# A. Factored Inference

When training the BJDE in the semi-supervised regime, we introduce a factored inference procedure that reduce the number of parameters in the recognition model.

In the semi-supervised regime, the 1-layer BJDE recognition model requires approximating three posteriors: $p(z|x,y) \propto p(z)p(x,y|z)$, $p(z|x) \propto p(z)p(x|z)$, and $p(z|y) \propto p(z)p(y|z)$. The standard approach would be to assign one recognition network for each approximate posterior. This approach, however, does not take advantage of the fact that these posteriors share the same likelihood functions, i.e., $p(x,y|z) = p(x|z)p(y|z)$.

Rather than learning the three approximate posteriors independently, we propose to learn the approximate likelihood functions $\hat{\ell}(z;x) \approx p(x|z)$, $\hat{\ell}(z;y) \approx p(y|z)$ and let $\hat{\ell}(z;x,y) = \hat{\ell}(z;x)\hat{\ell}(z;y)$. Consequently, this factorization of the recognition model enables parameter sharing within the joint recognition model (which is beneficial for semi-supervised learning) and eliminates the need for constructing a neural network that takes both $x$ and $y$ as inputs. The latter property is especially useful when learning a joint model over multiple, heterogeneous data types (e.g. image, text, and audio).

In practice, we directly learn recognition networks for $q(z|x)$ and $\hat{\ell}(z;y)$ and perform factored inference as follows

$$q(z|x,y) \propto q_{\tilde{\phi}}(z|x)\hat{\ell}_{\tilde{\phi}}(z;y),\ q(z|y) \propto p(z)\hat{\ell}_{\tilde{\phi}}(z;y), \tag{14}$$

where $\tilde{\phi}$ parameterizes the recognition networks. To ensure proper normalization in Eq. (14), it is sufficient for $\hat{\ell}$ to be bounded. If the prior $p(z)$ belongs to an exponential family with sufficient statistics $T(z)$, we can parameterize $\hat{\ell}_{\tilde{\phi}}(z;y) = \exp\left(\langle T(z), \eta_{\tilde{\phi}}(y)\rangle\right)$, where $\eta_{\tilde{\phi}}(y)$ is a network such that $\eta_{\tilde{\phi}}(y) \in \{\eta | \{\langle T(z), \eta\rangle \ \ \forall z\}$ is upper bounded$\}$. Then the approximate posterior can be obtained by simple addition in the natural parameter space of the corresponding exponential family. When the prior and approximate likelihood are both Gaussians, this is exactly precision-weighted merging of the means and variances (Sønderby et al., 2016).

# B. Derivation of the Hybrid Objective

We first provide the derivation of Eq. (11). We begin with the factorization proposed in Eq. (7), which we repeat here for self-containedness,

$$\begin{aligned} p(\mathbf{X}_l, \mathbf{Y}_l, \mathbf{X}_u, \tilde{\theta}, \theta) = p(\tilde{\theta}, \theta) \\ p(\mathbf{X}_u|\tilde{\theta})p(\mathbf{X}_l|\tilde{\theta})p(\mathbf{Y}_l|\mathbf{X}_l, \theta). \end{aligned} \tag{15}$$

Since our model includes unpaired $y$, we modify Eq. (15) to include

$$\begin{aligned} p(\mathbf{X}_l, \mathbf{Y}_l, \mathbf{X}_u, \mathbf{Y}_u, \tilde{\theta}, \theta) = p(\tilde{\theta}, \theta) \\ p(\mathbf{X}_u|\tilde{\theta})p(\mathbf{Y}_u|\tilde{\theta})p(\mathbf{X}_l|\tilde{\theta})p(\mathbf{Y}_l|\mathbf{X}_l, \theta). \end{aligned} \tag{16}$$

To account for the variational parameters, we include them in the joint density as well,

$$\begin{aligned} p(\mathbf{X}_l, \mathbf{Y}_l, \mathbf{X}_u, \mathbf{Y}_u, \tilde{\theta}, \tilde{\phi}, \theta, \phi) = p(\tilde{\theta}, \tilde{\phi}, \theta, \phi) \\ p(\mathbf{X}_u|\tilde{\theta}, \tilde{\phi})p(\mathbf{Y}_u|\tilde{\theta}, \tilde{\phi}) \\ p(\mathbf{X}_l|\tilde{\theta}, \tilde{\phi})p(\mathbf{Y}_l|\mathbf{X}_l, \theta, \phi) \end{aligned} \tag{17}$$

