DISSECT: DISentangle Sharable ConTent for Multimodal Integration and Crosswise-mapping

Geoffrey Schau, Erik Burlingame, Young Hwan Chang

Abstract—Deep learning systems have emerged as powerful mechanisms for learning domain translation models. However, in many cases, complete information in one domain is assumed to be necessary for sufficient cross-domain prediction. In this work, we motivate a formal justification for domain-specific information separation in a simple linear case and illustrate that a self-supervised approach enables domain translation between data domains while filtering out domain-specific data features. We introduce a novel approach to identify domain-specific information from sets of unpaired measurements in complementary data domains by considering a deep learning cross-domain autoencoder architecture designed to learn shared latent representations of data while enabling domain translation. We introduce an orthogonal gate block designed to enforce orthogonality of input feature sets by explicitly removing non-sharable information specific to each domain and illustrate separability of domain-specific information on a toy dataset.

I. INTRODUCTION

The task of domain translation broadly relates to identifying a pair or set of transfer functions that model a data point in one domain as an equivalent data point in some other domain. For many problems, data may be collected from two complementary modalities for the same instance or event, enabling familiar statistical approaches for inferring correlative or causal relationships between measurements made in two or more domains. However, in instances for which paired measurements are not available, mapping a representation of a measurement into another domain becomes a significantly more complex challenge.

These types of situations arise often in the field of biomedical research in which cellular assays inherently destroy or permanently modifies the sample under study such that no two assays can be applied to the same sample. For example, the recent emergence of single-cell genomic sequencing technologies enable molecular characterization of individual cells, but destroy those cells in the process, therefore their application to the same set of cells is generally mutually exclusive. Similarly, single-cell multiplexed imaging technologies enable deep spatial profiling of tissue, yet the permanent fixture of tissue limits the applicability of sequencing technologies following or preceding imaging. In these cases, unpaired single-cell resolution measurements are available, though the absence of paired measurements limits the degree to which biological insights may be translated between technological approaches.

In this work, we consider the case of unpaired domain translation. Presently, literature describing learning systems

for domain translation are largely applied to image-to-image translation settings. CycleGAN, built upon the generative adversarial framework [1], introduced the notion of cycle consistency to enable image-to-image domain translation in the absence of paired data [2], [3]. Although highly effective in certain settings, the authors recognize limitations of a purely GAN-based system, which experience common issues such as mode collapse [4]. Unlike GANs, variational autoencoders (VAEs) [5] learn optimal latent representations of input data by learning to generate reconstructions of an input data source rather than an arbitrary image capable of fooling a learned discriminator. Split-brain autoencoders [6] have recently been proposed as an alternative approach for learning cross-channel information of individual data samples, though the architecture has yet to be extended to the challenge of unpaired data samples from complementary domains. Shared latent representations between domains has been proposed using a pair of VAE models in both UNIT and MUNIT learning architectures [7], [8], though neither approach seeks to disentangle shared factors of variation.

In this paper, we consider the domain translation task between independent data domains that each describe the same underlying unit of analysis but from distinct, yet not strictly orthogonal, projections while estimating the degree to which complementary domains inform each other. First, we consider a simple linear case and illustrate the benefit of identifying domain-specific information. Second, motivated by theoretical derivation from a linear case, we propose a self-supervised deep learning-based approach that leverages unpaired data across two domains to learn crosswise mapping while disentangling mutual information content from distinct profiling modalities at single-cell resolution on a toy dataset.

II. PROBLEM FORMULATION

We consider a multimodal dataset where X_1 and X_2 are measurements from two different domains. For given $X_1 \in \mathbb{R}^m$ and $X_2 \in \mathbb{R}^n$, we are interested in identifying a crosswise mapping $\Phi(\cdot)$, $\Psi(\cdot)$ as shown in Figure 1.

