Supplementary Material for:

Bayes-Factor-VAE: Hierarchical Bayesian Deep Auto-Encoder Models for Factor Disentanglement

Minyoung Kim[1], Yuting Wang[1,2], Pritish Sahu[2], and Vladimir Pavlovic[1,2]

[1]Samsung AI Center, Cambridge, UK
[2]Dept. of Computer Science, Rutgers University, NJ, USA
mikim21@gmail.com, {yw632,ps851,vladimir}@cs.rutgers.edu, v.pavlovic@samsung.com

1. Overview

This supplement contains the following additional material:

- Sec. 2: Detailed derivations and proofs omitted from the main paper.
- Sec. 3: Details about datasets, our experimental setups, and the three disentanglement metrics used for quantitative performance assessment.
- Sec. 4: Additional experimental results.
- Sec. 5: Model architectures used in the experiments.

2. Derivations and Proofs

2.1. Invariance of KL Divergence under Nonlinear Invertible Transformation

Theorem 1. *Let* $\mathbf{x}$ *and* $\mathbf{z}$ *be two random variables related as* $\mathbf{z} = enc(\mathbf{x})$ *for a nonlinear invertible function* $enc(\cdot)$. *Then,*

$$KL(p_d(\mathbf{x})||p(\mathbf{x})) = KL(q(\mathbf{z})||p(\mathbf{z})). \tag{1}$$

Proof. We let $\text{dec}(\cdot) := \text{enc}^{-1}(\cdot)$, thus $\mathbf{x} = \text{dec}(\mathbf{z})$ and $\mathbf{z} = \text{dec}^{-1}(\mathbf{x})$. From the *theorem of density of function of random variable*, we can express $p(\mathbf{x})$ in terms of $p(\mathbf{z})$ as follows:

$$\begin{aligned} p(\mathbf{x}) &= p(\text{dec}^{-1}(\mathbf{x})) \cdot |\det \nabla_{\mathbf{x}} \text{dec}^{-1}(\mathbf{x})| && (2)\\ &= p(\mathbf{z}) \cdot \frac{1}{|\det \nabla_{\mathbf{z}} \text{dec}(\mathbf{z})|} && (3) \end{aligned}$$

In (2), ∇ is the Jacobian operator (yielding a square invertible matrix), and (3) is from the *theorem of derivative of inverse function*, namely $\nabla_{\mathbf{x}} \text{dec}^{-1}(\mathbf{x}) = (\nabla_{\mathbf{z}} \text{dec}(\mathbf{z}))^{-1}$ where the latter $(\cdot)^{-1}$ is the matrix inversion.

Similarly $p_d(\mathbf{x})$ can be written as:

$$p_d(\mathbf{x}) = q(\mathbf{z}) \cdot \frac{1}{|\det \nabla_{\mathbf{z}} \text{dec}(\mathbf{z})|} \tag{4}$$

Now the following completes the proof:

$$\begin{aligned} \text{KL}(p_d(\mathbf{x})||p(\mathbf{x})) &= \int p_d(\mathbf{x}) \cdot \log \frac{p_d(\mathbf{x})}{p(\mathbf{x})} \, \mathbf{dx} && (5)\\ &= \int q(\mathbf{z}) \cdot \frac{1}{|\det \nabla_{\mathbf{z}} \text{dec}(\mathbf{z})|} \cdot \log \frac{q(\mathbf{z})}{p(\mathbf{z})} \, \mathbf{dx} && (6)\\ &= \int q(\mathbf{z}) \cdot \log \frac{q(\mathbf{z})}{p(\mathbf{z})} \, \mathbf{dz} && (7)\\ &= \text{KL}(q(\mathbf{z})||p(\mathbf{z})), && (8) \end{aligned}$$

where in the third equality, we used $\mathbf{dx} = |\det \nabla_{\mathbf{z}} \text{dec}(\mathbf{z})| \mathbf{dz}$. □

2.2. Decomposition of $\mathrm{KL}_{\mathbf{z}}$

Theorem 2. *Let $\mathbf{R}$ and $\mathbf{N}$ be the sets of relevant and nuisance variables, respectively, in the latent space $\mathbf{z} \in \mathcal{Z} \subseteq \Re^d$ such that $\mathbf{R} \cup \mathbf{N} = \{1, \ldots, d\}$ and $\mathbf{R} \cap \mathbf{N} = \emptyset$. The KL divergence in the latent space $\mathcal{Z}$ can then be decomposed as*

$$KL_{\mathbf{z}} := KL\left(q(\mathbf{z})||\prod_{j=1}^{d} p(z_j)\right) = TC + \sum_{j\in\mathbf{R}} KL(q(z_j)||p(z_j)) + \sum_{j\in\mathbf{N}} KL(q(z_j)||p(z_j), \tag{9}$$

where

$$TC := KL\left(q(\mathbf{z})||\prod_{j=1}^{d} q(z_j)\right). \tag{10}$$

Proof.

$$\begin{aligned}
\mathrm{KL}_{\mathbf{z}} &:= \mathrm{KL}\left(q(\mathbf{z})||\prod_{j=1}^{d} p(z_j)\right) && (11)\\
&= \mathbb{E}_{q(\mathbf{z})}\left[\log \frac{q(\mathbf{z})}{\prod_{j=1}^{d} p(z_j)}\right] = \mathbb{E}_{q(\mathbf{z})}\left[\log \frac{q(\mathbf{z})\prod_{j=1}^{d} q(z_j)}{\prod_{j=1}^{d} p(z_j)\prod_{j=1}^{d} q(z_j)}\right] && (12)\\
&= \mathrm{KL}\left(q(\mathbf{z})||\prod_{j=1}^{d} q(z_j)\right) + \mathbb{E}_{q(\mathbf{z})}\left[\log \frac{\prod_{j=1}^{d} q(z_j)}{\prod_{j=1}^{d} p(z_j)}\right] && (13)\\
&= \mathrm{TC} + \sum_{j=1}^{d} \mathbb{E}_{q(\mathbf{z})}\left[\log \frac{q(z_j)}{p(z_j)}\right] = \mathrm{TC} + \sum_{j=1}^{d} \mathrm{KL}\big(q(z_j)||p(z_j)\big) && (14)\\
&= \mathrm{TC} + \sum_{j\in\mathbf{R}} \mathrm{KL}(q(z_j)||p(z_j)) + \sum_{j\in\mathbf{N}} \mathrm{KL}(q(z_j)||p(z_j)) && (15)
\end{aligned}$$

