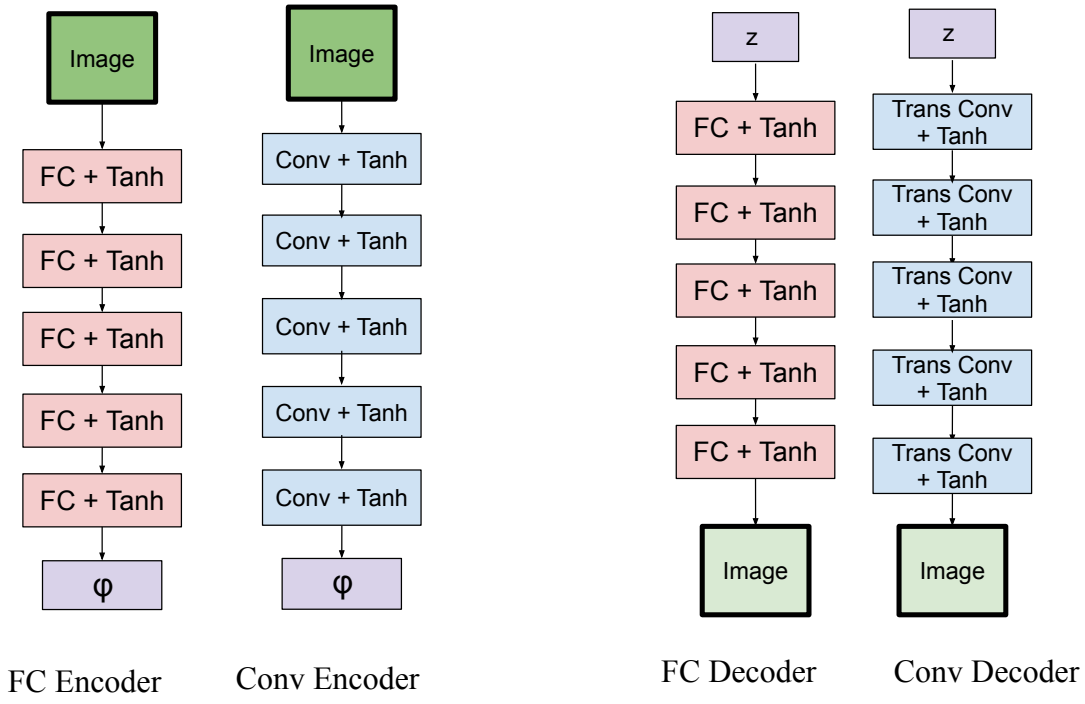


Figure S1: The encoder-decoder architecture for Harmony used in our experiments

S3.2 Choice of hyperparameters

There are a number of hyperparameters associated with Harmony. One of them is the dimension of semantic latent factor z. As 1-D z is enough to capture the variability of the semantic contents in the our datasets (except simulated cryo-ET), we set this to 1 in our experiments. In most of our experiments, increasing the dimension more than 1 does not significantly change the disentanglement. In Figure S3, we have demonstrated the effect of dimension of semantic latent factor z on disentanglement score for disentangling translation and rotation from the whole MNIST dataset. For simulated cryo-ET, setting the dimension as 50 provided best results. A hyperparameter introduced in our objective function is γ. We have observed setting γ too low (1) gives poor disentanglement. However, after a certain threshold, changing γ does not affect the overall disentanglement. The effect of γ on disentanglement score for disentangling translation and rotation from the whole MNIST dataset is demonstrated in Figure S2. In our experiments, we set γ as $\frac{M}{N \times 1000}$, where M is the number of data points and N is the batch size.

Another hyperparameter set is the transformation parameters $k \in K$, that we use to create distorted inputs for Harmony. In the case of rotation, the parameter sampled from Uniform$(-45°, 45°)$ or Uniform$(-90°, 90°)$ all resulted in similar results. For transformations like translation, scaling or lighting conditions, we use an appropriate uniform distribution range that does not change the semantic meaning of the images (translate it out of frame or contrast it so low that it becomes a blank image) as output.

S3.3 Effect of hyperparameters in disentangling

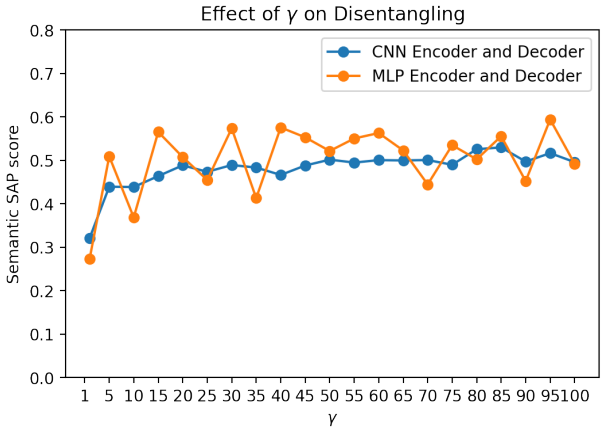


Figure S2: Plot of SAP score vs γ while keeping all other factors fixed.