A. Disentangling domain-specific and sharable features to identify a crosswise mapping

In each domain, there exist domain-specific (i.e., x_{ii}, x_{jj}), and sharable contents/features (i.e., x_{ij}, x_{ji}) and we assume that they are disjoint, i.e., $x_{ii} \perp x_{ij}$ or x_{ji} . To split domain-specific and sharable content, we introduce a simple transformation W and assume that it is an orthonormal matrix (i.e., $W^T W = I$ and inner product $(\cdot, \cdot)$, $(w_i, w_j) = 0$ for $i \neq j$ and $(w_i, w_j) = 1$ for $i = j$) where $W_1 \in \mathbb{R}^{m \times m}$ and

Fig. 1. A conceptual diagram of crosswise mapping: (top) crosswise mapping without dissecting domain-specific features (similar to split-brain autoencoder) (bottom) $\Phi(\cdot)$, $\Psi(\cdot)$ are any mappings (including autoencoder) where solid line represents shared network and dot line represent non-shared network.

$W_2 \in \mathbb{R}^{n \times n}$ are orthogonal matrices:

$$WX = \begin{bmatrix} W_1 & 0_{m \times n} \\ 0_{n \times m} & W_2 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} W_1 X_1 \\ W_2 X_2 \end{bmatrix} \triangleq \begin{bmatrix} x_{11} \\ x_{12} \\ x_{21} \\ x_{22} \end{bmatrix}$$

One could consider that W as a data sorting transform where each row of W consists of zeros and 1, i.e., standard basis, where corresponding nonzero component is chosen for either domain-specific or shared data set. Also, we consider the inverse mapping in a similar way:

$$\begin{aligned} \hat{X} = \begin{bmatrix} \hat{X}_1 \\ \hat{X}_2 \end{bmatrix} &= W^{-1} \begin{bmatrix} \hat{x}_{11} \\ \hat{x}_{12} \\ \hat{x}_{21} \\ \hat{x}_{22} \end{bmatrix} = W^\top \begin{bmatrix} \hat{x}_{11} \\ \hat{x}_{12} \\ \hat{x}_{21} \\ \hat{x}_{22} \end{bmatrix} \\ &= \begin{bmatrix} W_1^\top & 0 \\ 0 & W_2^\top \end{bmatrix} \begin{bmatrix} \hat{x}_{11} \\ \hat{x}_{12} \\ \hat{x}_{21} \\ \hat{x}_{22} \end{bmatrix} = \begin{bmatrix} W_1^\top & 0 \\ 0 & W_2^\top \end{bmatrix} \begin{bmatrix} \hat{x}_{11} \\ \hat{x}_{12} \\ \hat{x}_{21} \\ \hat{x}_{22} \end{bmatrix} \end{aligned} \quad (1)$$

where $\hat{X}$ represents the reconstructed data from input data X with a crosswise mapping ($\Phi(\cdot)$, $\Psi(\cdot)$) as shown in Figure 1.

Lemma 1: W transformation does not affect the two-norm, i.e., $\|\tilde{X}\|_2 = \|W\tilde{X}\|_2$

Proof : by definition, i.e., satisfies orthogonal invariance. $\|W\tilde{X}\|_2^2 = (W\tilde{X})^\top (W\tilde{X}) = \tilde{X}^\top W^\top W\tilde{X} = \tilde{X}^\top \tilde{X} = \|\tilde{X}\|_2^2$

To learn crosswise mapping across two domains, first we need to identify sharable content and split them into two disjoint models or networks as shown in Figure 1 (bottom). To identify crosswise mapping (solid arrow), it is trained to perform prediction on one subset of data from another subset, similar to split-brain autoencoder [6]. For non-shared model (dotted arrow), we basically bypass them (mapping is an identify matrix or a simple autoencoder, i.e., input is the same as output). Note that if we do not consider disentanglement of sharable and domain-specific information (Figure 1 (top)),

it is difficult to learn crosswise-mapping due to the non-sharable information across two domains.

Lemma 2: An orthonormal transformation W reduces reconstruction error ($\|\tilde{X}\|_2 \triangleq \|X - \hat{X}\|_2$) by bypassing disjoint information to non-shared network.