□

2.3. Bound on Factorized Divergence $\mathrm{KL}(q(z_j)||p(z_j))$

Theorem 3. *Let $q(z_j)$ and $p(z_j)$ denote the approximate and prior marginal densities, respectively, of factor j in latent space $\mathcal{Z}$. Let $q(z_j|\mathbf{x})$ be the posterior (encoder driven) density of the latent factor conditioned on input $\mathbf{x}$ with density $p_d(\mathbf{x})$. Then*

$$KL(q(z_j)||p(z_j)) \leq \mathbb{E}_{p_d(\mathbf{x})}[KL(q(z_j|\mathbf{x})||p(z_j))]. \tag{16}$$

Proof.

$$\begin{aligned}
&\mathbb{E}_{p_d(\mathbf{x})}[\mathrm{KL}(q(z_j|\mathbf{x})||p(z_j))] - \mathrm{KL}(q(z_j)||p(z_j)) && (17)\\
&= \mathbb{E}_{p_d(\mathbf{x})}\big[\mathbb{E}_{q(z_j|\mathbf{x})}[\log q(z_j|\mathbf{x}) - \log p(z_j)]\big] - \mathbb{E}_{p_d(\mathbf{x})}\big[\mathbb{E}_{q(z_j|\mathbf{x})}[\log q(z_j) - \log p(z_j)]\big] && (18)\\
&= \mathbb{E}_{p_d(\mathbf{x})}\big[\mathbb{E}_{q(z_j|\mathbf{x})}[\log q(z_j|\mathbf{x}) - \log q(z_j)]\big] && (19)\\
&= \mathbb{E}_{p_d(\mathbf{x})}[\mathrm{KL}(q(z_j|\mathbf{x})||q(z_j))] \;\geq\; 0. && (20)
\end{aligned}$$

□

2.4. Justification of the Prior Choice in BF-VAE-0

Recall that in BF-VAE-0, we adopted a 0-mean, free-precision Gaussian as a prior model $p(z_j)$, namely $p(z_j) = \mathcal{N}(z_j; 0, \alpha_j^{-1})$, and asserted that it is as flexible as a free-mean, free-precision Gaussian $\mathcal{N}(z_j; \mu_j, \alpha_j^{-1})$, in terms of matching $p(z_j)$ and $q(z_j)$, i.e., minimizing $\mathbb{E}_{p_d(\mathbf{x})}[\mathrm{KL}(q(z_j|\mathbf{x})||p(z_j))]$, the upper bound of $\mathrm{KL}(q(z_j)||p(z_j))$.

We prove this assertion by showing that the following minimization problem wrt $p(z_j) = \mathcal{N}(z_j; \mu_j, \alpha_j^{-1})$ attains the optimum $\mu_j^* = 0$ with arbitrary α_j^* (i.e., either $\alpha_j^* > 1$ or $\alpha_j^* < 1$ or even $\alpha_j^* = 1$).

$$\min_{\mu_j, \alpha_j > 0} \mathbb{E}_{p_d(\mathbf{x})}[\mathrm{KL}(q(z_j|\mathbf{x})||\mathcal{N}(z_j; \mu_j, \alpha_j^{-1}))], \tag{21}$$

where $q(z_j|\mathbf{x}) = \mathcal{N}(z_j; m_j(\mathbf{x}), s_j(\mathbf{x})^2)$.

Note that the objective of (21) is essentially an *average of KL divergences over different* $\mathbf{x}$*'s*, and we can write (21) as (dropping the subscript j for simplicity):

$$\min_{\mu,\alpha>0} \mathbb{E}_{p(m,s^2)}[\mathrm{KL}(\mathcal{N}(m,s^2)||\mathcal{N}(\mu,\alpha^{-1}))], \tag{22}$$

for some distribution $p(m, s^2)$. Since the KL term admits a closed-form, (22) is equivalent to (up to constant):

$$\min_{\mu,\alpha>0} \mathbb{E}_{p(m,s^2)}\Big[\alpha\big((m-\mu)^2+s^2\big)\Big] - \log\alpha, \tag{23}$$

Letting $\overline{s}^2 = \mathbb{E}[s^2]$ leads to:

$$\min_{\mu,\alpha>0} \alpha\mathbb{E}_{p(m)}\big[(m-\mu)^2\big] + \alpha\overline{s}^2 - \log\alpha. \tag{24}$$

Without loss of generality, we can assume that $p(m)$ is symmetric (i.e., $p(m) = p(-m)$). Then (24) can be written as (using $\mathbb{E}[m] = 0$):

$$\min_{\mu,\alpha>0} \alpha\Big(\overline{m}^2 + \mu^2 + \overline{s}^2\Big) - \log\alpha, \tag{25}$$

where we let $\overline{m}^2 = \mathbb{E}[m^2]$.

From (25), it is obvious that the optimal $\mu^* = 0$ and $\alpha^* = \big(\overline{m}^2 + \overline{s}^2\big)^{-1}$, where the latter can take arbitrary value depending on $\overline{m}$ and $\overline{s}$. This justifies our choice of 0-mean and free-precision Gaussians as latent prior distributions.