By taking the log and replacing the necessary densities with their variational lower bound,

$$\begin{aligned} \ln p(\mathbf{X}_l, \mathbf{Y}_l, \mathbf{X}_u, \mathbf{Y}_u, \tilde{\theta}, \tilde{\phi}, \theta, \phi) \geq \ln p(\tilde{\theta}, \tilde{\phi}, \theta, \phi) + \\ \mathcal{J}_x(\tilde{\theta}, \tilde{\phi}; \mathbf{X}_u) + \mathcal{J}_y(\tilde{\theta}, \tilde{\phi}; \mathbf{Y}_u) + \\ \mathcal{J}_x(\tilde{\theta}, \tilde{\phi}; \mathbf{X}_l) + \mathcal{C}(\theta, \phi; \mathbf{X}_l, \mathbf{Y}_l), \end{aligned} \tag{18}$$

we arrive at Eq. (11). We note, however, that a more general hybrid objective Eq. (13) is achievable. To derive the general objective, we consider an alternative factorization of the joint density in Eq. (17),

$$\begin{aligned} p(\mathbf{X}_l, \mathbf{Y}_l, \mathbf{X}_u, \mathbf{Y}_u, \tilde{\theta}, \tilde{\phi}, \theta, \phi) = p(\tilde{\theta}, \tilde{\phi}, \theta, \phi, ) \\ p(\mathbf{X}_l, \mathbf{Y}_l, \mathbf{X}_u, \mathbf{Y}_u|\tilde{\theta}, \tilde{\phi}, \theta, \phi). \end{aligned} \tag{19}$$

We factorize the likelihood term such that $\mathbf{X}_u$ and $\mathbf{Y}_u$ are always explained by the joint parameters $\tilde{\theta}, \tilde{\phi}$,

$$\begin{aligned} p(\mathbf{X}_l, \mathbf{Y}_l, \mathbf{X}_u, \mathbf{Y}_u|\tilde{\theta}, \tilde{\phi}, \theta, \phi) = p(\mathbf{X}_u|\tilde{\theta}, \tilde{\phi})p(\mathbf{Y}_u|\tilde{\theta}, \tilde{\phi}) \\ p(\mathbf{X}_l, \mathbf{Y}_l|\tilde{\theta}, \tilde{\phi}, \theta, \phi). \end{aligned} \tag{20}$$

We then introduce an auxiliary variable $s = \{0, 1\}$,

$$\begin{aligned} & p(\mathbf{X}_l, \mathbf{Y}_l|\tilde{\theta}, \tilde{\phi}, \theta, \phi) \\ & \quad = \sum_s p(s)p(\mathbf{X}_l, \mathbf{Y}_l|\tilde{\theta}, \tilde{\phi}, \theta, \phi, s), \end{aligned} \tag{21}$$

where

$$p(\mathbf{X}_l, \mathbf{Y}_l|\tilde{\theta}, \tilde{\phi}, \theta, \phi, s^0) = p(\mathbf{X}_l, \mathbf{Y}_l|\tilde{\theta}, \tilde{\phi}) \tag{22}$$

$$p(\mathbf{X}_l, \mathbf{Y}_l|\tilde{\theta}, \tilde{\phi}, \theta, \phi, s^1) = p(\mathbf{X}_l|\tilde{\theta}, \tilde{\phi})p(\mathbf{Y}_l|\mathbf{X}_l, \theta, \phi). \tag{23}$$

Using Jensen's inequality, we can lower bound $\ln p(\mathbf{X}_l, \mathbf{Y}_l|\tilde{\theta}, \tilde{\phi}, \theta, \phi)$ with

$$p(s^0)\ln p(\mathbf{X}_l, \mathbf{Y}_l|\tilde{\theta}, \tilde{\phi}) + p(s^1)\ln p(\mathbf{X}_l|\tilde{\theta}, \tilde{\phi})p(\mathbf{Y}_l|\mathbf{X}_l, \theta, \phi). \tag{24}$$

By taking the log of Eq. (19), replacing all remaining densities with their variational lower bound, and setting