Proof : With W transformation, assume that ϕ and ψ are the optimal crosswise mappings to minimize the reconstruction error:

$$(P1) : \min_{W, \phi, \psi} \|\tilde{X}\|_2^2$$

Then, with two disjoint networks, the reconstruction error is as follow:

$$\begin{aligned} \|\tilde{X}\|_2^2 &= \|W\tilde{X}\|_2^2 = \|W(X - \hat{X})\|_2^2 = \left\| \begin{bmatrix} x_{11} - \hat{x}_{11} \\ x_{12} - \phi(x_{21}) \\ x_{21} - \psi(x_{12}) \\ x_{22} - \hat{x}_{22} \end{bmatrix} \right\|_2^2 \\ &= \underbrace{\|x_{11} - \hat{x}_{11}\|_2^2}_0 + \|x_{12} - \phi(x_{21})\|_2^2 + \|x_{21} - \psi(x_{12})\|_2^2 \\ &\quad + \underbrace{\|x_{22} - \hat{x}_{22}\|_2^2}_0 = \|x_{12} - \phi(x_{21})\|_2^2 + \|x_{21} - \psi(x_{12})\|_2^2 \end{aligned}$$

Without splitting into two disjoint networks (i.e., no bypass term), assume that $\Phi = (\phi_1, \phi_2)$ and $\Psi = (\psi_1, \psi_2)$ are the optimal mappings to minimize the reconstruction error:

$$(P2) : \min_{\Phi, \Psi} \|\tilde{X}\|_2^2$$

$$\begin{aligned} \|\tilde{X}\|_2^2 &= \|W\tilde{X}\|_2^2 = \left\| \begin{bmatrix} x_{11} \\ x_{12} \\ x_{21} \\ x_{22} \end{bmatrix} - \begin{bmatrix} \Phi \begin{bmatrix} x_{21} \\ x_{22} \end{bmatrix} \\ \Psi \begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix} \end{bmatrix} \right\|_2^2 \\ &= \left\| \begin{bmatrix} x_{11} - \phi_1 \begin{bmatrix} x_{21} \\ x_{22} \end{bmatrix} \\ x_{12} - \phi_2 \begin{bmatrix} x_{21} \\ x_{22} \end{bmatrix} \\ x_{21} - \psi_1 \begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix} \\ x_{22} - \psi_2 \begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix} \end{bmatrix} \right\|_2^2 \\ &= \left\| x_{11} - \phi_1 \begin{bmatrix} x_{21} \\ x_{22} \end{bmatrix} \right\|_2^2 + \left\| x_{12} - \phi_2 \begin{bmatrix} x_{21} \\ x_{22} \end{bmatrix} \right\|_2^2 \\ &\quad + \left\| x_{21} - \psi_1 \begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix} \right\|_2^2 + \left\| x_{22} - \psi_2 \begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix} \right\|_2^2 \\ &\geq \left\| x_{12} - \phi_2 \begin{bmatrix} x_{21} \\ x_{22} \end{bmatrix} \right\|_2^2 + \left\| x_{21} - \psi_1 \begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix} \right\|_2^2 \\ &= \|x_{12} - \phi(x_{21})\|_2^2 + \|x_{21} - \psi(x_{12})\|_2^2 \quad (P1) \end{aligned} \quad (2)$$

In the last equality sign, since we assume that x_{22} and x_{11} are disjoint from x_{12} and x_{21} (i.e., domain specific content) by definition, we can ignore these components.

Lemma 2 shows that (P1) provides the lower reconstruction error compared to (P2). Note that in Equation (2) (before the inequality sign), the first and last terms can cause huge reconstruction error since x_{11} and x_{22} are domain-specific

features by definition and x_{ii} and x_{ij} (or x_{ji}) are disjoint by the assumption, and thus this can be infinite value (i.e., no crosswisemapping from domain i to domain j). On the other hand, W allows us to dissect domain-specific features (x_{11}, x_{22}) and identify the shared (x_{12}, x_{21}) information across multimodal datasets by using two disjoint networks, rather than attempting to force highly divergent datasets into a completely shared latent space mapping (P2).

One may argue that (P1) may bypass the shared information to reduce the reconstruction error, for example, making x_{12} and x_{21} are empty set since we do not penalize any reconstruction error in non-shared part. To address this question, we consider the following case.

First, we denote $x_{12} = \begin{bmatrix} x_{12}^b \\ x_{12}^s \end{bmatrix}$, $x_{21} = \begin{bmatrix} x_{21}^b \\ x_{21}^s \end{bmatrix}$ where superscript $\{\cdot\}^b$ and $\{\cdot\}^s$ represent the shared feature which go through “bypass” and “shared” networks respectively. Second, we assume that there exists $(\bar{\phi}, \bar{\psi})$ such that $\|x_{12}^b - \hat{x}_{12}^b\| = \left\| x_{12}^b - \bar{\phi} \left(\begin{bmatrix} x_{21}^b \\ x_{21}^s \end{bmatrix} \right) \right\| = 0$ and $\|x_{21}^b - \hat{x}_{21}^b\| = \left\| x_{21}^b - \bar{\psi} \left(\begin{bmatrix} x_{12}^b \\ x_{12}^s \end{bmatrix} \right) \right\| = 0$, i.e., there exist a perfect crosswise mapping.