2.5. Bound on Average Marginal Data Log-Likelihood in BF-VAE-1

Theorem 4. *Let* $\{\mathbf{x}^n\}_{n=1}^N$ *be the set of iid ambient observations. Then, the negative log likelihood of these observations can be upper bounded by the BF-VAE-1 model as*

$$-\log p\big(\{\mathbf{x}^n\}_{n=1}^N\big) \quad\leq\quad \mathit{Rec}(\theta,\nu) \quad+\quad \frac{1}{N}\mathit{KL}(q(\boldsymbol{\alpha})||p(\boldsymbol{\alpha})) \quad+\quad \mathbb{E}_{p_d(\mathbf{x})}\mathbb{E}_{q(\boldsymbol{\alpha})}\big[\mathit{KL}(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\boldsymbol{\alpha}))\big]. \tag{26}$$

Proof. We begin with the KL divergence between the true posterior distribution and our variational approximation.

$$\mathrm{KL}\left(q(\boldsymbol{\alpha})\prod_{n=1}^N q(\mathbf{z}^n|\mathbf{x}^n)\;\middle\|\; p\big(\boldsymbol{\alpha},\{\mathbf{z}^n\}_{n=1}^N|\{\mathbf{x}^n\}_{n=1}^N\big)\right) \tag{27}$$

$$= \mathbb{E}_{q(\boldsymbol{\alpha})\prod_n q(\mathbf{z}^n|\mathbf{x}^n)}\left[\log\frac{q(\boldsymbol{\alpha})\prod_n q(\mathbf{z}^n|\mathbf{x}^n)}{p\big(\boldsymbol{\alpha},\{\mathbf{z}^n\}_{n=1}^N|\{\mathbf{x}^n\}_{n=1}^N\big)}\right] \tag{28}$$

$$= \mathbb{E}_{q(\boldsymbol{\alpha})\prod_n q(\mathbf{z}^n|\mathbf{x}^n)}\left[\log\frac{q(\boldsymbol{\alpha})\prod_n q(\mathbf{z}^n|\mathbf{x}^n)}{p\big(\boldsymbol{\alpha},\{\mathbf{z}^n\}_{n=1}^N,\{\mathbf{x}^n\}_{n=1}^N\big)/p\big(\{\mathbf{x}^n\}_{n=1}^N\big)}\right] \tag{29}$$

$$= \mathbb{E}_{q(\boldsymbol{\alpha})\prod_n q(\mathbf{z}^n|\mathbf{x}^n)}\left[\log\frac{p\big(\{\mathbf{x}^n\}_{n=1}^N\big)q(\boldsymbol{\alpha})\prod_n q(\mathbf{z}^n|\mathbf{x}^n)}{p(\boldsymbol{\alpha})\prod_n p(\mathbf{z}^n|\boldsymbol{\alpha})p(\mathbf{x}^n|\mathbf{z}^n)}\right] \tag{30}$$

$$= \mathbb{E}_{q(\boldsymbol{\alpha})\prod_n q(\mathbf{z}^n|\mathbf{x}^n)}\left[\log p\big(\{\mathbf{x}^n\}_{n=1}^N\big) + \log\frac{q(\boldsymbol{\alpha})}{p(\boldsymbol{\alpha})} + \log\frac{\prod_n q(\mathbf{z}^n|\mathbf{x}^n)}{\prod_n p(\mathbf{z}^n|\boldsymbol{\alpha})} - \log p(\mathbf{x}^n|\mathbf{z}^n)\right] \tag{31}$$

$$= \log p\big(\{\mathbf{x}^n\}_{n=1}^N\big) + \mathrm{KL}(q(\boldsymbol{\alpha})||p(\boldsymbol{\alpha})) + \mathbb{E}_{q(\boldsymbol{\alpha})}\left[\sum_{n=1}^N\Big\{\mathrm{KL}(q(\mathbf{z}^n|\mathbf{x}^n)||p(\mathbf{z}^n|\boldsymbol{\alpha})) - \mathbb{E}_{q(\mathbf{z}^n|\mathbf{x}^n)}\big[\log p(\mathbf{x}^n|\mathbf{z}^n)\big]\Big\}\right] \tag{32}$$

Using the fact that KL divergence (27) is non-negative, we have the following inequality:

$$-\log p\big(\{\mathbf{x}^n\}_{n=1}^N\big) \;\leq\; \mathrm{KL}(q(\boldsymbol{\alpha})||p(\boldsymbol{\alpha})) + \mathbb{E}_{q(\boldsymbol{\alpha})}\left[\sum_{n=1}^N\Big\{\mathrm{KL}(q(\mathbf{z}^n|\mathbf{x}^n)||p(\mathbf{z}^n|\boldsymbol{\alpha})) - \mathbb{E}_{q(\mathbf{z}^n|\mathbf{x}^n)}\big[\log p(\mathbf{x}^n|\mathbf{z}^n)\big]\Big\}\right] \tag{33}$$

Dividing both sides by N leads to our upper bound $\mathcal{U}_1$ in (26).

□

2.6. Explicit Parametric Forms of KL Divergences for BF-VAE-1 Likelihood Bound

Theorem 5. *Let $q(\boldsymbol{\alpha})$ and $p(\boldsymbol{\alpha})$ be parameterized as Gamma densities:*

$$q(\boldsymbol{\alpha}) = \prod_j \mathcal{G}(\alpha_j; \hat{a}_j, \hat{b}_j), \tag{34}$$

$$p(\boldsymbol{\alpha}) = \prod_j \mathcal{G}(\alpha_j; a_j, b_j), \tag{35}$$

where $\mathcal{G}(y; a, b) = \frac{b^a}{\Gamma(a)} y^{a-1} e^{-by}$, and $\Gamma(\cdot)$ is the gamma function. Then,

$$KL(q(\boldsymbol{\alpha})||p(\boldsymbol{\alpha})) = \sum_{j=1}^{d} \left((\hat{a}_j - a_j)\psi(\hat{a}_j) + a_j \log \frac{\hat{b}_j}{b_j} + \log \frac{\Gamma(a_j)}{\Gamma(\hat{a}_j)} + \frac{\hat{a}_j(b_j - \hat{b}_j)}{\hat{b}_j} \right). \tag{36}$$

Proof. We will use the following well-known fact of the expectations of Gamma and log-Gamma random variables. For $y \sim \mathcal{G}(y; a, b)$,

$$\mathbb{E}[y] = \frac{a}{b}, \quad \mathbb{E}[\log y] = \psi(a) - \log b, \tag{37}$$

where $\psi(\cdot)$ is the digamma function (i.e., $\psi(y) = \frac{\Gamma'(y)}{\Gamma(y)}$).