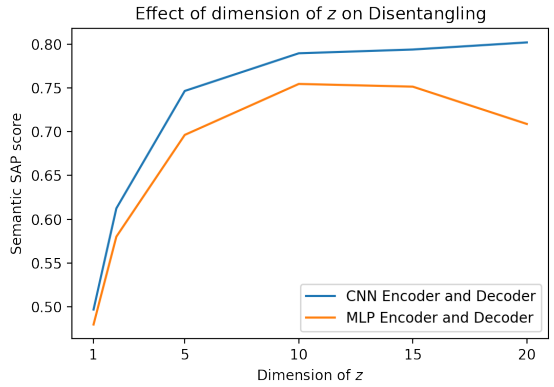


Figure S3: Plot of SAP score vs dimension of latent factor while keeping all other factors fixed.

S4 Ablation Experiments

We performed ablation experiments to demonstrate the contribution of the individual modules in Harmony. Particularly, we demonstrated the importance of per dimensional KL divergence in objective function and the cross-contrastive setting. Hence we evaluated two models- **(i)** Harmony without per dimensional KL loss in objective, **(ii)** Harmony without any augmented branch (with objective as $\mathcal{D}^{(1)}(x_z, x_k)$). In Table S1, we demonstrate the results of our ablation experiments against the 30,000 rotated, translated, and scaled version of MNIST digits and 10,000 diversely contrasted version of celebA images of 10 distinct facial identities. The results confirm the contribution of each individual modules in Harmony.

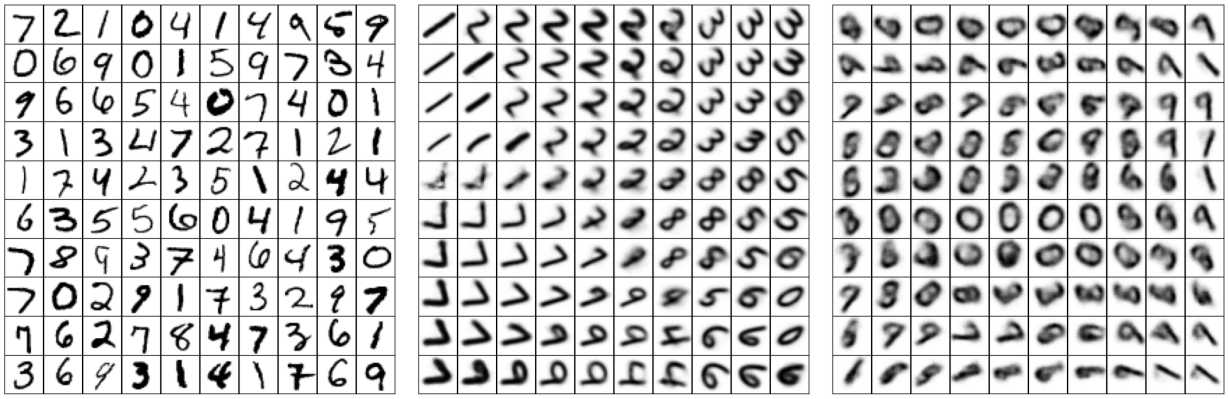
Method	**Dataset**	**SAP**	$P(\mathbf{c} \mid \mathbf{z})$
Harmony (no cross-contrastive)	MNIST (rotated, translated, scaled)	0.46	0.8796
	CelebA (contrast)	0.1628	0.7473
Harmony (no KL)	MNIST (rotated, translated, scaled)	0.5283	0.9118
	CelebA (contrast)	0.5554	0.9981
Harmony	MNIST (rotated, translated, scaled)	**0.55**	**0.944**
	CelebA (contrast)	**0.5586**	**1.0**

Table S1: Results from Ablation Experiments

S5 Additional Results

S5.1 Additional Results on Full Vanilla MNIST

We used Harmony to disentangle translation and rotation from full mnist dataset. From the images interpolated from semantic latent manifold of Harmony, it is evident that Harmony could successfully disentangle rotation and translation from digits (please note, it does not aim to disentangle digit writing style). From images interpolated from β-VAE or vanilla-VAE (β=1) it is evident that traversing along any of the latent axis would not generate digits unaffected by rotation and translation.



(a) (b) (c)

Figure S4: Disentangling rotation and translation from full original MNIST dataset. (a) shows some sample input images. (b) shows the images generated from semantic latent space generated by Harmony (with $\gamma = 70$ and dimension of z set as 2) (c) shows the images generated from the latent space of β-VAE.