If we bypass this shared information, the reconstruction error can be written as follows: $J_1 = \|x_{12}^s - \phi(x_{21}^s)\|_2^2 + \|x_{21}^s - \psi(x_{12}^s)\|_2^2$. If we do not bypass this shared information, the reconstruction error can be written as follows:

$$\begin{aligned} J_2 &= \|x_{12} - \hat{x}_{12}\|_2^2 + \|x_{21} - \hat{x}_{21}\|_2^2 \\ &= \left\| \begin{bmatrix} x_{12}^b \\ x_{12}^s \end{bmatrix} - \phi \left(\begin{bmatrix} x_{21}^b \\ x_{21}^s \end{bmatrix} \right) \right\|_2^2 + \left\| \begin{bmatrix} x_{21}^b \\ x_{21}^s \end{bmatrix} - \psi \left(\begin{bmatrix} x_{12}^b \\ x_{12}^s \end{bmatrix} \right) \right\|_2^2 \\ &= \left\| x_{12}^b - \phi_b \left(\begin{bmatrix} x_{21}^b \\ x_{21}^s \end{bmatrix} \right) \right\|_2^2 + \left\| x_{12}^s - \phi_s \left(\begin{bmatrix} x_{21}^b \\ x_{21}^s \end{bmatrix} \right) \right\|_2^2 \\ &\quad + \left\| x_{21}^b - \psi_b \left(\begin{bmatrix} x_{12}^b \\ x_{12}^s \end{bmatrix} \right) \right\|_2^2 + \left\| x_{21}^s - \psi_s \left(\begin{bmatrix} x_{12}^b \\ x_{12}^s \end{bmatrix} \right) \right\|_2^2 \end{aligned}$$

where we denote $\phi = (\phi_b, \phi_s)$ and $\psi = (\psi_b, \psi_s)$.

Lemma 3: If there exists a crosswise mapping $(\bar{\phi}, \bar{\psi})$ such that it guarantees zero or negligible reconstruction error (i.e., exact crosswise mapping), then bypassing those shared contents will be suboptimal, i.e., a larger reconstruction error, $J_1 \geq J_2$.

Proof:

$$\begin{aligned} J_1 &= \|x_{12}^s - \phi(x_{21}^s)\|_2^2 + \|x_{21}^s - \psi(x_{12}^s)\|_2^2 \\ &\geq \left\| x_{12}^s - \phi' \left(\begin{bmatrix} x_{21}^b \\ x_{21}^s \end{bmatrix} \right) \right\|_2^2 + \left\| x_{21}^s - \psi' \left(\begin{bmatrix} x_{12}^b \\ x_{12}^s \end{bmatrix} \right) \right\|_2^2 \\ &= \left\| x_{12}^s - \phi' \left(\begin{bmatrix} x_{21}^b \\ x_{21}^s \end{bmatrix} \right) \right\|_2^2 + \left\| x_{21}^s - \psi' \left(\begin{bmatrix} x_{12}^b \\ x_{12}^s \end{bmatrix} \right) \right\|_2^2 \\ &\quad + \|x_{12}^b - \hat{x}_{12}^b\| + \|x_{21}^b - \hat{x}_{21}^b\| \\ &= \left\| x_{12}^s - \phi' \left(\begin{bmatrix} x_{21}^b \\ x_{21}^s \end{bmatrix} \right) \right\|_2^2 + \left\| x_{21}^s - \psi' \left(\begin{bmatrix} x_{12}^b \\ x_{12}^s \end{bmatrix} \right) \right\|_2^2 \\ &\quad + \left\| x_{12}^b - \bar{\phi} \left(\begin{bmatrix} x_{21}^b \\ x_{21}^s \end{bmatrix} \right) \right\|_2^2 + \left\| x_{21}^b - \bar{\psi} \left(\begin{bmatrix} x_{12}^b \\ x_{12}^s \end{bmatrix} \right) \right\|_2^2 \\ &= J_2 \text{ (i.e., } \phi' = \phi_s, \bar{\phi} = \phi_b, \psi' = \psi_s, \bar{\psi} = \psi_b) \end{aligned} \quad (3)$$