First we tackle $\mathrm{KL}(q(\boldsymbol{\alpha})||p(\boldsymbol{\alpha}))$, the second term in $\mathcal{U}_1$ in (26). It essentially involves in the KL divergence between two univariate Gamma distributions, which can be derived as:

$$\mathrm{KL}(\mathcal{G}(y; \hat{a}, \hat{b})||\mathcal{G}(y; a, b)) = \mathbb{E}_{\mathcal{G}(y;\hat{a},\hat{b})}\left[\log \frac{\mathcal{G}(y; \hat{a}, \hat{b})}{\mathcal{G}(y; a, b)} \right] \tag{38}$$

$$= \mathbb{E}_{\mathcal{G}(y;\hat{a},\hat{b})}\left[\Big(\hat{a} \log \hat{b} - \log \Gamma(\hat{a}) + (\hat{a} - 1) \log y - \hat{b} y \Big) - \Big(a \log b - \log \Gamma(a) + (a - 1) \log y - by \Big) \right] \tag{39}$$

$$= \hat{a} \log \hat{b} - a \log b - \log \Gamma(\hat{a}) + \log \Gamma(a) + (\hat{a} - a)(\psi(\hat{a}) - \log \hat{b}) + (b - \hat{b}) \frac{\hat{a}}{\hat{b}} \tag{40}$$

$$= (\hat{a} - a)\psi(\hat{a}) + a \log \frac{\hat{b}}{b} + \log \frac{\Gamma(a)}{\Gamma(\hat{a})} + \frac{\hat{a}(b - \hat{b})}{\hat{b}} \tag{41}$$

Since $\mathrm{KL}(q(\boldsymbol{\alpha})||p(\boldsymbol{\alpha})) = \sum_{j=1}^{d} \mathrm{KL}\big(\mathcal{G}(\alpha_j; \hat{a}_j, \hat{b}_j)||\mathcal{G}(\alpha_j; a_j, b_j)\big)$, we directly arrive at (36). □

Theorem 6. *Let $q(\boldsymbol{\alpha})$ and $p(\boldsymbol{\alpha})$ be parameterized as Gamma densities in (34) and (35). Then,*

$$\mathbb{E}_{q(\boldsymbol{\alpha})}\Big[KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\boldsymbol{\alpha})) \Big] = \frac{1}{2} \sum_{j=1}^{d} \left(-\psi(\hat{a}_j) + \log \frac{\hat{b}_j}{s_j(\mathbf{x})^2} + \frac{\hat{a}_j \big(m_j(\mathbf{x})^2 + s_j(\mathbf{x})^2 \big)}{\hat{b}_j} - 1 \right). \tag{42}$$

Proof. Here we derive $\mathbb{E}_{p_d(\mathbf{x})}\mathbb{E}_{q(\boldsymbol{\alpha})}\big[\mathrm{KL}(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\boldsymbol{\alpha}))\big]$, the third term in $\mathcal{U}_1$ in (26). The outer expectation over $p_d(\mathbf{x})$ is computed by Monte Carlo estimation with minibatches, while for a given $\mathbf{x}$, the inner expectation of the Gaussian KL over the Gamma also admits a closed form as follows:

$$\mathbb{E}_{q(\boldsymbol{\alpha})}\Big[\mathrm{KL}(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\boldsymbol{\alpha})) \Big] = \sum_{j=1}^{d} \mathbb{E}_{q(\alpha_j)}\Big[\mathrm{KL}\big(\mathcal{N}(m_j(\mathbf{x}), s_j(\mathbf{x})^2) \,||\, \mathcal{N}(0, \alpha_j^{-1}) \big) \Big] \tag{43}$$

$$= \sum_{j=1}^{d} \mathbb{E}_{q(\alpha_j)}\left[\frac{1}{2} \Big(\alpha_j \big(m_j(\mathbf{x})^2 + s_j(\mathbf{x})^2 \big) - 1 - \log s_j(\mathbf{x})^2 - \log \alpha_j \Big) \right] \tag{44}$$

$$= \frac{1}{2} \sum_{j=1}^{d} \left(-\psi(\hat{a}_j) + \log \frac{\hat{b}_j}{s_j(\mathbf{x})^2} + \frac{\hat{a}_j \big(m_j(\mathbf{x})^2 + s_j(\mathbf{x})^2 \big)}{\hat{b}_j} - 1 \right). \tag{45}$$

□

2.7. Decomposition of Approximate Marginal $q(\boldsymbol{z})$ in BF-VAE-2

Theorem 7. *Let $q(\boldsymbol{z})$ be the marginal density of given by encoder $q(\boldsymbol{z}|\boldsymbol{x})$, where the ambient data is distributed according to $p_d(\boldsymbol{x})$. Then,*

$$q(\mathbf{z}) = q(\mathbf{z_R}) \cdot \prod_{j \in \mathbf{N}} q(z_j), \tag{46}$$

where $\mathbf{R}$ and $\mathbf{N}$ are the sets of relevant and nuisance variables, respectively.