S5.2 Additional Results on cryo-EM datasets

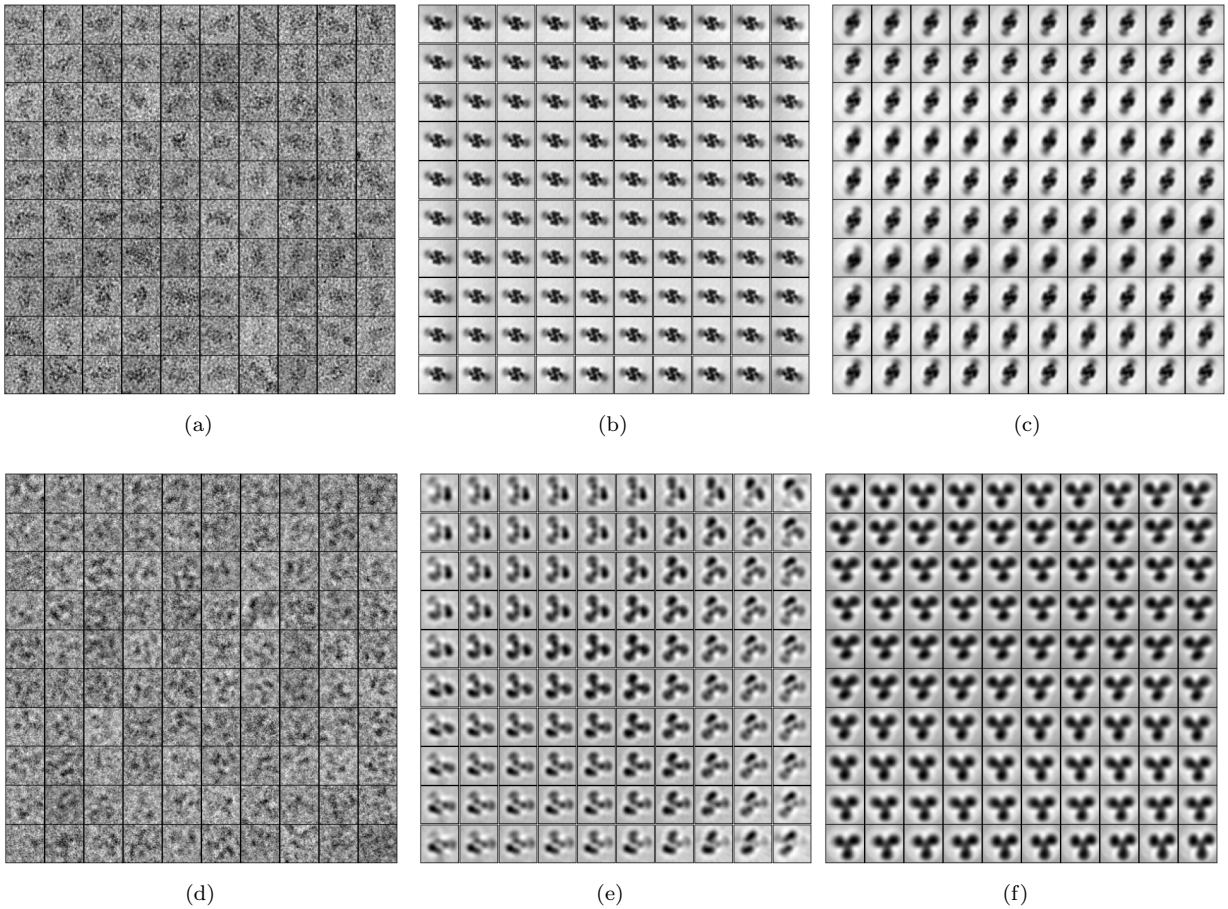


Figure S5: Learning Transformation Invariant Representation of protein complexes from cryo-EM images using Harmony. (a) and (d) shows exemplary cryo-EM images of CODH/ACS protein and antibody protein conformations respectively. (b) and (e) demonstrates the interpolated protein conformations from latent manifold learned by SpatialVAE [1] (dimension of semantic latent factors=2) for CODH/ACS dataset and antibody dataset respectively. (c) and (f) depicts the interpolated protein conformations from latent manifold learned by Harmony (dimension of semantic latent factors=1) for CODH/ACS dataset and antibody dataset respectively. Harmony can extract and align the conformations very well from the noisy images while performing denoising by the decoder. The SpatialVAE visualizations are generated by removing the orientation from the observed images. Further, random rotation augmentations were applied to the inference network training in Spatial-VAE to ensure its robustness. Nevertheless, spatial-VAE tends to capture structure in the image background and sometimes learn inconsistent representations for antibody dataset. None of these problems are encountered in Harmony.