In the first inequality, we consider a mapping ϕ', ψ' with keeping bypass term (x_{21}^b, x_{12}^b) and without loss of generality, this mapping provides a lower error bound. Therefore, bypassing the shared information will result in the suboptimal solution where the shared information is necessary for reducing the reconstruction error in shared network. By Lemma 3, the optimization problem (P1) will identify W which only allows to bypass non-shared content information.

B. Cross-domain cycle consistency to deal with unpaired dataset

The work presented herein leverages unpaired data across two domains to learn crosswise mapping while disentangling mutual information content from distinct profiling modalities. To address the challenge of unpaired data samples from complementary domains, we consider a cycle-consistency concept as free supervisory signal for learning a crosswise mapping across two domains. If we identify the exact crosswise mapping for the identified shared information, both $\Phi(\cdot)$ and $\Psi(\cdot)$ satisfy $\Phi(\Psi(\cdot)) = I_m(\cdot)$ and $\Psi(\Phi(\cdot)) = I_n(\cdot)$ where $\Phi: \mathbb{R}^n \rightarrow \mathbb{R}^m$ and $\Psi: \mathbb{R}^m \rightarrow \mathbb{R}^n$. Here we consider a simple linear case in order to build a theoretical framework to address questions of why we need disentanglement and cross-domain consistency for building deep learning models.

Example: a linear mapping case

Consider $u = u_1 \oplus u_2 \in \mathbb{R}^m$ and $v = v_1 \oplus v_2 \in \mathbb{R}^n$ where subscript $\{\cdot\}_1$ represents sharable content and subscript $\{\cdot\}_2$ represent domain specific information, and $\oplus$ represents the direct sum. Define continuous linear maps $f: u_1 \rightarrow v_1$ and $g: v_1 \rightarrow u_1$ and consider cyclic consistency loss:

$$\mathcal{L}_{cyc} \triangleq \|u_1 - g(f(u_1))\|^2 + \|v_1 - f(g(v_1))\|^2 \quad (4)$$

Proposition 1: $u_1 = g(v_1)$ and $v_1 = f(u_1)$ is the optimal solution for minimizing cyclic consistency loss (4).

Proof:

$$\begin{aligned} \mathcal{L}_{cyc} &= \|u_1 - g(v) + g(v) - g(f(u_1))\|^2 \\ &\quad + \|v_1 - f(u) + f(u) - f(g(v_1))\|^2 \\ &= \|u_1 - g(v)\|^2 + \|g(v) - g(f(u_1))\|^2 + \|v_1 - f(u)\|^2 \\ &\quad + \|f(u) - f(g(v_1))\|^2 + 2\langle u_1 - g(v), g(v) - g(f(u_1)) \rangle \\ &\quad + 2\langle v_1 - f(u), f(u) - f(g(v_1)) \rangle \end{aligned}$$

Plug in $u = u_1 + u_2 = g(v_1) + u_2$ and $v = v_1 + v_2 = f(u_1) + v_2$, then

$$\begin{aligned} \mathcal{L}_{cyc}|_{u_1=g(v_1), v_1=f(u_1)} &= \\ &= \|-g(v_2)\|^2 + \|g(v_2)\|^2 - 2\langle g(v_2), g(v_2) \rangle \\ &\quad + \|-f(u_2)\|^2 + \|f(u_2)\|^2 - 2\langle f(u_2), f(u_2) \rangle = 0 \end{aligned}$$

Remark 1. Since $u_1 \in \mathcal{R}(g)$ and u_2 is the orthogonal complement in $\mathbb{R}^m$, $u_2 \in \mathcal{R}(g)^\perp = \mathcal{N}(g^*)$ by definition where $\mathcal{R}(\cdot)$ represents the range space, $\mathcal{N}(\cdot)$ represents the null space, and $g^*(\cdot)$ represents the adjoint of $g(\cdot)$. Similarly, $v_2 \in \mathcal{R}(f)^\perp = \mathcal{N}(f^*)$.