Proof. From the definition of the marginal $q(\boldsymbol{z})$ we get:

$$\begin{aligned}
q(\mathbf{z}) &= \int q(\mathbf{z}|\mathbf{x}) p_d(\mathbf{x}) d\mathbf{x} = \int \prod_{j=1}^{d} q(z_j|\mathbf{x}) p_d(\mathbf{x}) d\mathbf{x} && (47)\\
&= \int \prod_{j \in \mathbf{R}} q(z_j|\mathbf{x}) \prod_{j \in \mathbf{N}} q(z_j|\mathbf{x}) p_d(\mathbf{x}) d\mathbf{x} && (48)\\
&= \int \prod_{j \in \mathbf{R}} q(z_j|\mathbf{x}) \prod_{j \in \mathbf{N}} q(z_j) p_d(\mathbf{x}) d\mathbf{x} && (49)\\
&= \left(\int \prod_{j \in \mathbf{R}} q(z_j|\mathbf{x}) p_d(\mathbf{x}) d\mathbf{x} \right) \cdot \prod_{j \in \mathbf{N}} q(z_j) && (50)\\
&= q(\mathbf{z_R}) \cdot \prod_{j \in \mathbf{N}} q(z_j). && (51)
\end{aligned}$$

□

3. Details of Datasets, Experimental Setups, and Disentanglement Metrics

3.1. Datasets

In this work, we consider the following datasets:

- **3D-Face** [7]. This dataset contains synthetic face images rendered from a 3D face mode, with four factors of variation (number of modes): subject ID (50), azimuth (21), elevation (11), and lighting along the horizon (11).
- **Sprites** [6]. The dataset consists of $737,280$ binary images of three shapes (oval, square, and heart), undergoing variations in four geometric factors: scale (6 variation modes), rotation (40), and X, Y translation (32 modes each), resulting in five factors total.
- **Color-Extended Sprites (`C-Spr`)** [5]. The dataset extends `Sprites` by filling the sprites with some randomly chosen colors in the images. Since the color values are not fully explored for every configuration of the factors, we regard it as simple noise, not a factor, as is also intended by the dataset creators.
- **Oval Shape Subset (`O-Spr`)** . From the original dataset, we select the oval shape images, yielding four independent factors of variation across $245,760$ images.
- **Teapots** [1]. The dataset contains 200,000 images of a teapot across five different sources of variation: azimuth, elevation, and the color of the teapot object. Since the factor labels provided in the dataset are real and uniquely valued, to evaluate metrics I and II, we discretized them into ten bins.
- **Celeb-A** [4]. This popular dataset captures variations in 40 binary attributes including azimuth, hair color, bang, gender, mustache, brightness, etc. The is a challenging dataset because the labels (attributes) are biased[1] and noisy.

[1] Celeb-A attributes are not fully disentangled. For instance, the attributes *male* and *bald* (or *mustache*) are highly correlated, and the dataset does not fully/uniformly cover the entire joint factor space (e.g., $(male = 0, bald = 1)$ is rare).

3.2. Experimental Setup

The latent dimension d for all models is set to 10 for all datasets except $d = 20$ for `Celeb-A`. For the trade-off parameters, we have used the optimal values reported in the publications: i) β-VAE; $\beta = 8.0$ for `3D-Face`, $\beta = 4.0$ for `Sprites`, `C-Spr`, and `O-Spr`, and $\beta = 6.0$ for `Teapots`, ii) Factor-VAE: $\gamma = 10.0$ for `3D-Face` and `Teapots`, and $\gamma = 35.0$ for `Sprites`, `C-Spr`, and `O-Spr`. For our BF-VAE-0 and BF-VAE-2 models, the sparseness trade-off parameters (η's) are obtained through grid search to yield the best performance. They are: i) BF-VAE-0: $\eta = 1.0$ (`3D-Face`), $\eta = 0.05$ (`Sprites`), $\eta = 0.01$ (`C-Spr`), $\eta = 0.05$ (`O-Spr`), and $\eta = 0.03$ (`Teapots`), ii) BF-VAE-2: $\eta_S = \eta_H = 1.0$ (`3D-Face`), $\eta_S = 0.05, \eta_H = 0.1$ (`Sprites`), $\eta_S = 0.01, \eta_H = 0.1$ (`C-Spr`), $\eta_S = \eta_H = 0.1$ (`O-Spr` and `Teapots`).

3.3. Disentanglement Metrics

Metric I [2] - One factor fixed. The goal of this metric is to assess the variability of the discovered latent factors as a function of the true factor variance. Let $\mathbf{v}$ be the vector of ground-truth factors for $\mathbf{x}$. For each factor index j, a set of L samples $\mathbf{v}^{(i)} = (v_j, v_{-j}^{(i)})$, $i = 1, \ldots, L$, is collected, corresponding to clamping of factor j (no variance) and free variation of the remaining factors. The existence of latent factor(s) z_u with similar, vanishing variation would then indicate the discovery of known factor j. In practice, this is accomplished by evaluating the encoder's outputs, $\mathbf{z}^{(i)} \sim q(\mathbf{z}|\mathbf{x}^{(i)})$ for images $\mathbf{x}^{(i)}$ corresponding to samples $\mathbf{v}^{(i)}$. The (sample) variance $\mathbb{V}(\mathbf{z})$ is used to find the index u of the factor with the smallest variance,

$$u := \arg \min_{1 \le j \le d} \mathbb{V}(z_j). \tag{52}$$

u then serves as the covariate for predicting the true factor index j: the metric is defined to be the accuracy of a simple classifier that predicts j from u, among all $j = 1, \ldots, d$. If a model achieves strong disentanglement, we can expect $u = j$ (up to a permutation), making the classification easy. In particular, since $u, j \in \{1, \ldots, d\}$, the data pairs $\{(u, j)\}$ can be represented as a simple contingency table, in which a majority vote classifier is used for prediction. We use $L = 100$ samples to form each pair (u, j), and collect 800 pairs to compute the accuracies of the majority vote classifiers. Since the metric is based on random samples, we repeat the evaluation ten times randomly to report the means and standard deviations.