S5.3 Additional Results on cryo-ET datasets

To cluster the simulated subtomograms, we have applied simple K-neareast neighbor algorithm (with k=5) to the semantic latent factor encoded by Harmony. In Table S2, we have compared the clustering results with state-of-the-art clustering methods. The accuracy values for other methods are taken from recent work by Zeng et al [14]. Cluster-only methods are taken into account only. After disentangling semantic content by Harmony, just performing a simple K-nearest neighbor on semantic latent factor gives better performance than many sophisticated deep clustering methods like DeepCluster and PICA. The results demonstrate the importance of disentangled representation learning in semantically meaningful feature extraction.

Method	**SNR 100**	**SNR 0.1**	**SNR 0.05**	**SNR 0.03**	**SNR 0.01**
DeepCluster [3]	68.7	48.8	39.4	34.0	27.2
PICA [9]	**100**	**86.5**	55.8	29.2	28.4
JimNet (cluster) [14]	99.5	77.5	62.8	*51.3*	**35.3**
Harmony + KNN	*99.6*	*86.1*	**63.0**	**52.7**	**47**

Table S2: Comparison of Harmony+KNN with state of the art deep clustering methods on simulated subtomograms.

The disentanglement by Harmony can be also used for coarse and fast alignment. However, unlike the subtomogram alignment methods, that align subtomogram to a density map or high resolution subtomogram, Harmony performs coarse and fast group wise alignment of subtomograms. For some specific classes, we have observed (Figure S6) the alignment is close to fine alignment. By performing groupwise alignment, it can create initial models for structural refinement [16].

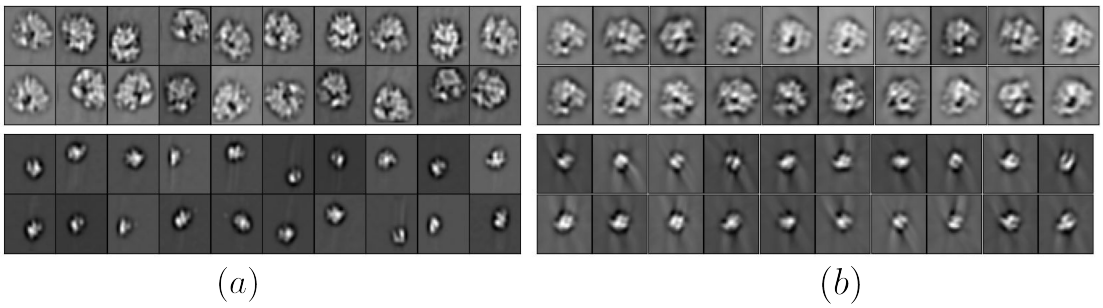


Figure S6: Detailed view of Unsupervised Coarse and Fast Subtomogram Alignment through disentangling 3D rigid body transformations from semantic content. (a) Sample 2D central slice images of RNA polymerase (PDB 1I6V) (above) and Ribosome (PDB 5T2C) (below) subtomograms from simulated dataset with SNR=100. (b) Corresponding output images from semantic latent space only (disentangled from transformations)

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.

[2] Jochen Böhm, Achilleas S Frangakis, Reiner Hegerl, Stephan Nickell, Dieter Typke, and Wolfgang Baumeister. Toward detecting and identifying macromolecules in a cellular context: template matching applied to electron tomograms. *Proceedings of the National Academy of Sciences*, 97(26):14245–14250, 2000.

[3] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features, 2019.

[4] 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.

[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.

[6] 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.

[7] 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.

[8] Chyongere Hsieh, Thomas Schmelzer, Gregory Kishchenko, Terence Wagenknecht, and Michael Marko. Practical workflow for cryo focused-ion-beam milling of tissues and cells for cryo-tem tomography. *Journal of structural biology*, 185(1):32–41, 2014.

[9] Jiabo Huang, Shaogang Gong, and Xiatian Zhu. Deep semantic clustering by partition confidence maximisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8849–8858, 2020.

[10] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2*, pages 2017–2025, 2015.

[11] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[12] Edgar Riba, Dmytro Mishkin, Daniel Ponsa, Ethan Rublee, and Gary Bradski. Kornia: an open source differentiable computer vision library for pytorch. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3674–3683, 2020.

[13] Jonathan Wagner, Miroslava Schaffer, and Rubén Fernández-Busnadiego. Cryo-electron tomography—the cell biology that came in from the cold. *FEBS letters*, 591(17):2520–2533, 2017.

[14] Xiangrui Zeng, Gregory Howe, and Min Xu. End-to-end robust joint unsupervised image alignment and clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3854–3866, 2021.

[15] 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.

[16] Jasenko Zivanov, Takanori Nakane, Björn O Forsberg, Dari Kimanius, Wim JH Hagen, Erik Lindahl, and Sjors HW Scheres. New tools for automated high-resolution cryo-em structure determination in relion-3. *elife*, 7:e42166, 2018.