Fig. 2. A continuous linear mapping case where crosswise mappings $f : u_1 \rightarrow v_1$ and $g : v_1 \rightarrow u_1$ are adjoint to each other.

Remark 2. Since the optimal solution of (4) is $u_1 = g(v_1)$, $g(v_2) = 0$ holds, i.e., $v_2 \in \mathcal{N}(g)$ and thus $g = f^*$. This also satisfies $f(u_2) = 0$ since $u_2 \in \mathcal{N}(g^*) = \mathcal{N}(f)$.

Theorem 1: For a linear crosswise mapping case, the optimal crosswise mappings (f, g) for (4) are the adjoint to each other, i.e., $f = g^*$ and $g = f^*$.

Proof: By Proposition 1 and Remark 1 and 2.

III. EXTENSION TO DEEP LEARNING MODEL

Motivated by the theoretical derivation and result from the linear mapping case, we extend this conceptual framework to deep learning practice by utilizing a cross-domain autoencoding architecture and a separate gate layer to identify domain-specific information in a self-supervised manner. We also consider a cycle-consistency concept as free supervisory signal for learning a crosswise mapping (f, g) across two domains.

A. XAE Learning Architecture

To address cross-domain translation in a deep learning setting, we consider crossed autoencoder (XAE) architecture

Fig. 3. Conceptual illustration of architecture: we introduce an Orthogonal Gate Block to disentangle shared (A^* , B^*) and unshared ($A^\perp$, $B^\perp$) information in complementary domains during domain translation. Our approach seeks to minimize the cycle consistency loss between an input and its autoencoded reconstruction (red dot to red dot, in A, Eq. (5) and cycled reconstruction (red dot to purple dot in A, Eq. (6)). Concordant latent representation is encouraged by minimizing cross-domain representations (red dot and blue dot in z , Eq. (7)).

as a pair of domain-specific autoencoder networks composed of encoders Φ_A and Φ_B and decoders Ψ_A and Ψ_B that independently learn latent representations of data from independent domains A and B in a joint latent space z as shown in Fig. 3. The self-supervised autoencoder system learns to minimize domain-specific autoencoder reconstruction loss while concurrently learning to minimize cross-domain cycle consistency loss following domain translation through the shared latent representation and performs cross-domain translation by swapping the decoder networks during evaluation such that $f_{A \rightarrow B}(A_i) = \Psi_B(\Phi_A(A_i))$ and $f_{B \rightarrow A}(B_i) = \Psi_A(\Phi_B(B_i))$.

The learning objective for the XAE model includes terms to jointly optimize both domain-specific autoencoders as well as the bi-directional cycle consistent domain translation. Each domain-specific autoencoder learns to minimize a variational autoencoder objective which is composed of two terms in which the first term seeks to minimize the error between an input and its reconstruction while the second term seeks to minimize the Kullback-Leibler (KL) divergence between the distribution of learned latent features and an expected normal prior, such that

$$\mathcal{L}_{\text{VAE}_A} = \|A_i - \Psi_A(\Phi_A(A_i))\|^2 + \text{KL}(p(\Phi_A(A_i)) \| \mathcal{N}(0, 1)) \quad (5)$$

Further, the model learns to minimize domain cycle consistency by minimizing the reconstruction error of an input sample and its cycled reconstruction such that

$$\mathcal{L}_{\text{CCA}} = \|A_i - \Psi_A(\Phi_B(\Psi_B(\Phi_A(A_i))))\|^2 \quad (6)$$

Lastly, because a single data point represented by either encoder should be represented by approximately the same latent representations, we introduce a novel mutual encoding divergence loss term to penalize mutual representational error such that

$$\mathcal{L}_{\text{MED}} = \|(\Phi_A(A_i) - \Phi_B(\Psi_B(\Phi_A(A_i))))\|^2 \quad (7)$$

Lambda coefficients are chosen to approximate equally weighted loss term contributions, such that the full XAE learning objective is given by

$$\mathcal{L}_{\text{XAE}} = \lambda_{\text{vae}}(\mathcal{L}_{\text{VAE}_A}) + \lambda_{\text{vae}}(\mathcal{L}_{\text{VAE}_B}) + \lambda_{\text{cc}}(\mathcal{L}_{\text{CCA}}) + \lambda_{\text{cc}}(\mathcal{L}_{\text{CC}_B}) + \lambda_{\text{med}}(\mathcal{L}_{\text{MED}_A}) + \lambda_{\text{med}}(\mathcal{L}_{\text{MED}_B}) \quad (8)$$

In this example the naive model assumes that the full information content provided in a given sample should be employed to generate cross-domain predictions, however in cases where this is not expected, we desire a mechanism to separate domain-specific from domain-shared information. Models are optimized with the Adam optimizer [9] with base learning rate of $1e-3$ for 10 epochs.