Metric II - One factor varied. Following the notion of disentanglement, another reasonable approach was proposed in [3] to collect samples with *only one factor varied*, instead of one factor fixed as in Metric I. That is, we collect images with $\mathbf{v}^{(i)} = (v_j^{(i)}, v_{-j})$ for $i = 1, \ldots, L$. (52) is then modified to $\arg\max$, and we can use the same majority vote classification to report the accuracy. The evaluation results in [3] demonstrate that this metric shows higher agreement with qualitative assessment of disentanglement than Metric I. However, note that to compute Metric II the dataset needs to contain *dense* joint variations in all true factors, typically a reasonable assumption for large, diverse datasets.

Metric III [1] proposed three metrics: 1) *Disentanglement*, 2) *Completeness*, and 3) *Informativeness*. These scores are regression-prediction based, using the latent vector $\mathbf{z}$ as the covariate for individual ground-truth factors v_j. Specifically, D measures the degree of dedication of each latent variable z_k in predicting v_j against others v_{-j} (the higher, the better), C captures the degree of exclusive contribution of z_k in predicting v_j against others z_{-k} (the higher, the better), and I measures the prediction error (the smaller, the better). For the regressors, both LASSO and Random Forests are used.

4. Additional Experimental Results

As we stated in the main paper, we place additional experimental results in this section: They are: i) Qualitative results (latent space traversal) on all datasets, and ii) Comparison with high-capacity prior models, the mixtures of Gaussians (Sec. 4.2).

4.1. Qualitative Results: Latent Space Traversal and Learned Relevance Indicators

As we did in the main paper, we depict images synthesized by traversing a single latent variable at a time and fixing the rest, while showing the accuracy of variable relevance indicator. Recall that our models have implicit/explicit indicators that point to relevant and nuisance variables. Specifically, i) BF-VAE-0 (learned α_j): j relevant if α_j is away from 1, while j is nuisance if $\alpha_j \approx 1$, ii) BF-VAE-1 (DOF of the corrected prior $\overline{p}(z_j)$, equal to $2\hat{a}_j$): j is relevant if $\hat{a}_j$ is small (distant from Gaussian), and vice versa, iii) BF-VAE-2 (learned relevance indicator variable r_j): j is relevant if r_j is large, and vice versa.

As the relevance indicators have different interpretation across the three models, we define *proxy indicators* $\hat{\mathbf{r}}$ for BF-VAE-0 and BF-VAE-1, which have consistent interpretation as that of BF-VAE-2, i.e., high/low r_j or $\hat{r}_j$ indicates relevance/nuisance of the variable z_j. Specifically, we define: $\hat{r}_j = |1 - \alpha_j^{-1}|$ for BF-VAE-0. For BF-VAE-1, we normalize the DOF ($2\hat{a}_j$) values to $[0, 1]$-scale by a linear transformation.

The results are shown, with the discussions in the captions, for:

- `3D-Face` in Fig. 1, Fig. 2, Fig. 3
- `Sprites` in Fig. 4, Fig. 5, Fig. 6
- `C-Spr` in Fig. 7, Fig. 8, Fig. 9
- `O-Spr` in Fig. 10, Fig. 11, Fig. 12
- `Teapots` in Fig. 13, Fig. 14, Fig. 15
- `Celeb-A` in Fig. 16, Fig. 17, Fig. 18.

Also, the enlarged version of the latent traversal images for BF-VAE-2 in the main paper is shown in Fig. 19. As detailed in the caption, adopting large η leads only strong factors to be detected, while having small η allows many weak factors identified.

4.2. Comparison with High Capacity Prior Model (Mixture of Gaussians)

The results on all labeled datasets are shown in Fig. 20. As evident, the MoG overfits as K increases for all datasets, implying that overly flexible prior model can be detrimental to the disentanglement performance. Equally significantly, MoG underperforms across the entire range of K, compared to our BF-VAE models.

5. Model Architectures and Optimization

In this section, we provide the detailed specifications of the structures of different deep VAE models used in our experiments as well as the optimization strategies.

5.1. Model Architectures

We adopt the model architectures and optimization parameters similar to those in [2]. The encoders consist of 5-layer conv-nets followed by two fully connected layers, and the decoders are 4-layer deconv-nets after two fully connected layers. We apply (4×4) filters for the convolution and the transposed convolution (deconv) in both models. For the adversarial discriminator D used for optimizing the TC loss, we use a 6-layer MLP model with 1000 hidden units per layer and the leaky ReLU nonlinearity. Specifically,

- Model architecture for 3D Faces and Sprites(Sprites, C-Spr, and O-Spr) is specified in Tab. 1.
- Model architecture for Teapots and Celeb-A is listed in Tab. 2.

Tab. 1. Encoder and Decoder architecture for `3D-Faces` and `Sprites` (`Sprites`, `C-Spr`, and `O-Spr`) dataset.

ENCODER	DECODER
INPUT 64×64 GREYSCALE IMAGE (3D-FACES) INPUT 64×64 BINARY IMAGE (SPRITES)	INPUT $\in \mathbb{R}^{10}$
4×4 CONV. 32 RELU. STRIDE 2	FC. 256 RELU (3D-FACES) FC. 128 RELU(SPRITES)
4×4 CONV. 32 RELU. STRIDE 2	FC. $4 \times 4 \times 64$ RELU.
4×4 CONV. 64 RELU. STRIDE 2	4×4 UPCONV. 64 RELU. STRIDE 2
4×4 CONV. 64 RELU. STRIDE 2	4×4 UPCONV. 32 RELU. STRIDE 2
FC. 256 RELU. FC. 2×10 (3D-FACES) FC. 128 RELU. FC. 2×10 (SPRITES)	4×4 UPCONV. 32 RELU. STRIDE 2
	4×4 UPCONV. 1. STRIDE 2

5.2. Optimization

The optimization parameters are chosen similarly as those in [2]. We use Adam with the batch size 64. We run 3×10^5 batch iterations with learning rates either 10^{-4} or 10^{-5}. In RF-VAE [3] and our BF-VAE-2, the relevance vector $\mathbf{r}$ is initialized as all-0.5 vector.