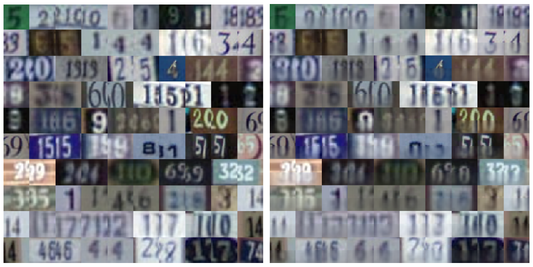
$p(s^0) = \alpha$,

$$\begin{aligned}
&\ln p(\mathbf{X}_l, \mathbf{Y}_l, \mathbf{X}_u, \mathbf{Y}_u, \tilde{\theta}, \tilde{\phi}, \theta, \phi) \\
&\quad \geq \mathcal{H}(\tilde{\theta}, \tilde{\phi}, \theta, \phi; \mathbf{X}_l, \mathbf{Y}_l, \mathbf{X}_u, \mathbf{Y}_u) \qquad (25) \\
&\quad = \ln p(\tilde{\theta}, \tilde{\phi}, \theta, \phi) + \\
&\qquad\quad \mathcal{J}_x(\tilde{\theta}, \tilde{\phi}; \mathbf{X}_u) + \mathcal{J}_y(\tilde{\theta}, \tilde{\phi}; \mathbf{Y}_u) + \\
&\qquad\quad \alpha \cdot \mathcal{J}_{xy}(\tilde{\theta}, \tilde{\phi}; \mathbf{X}_l, \mathbf{Y}_l) + \\
&\qquad\quad (1 - \alpha) \cdot \left[\mathcal{J}_x(\tilde{\theta}, \tilde{\phi}; \mathbf{X}_l) + \mathcal{C}(\theta, \phi; \mathbf{X}_l, \mathbf{Y}_l)\right], \quad (26)
\end{aligned}$$

we arrive at the general hybrid objective. Note that when $\alpha = 0$, Eq. (26) reduces to Eq. (18).

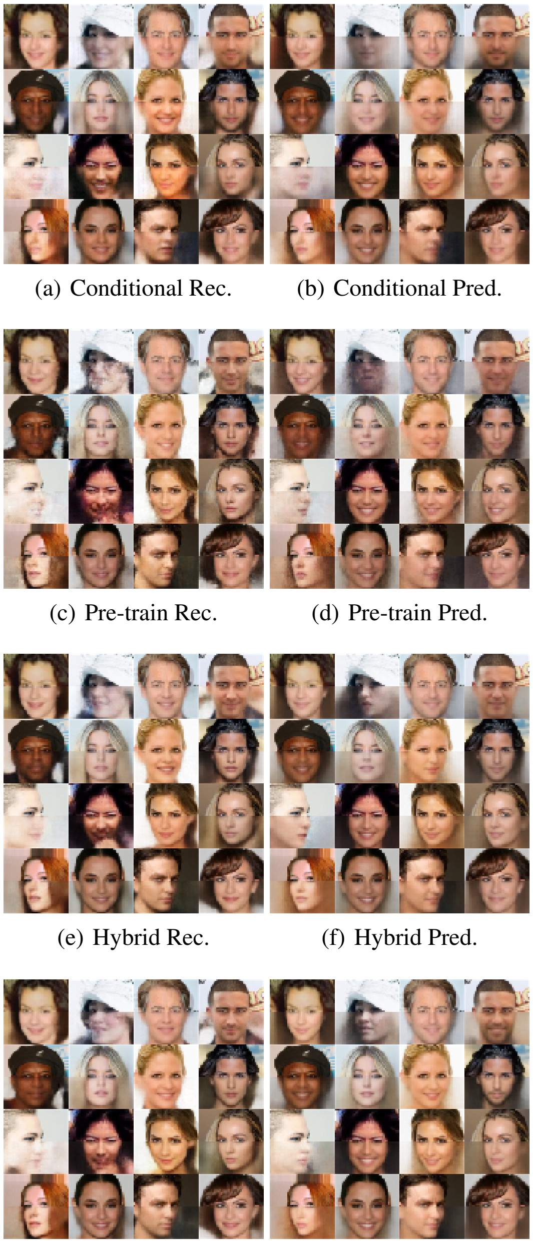
## C. Visualizations for CelebA and SVHN

We show visualizations of the hybrid BCDE predictions for CelebA and SVHN on the top-down prediction task in the $n_l = 10000$ semi-supervised regime. For each data set, we visualize both the images sampled during reconstruction as well as prediction using an approximation of the MAP estimate by greedily sampling the mode of each conditional distribution in the generative path.



(a) Hybrid Rec. (b) Hybrid Pred.

*Figure 6.* Visualization of the reconstructed and predicted bottom half of SVHN test set images when conditioned on the top half.



(a) Conditional Rec. (b) Conditional Pred.

(c) Pre-train Rec. (d) Pre-train Pred.

(e) Hybrid Rec. (f) Hybrid Pred.

(g) Hybrid Factored Rec. (h) Hybrid Factored Pred.

*Figure 7.* Visualization of the reconstructed and predicted bottom half of CelebA test set images when conditioned on the top half.