B. Orthogonal Gate Block

To consider autoencoder type mappings for ϕ and ψ , it is not straightforward to deal with dynamic changes of architectures for domain-specific and sharable content. To disentangle domain-specific information from information shared between domains, mutually informative data features

are estimated through the introduction of a novel Orthogonal Gate Block (OGB). As shown in Figure 3, the OGB is composed of a single regularized element-wise multiplication layer $\mathcal{W}$, a non-trainable yet programmable binary element-wise multiplication layer $\mathcal{B} \in [0, 1]$, and a hyperbolic tangent activation function. The OGB appends noise sampled from the input domain's respective data space to the input tensor and performs element-wise multiplication operations on an input data point X such that the layer's function is defined as $f(X) = \tanh(X \odot \mathcal{W} \odot \mathcal{B})$ in which $\mathcal{B}$ enforces orthogonality between separable and non-separable content.

Models are initiated with random weights, though after initial epochs of training, the magnitudes of the learned weights are used to discriminate between meaningful data features and noise to enable cross-domain translation given an information bottleneck in z . First, features are randomly selected from the relevant data spaces and appended to the input data tensor. Then, after a fixed number of training epochs, the OGB evaluates the distribution of the gate weights associated with noise features and sets a threshold cutoff, which we set experimentally to be the 50% quantile of the noise-associated weight distribution, though this value is dependent on the question under inquiry. Lastly, the OGB explicitly sets its binary layer to remove either data or noise features whose learned gate weight magnitude is lower than the established cutoff and in so doing enforces orthogonality onto the feature set that explicitly passes through relevant features while filtering out features that are learned to be lowly weighted in performing the cross-domain translation task.

Lemma 4: Magnitude-based Weight Pruning satisfies an orthonormal transformation condition.

Proof : (by definition) Magnitude-based weight pruning result in sparse connection, a fully connected layer whose parameters' values are either zero or 1 and thus, it only allows bijective connection. Then we can consider it as W where each row of W , w_i has only one non-zero component and $w_i^\top w_j = 0$ where $i \neq j$.

$$W^\top W = \sum_{i=1}^n e_i^\top e_i = I_n \quad (9)$$

where e_i represents the standard basis or unit vector of n -dim space.

IV. RESULTS AND DISCUSSION

To evaluate the efficacy of our proposed approach, we consider a simple dataset with perfectly overlapping mutual information which is modeled using the MNIST dataset of hand-written digits in 28x28 pixel image (2D) form, and a vectorized (1D) 784x1 pixel form. We perform a series of experiments by adding increasing volumes of noise to the original 784 pixel one-dimensional input tensor. We train several XAE models for 100 epochs on the two unpaired data sets described above. Figure 4.A illustrates the magnitude of learned weights in the orthogonal gate block at the end of training for five experiments each run with a different proportion of noise added to the data (0, 20%, 40%, 60%, and 80%). In all cases, we observe attenuation of noise features

by their learned respective weights which may indicate that the OGB is learning to limit the importance of these features to the domain translation task. The squared magnitude of the learned weights serve as a reasonable discriminatory variable, as shown by the receiver operating characteristic curves in Figure 4.B, which we do not expect to perfectly discriminate between data and noise because several features of the input data such as the extreme perimeter of the MNIST images, are not expected to be important for domain translation.

Because mutual information between the two domains is perfectly known, we fairly compare the domain translation prediction with knowable ground truth. Figure 4.C compares the cross-entropy between the predicted translations and their corresponding ground truth with and without the addition of the gate layer across experiments with varying noise burdens. Although slight, we observe a consistent improvement in the cross-entropy between prediction and ground truth with the addition of the OGB, suggesting that the layer's removal of less-informative features facilitates better cross-domain prediction. In this example, the added noise was applied only to the 1D input data to illustrate a domain-specific affect of additive noise. Figure 4.D illustrate the resulting weights and binary layer of the OGB after training, which clearly illustrate that in general the OGB appears to automatically identify the center of the image as being relevant to the domain translation task. The accompanying figure illustrates four examples of a single input image and its various representations through the learning framework.