Figure 1. Latent space traversal in BF-VAE-0 on the `3D-Face` dataset. The learned prior variances $\boldsymbol{\alpha}^{-1}$ (the value 1.0 depicted as the red dotted line) and the (proxy) relevance indicator $\hat{r}_j = |1 - \alpha_j^{-1}|$ are shown at the bottom. The four visually evident dimensions of variability (z_4, z_5, z_8, z_9) are highlighted within colored boxes, where each exactly matches one of the four ground-truth factors (z_4 = azimuth, z_5 = lighting, z_8 = elevation, and z_9 = subject ID). The learned α_j for all these four dims are away from 1, and the proxy indicator $\hat{r}_j$ for these dimensions are high (away from 0). Note that z_6 is emphasized as relevant by our model, but the traversal does not seem to depict significant changes. Upon further investigation, it is revealed that this factor correlates with the subject ID, but less so than z_9. The chosen traversal range suppressed visual variability of z_6, but not z_9, suggesting the need to carefully determine the range bounds.

Tab. 2. Encoder and Decoder architecture for `Teapots` and `Celeb-A` dataset.

ENCODER	DECODER
INPUT $64 \times 64 \times 3$ RGB IMAGE	INPUT $\in \mathbb{R}^{10}$
4×4 CONV. 32 RELU. STRIDE 2	FC. 256 RELU.
4×4 CONV. 32 RELU. STRIDE 2	FC. $2 \times 2 \times 256$ RELU.
4×4 CONV. 64 RELU. STRIDE 2	4×4 UPCONV. 64 RELU. STRIDE 2
4×4 CONV. 64 RELU. STRIDE 2	4×4 UPCONV. 64 RELU. STRIDE 2
4×4 CONV. 256 RELU. STRIDE 2	4×4 UPCONV. 32 RELU. STRIDE 2
FC. 256 RELU. FC. 2×10.	4×4 UPCONV. 32 RELU. STRIDE 2
	4×4 UPCONV. 3. STRIDE 2

References

[1] C. Eastwood and C. K. I. Williams. A framework for the quantitative evaluation of disentangled representations, 2018. In Proceedings of the Second International Conference on Learning Representations, ICLR. 5, 6, 21

[2] H. Kim and A. Mnih. Disentangling by factorising. International Conference on Machine Learning, 2018. 6, 7

[3] M. Kim, Y. Wang, P. Sahu, and V. Pavlovic. Relevance Factor VAE: Learning and Identifying Disentangled Factors, 2019. arXiv:1902.01568. 6, 7

[4] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 5

[5] F. Locatello, S. Bauer, M. Lucic, G. Rätsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations, 2018. arXiv:1811.12359. 5

[6] L. Matthey, I. Higgins, D. Hassabis, and A. Lerchner. dSprites: Disentanglement testing Sprites dataset, 2017. 5

[7] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3D Face Model for Pose and Illumination Invariant Face Recognition, 2009. Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance. 5