For visualization, we consider the original MNIST digit at the far left, followed by its direct autoencoded reconstruction, which is analogous to a typical variational autoencoder with characteristic softening of high-frequency detail. The third column illustrates the domain-translated sample, generated as a 1D vector but rearranged here to illustrate a comparable 2D representation, while the fourth column illustrates a cycle-consistent reconstruction of the input data point following domain translation. Note that these images are created by reconstructing a 2D image from the vectorized 1D OGB layer.

This work may be applicable to many computational tasks for which paired measurements from two complementary domains are absent, and in situations where measurements from either domain are expected to contain both domain-specific and sharable information content between the two. As deep learning methods continue to facilitate improved methods for domain translation, we anticipate a growing need for quantifying and identifying which features are mutually informative between complementary data domains. The approach presented herein is presently applied to a familiar benchmark dataset, but additional work is necessary to establish the generalizability of the principles set forth in this work.

With expected applications in the biomedical domain, we anticipate the need for synthetic biological data that accurately models unpaired domain-specific measurements with which to evaluate our approach. Future studies will

Fig. 4. Evaluation of the XAE learning system. (A) learned gate weights of actual data (green) and injected noise (orange) along the long axis of vectorized images illustrates a tendency for the learning system to lowly weight features associated with noise. (B) ROC curves of data/noise separability strictly by learned gate weight. (C) Cross-entropy between an input in one domain and its prediction in the other illustrates an advance of utilizing the OGB in this particular domain translation task. (D) Illustration of learned gate weights and the learned corresponding binary layer illustrates that the OGB learned to identify features associated with the center of the images while ignoring the perimeter of the MNIST data set.

incorporate additional synthetic data to further evaluate the capacity for this class of learning system to identify domain-specific features from complementary profiling technologies.

V. CONCLUSION

In this work, we consider the domain translation problem in the absence of paired measurements. We motivate the need for identifying sharable and non-sharable information between complementary data domains when considering cross-domain translation. We introduce a neural network layer designed to separate domain-specific information when learning optimal domain translation models by appending noisy data features and utilizing the magnitude of feature-specific weighting to remove noise and non-informative features. On a benchmark dataset, we illustrate that the novel orthogonal gate block incorporated into a neural network learning architecture appears able to distinguish between sharable and non-sharable information in a purely data-driven manner.

ACKNOWLEDGEMENT

This work was supported in part by the National Cancer Institute (U54CA209988, U2CCA233280). YHC acknowledges the OHSU Center for Spatial Systems Biomedicine, Brenden Colson Center for Pancreatic Care and Biomedical

Innovation Program Award from the Oregon Clinical & Translational Research Institute.

REFERENCES

- [1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Networks," *arXiv*, pp. 1–9, 2014.
- [2] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," *arXiv*, 2017.
- [3] J. Hoffman, E. Tzeng, T. Park, J.-y. Zhu, U. C. Berkeley, P. Isola, K. Saenko, A. A. Efros, T. Darrell, and U. C. Berkeley, "CyCADAL: Cycle-Consistent Adversarial Domain Adaptation," *arXiv*, pp. 1–15, 2017.
- [4] A. Srivastava, L. Valkov, C. Russell, M. U. Gutmann, and C. Sutton, "Veegan: Reducing mode collapse in gans using implicit variational learning," in *Advances in Neural Information Processing Systems 30* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), pp. 3308–3318, Curran Associates, Inc., 2017.
- [5] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *arXiv*, no. ML, pp. 1–14, 2013.
- [6] R. Zhang, P. Isola, and A. A. Efros, "Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction," *arXiv*, 2017.
- [7] M.-y. Liu, T. Breuel, and J. Kautz, "Unsupervised Image-to-Image Translation Networks," *NIPS*, no. Nips, 2017.
- [8] X. Huang, M.-y. Liu, S. Belongie, and J. Kautz, "Multimodal Unsupervised Image-to-Image Translation," *ECCV*, 2018.
- [9] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *ICLR*, pp. 1–15, 2015.