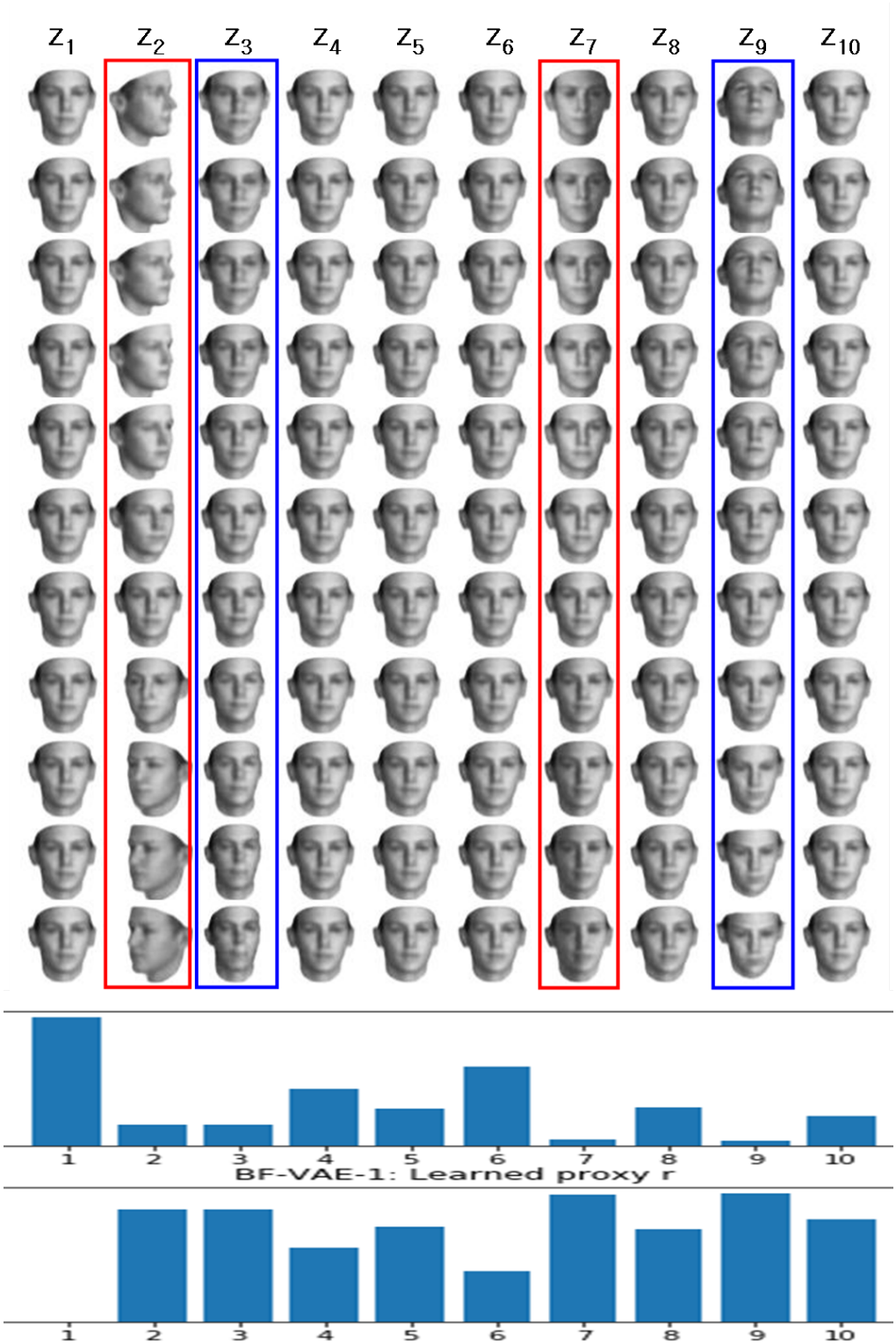

Figure 2. Latent space traversal in BF-VAE-1 on the `3D-Face` dataset. The DOF ($2\hat{a}_j$) of the corrected prior $\overline{p}(z_j)$ and the (proxy) relevance indicator $\hat{r}_j$ (obtained by linearly transforming/interpolating the DOF into $[0, 1]$) are shown at the bottom. The four recovered, highlighted, dimensions match the ground-truth factors, and their $\overline{p}(z_j)$'s also have relatively small DOFs, as expected, while the proxy indicator $\hat{r}_j$ for these dimensions are high (away from 0).

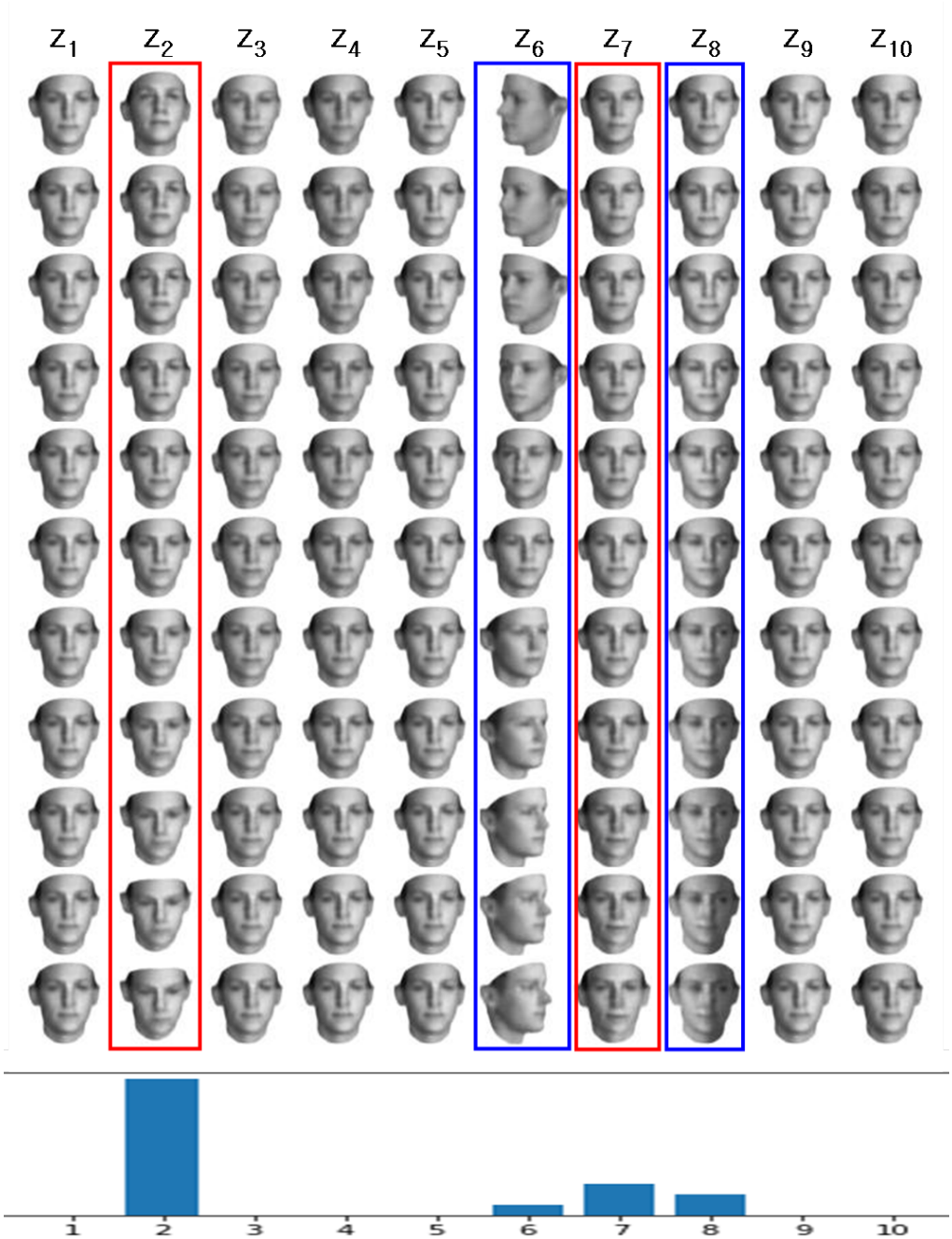

Figure 3. Latent space traversal in BF-VAE-2 on the `3D-Face` dataset. The learned relevance vector $\mathbf{r}$ is shown at the bottom. Again the four factors are nearly correctly identified, corresponding to the high values in the indicator variables r_j's.

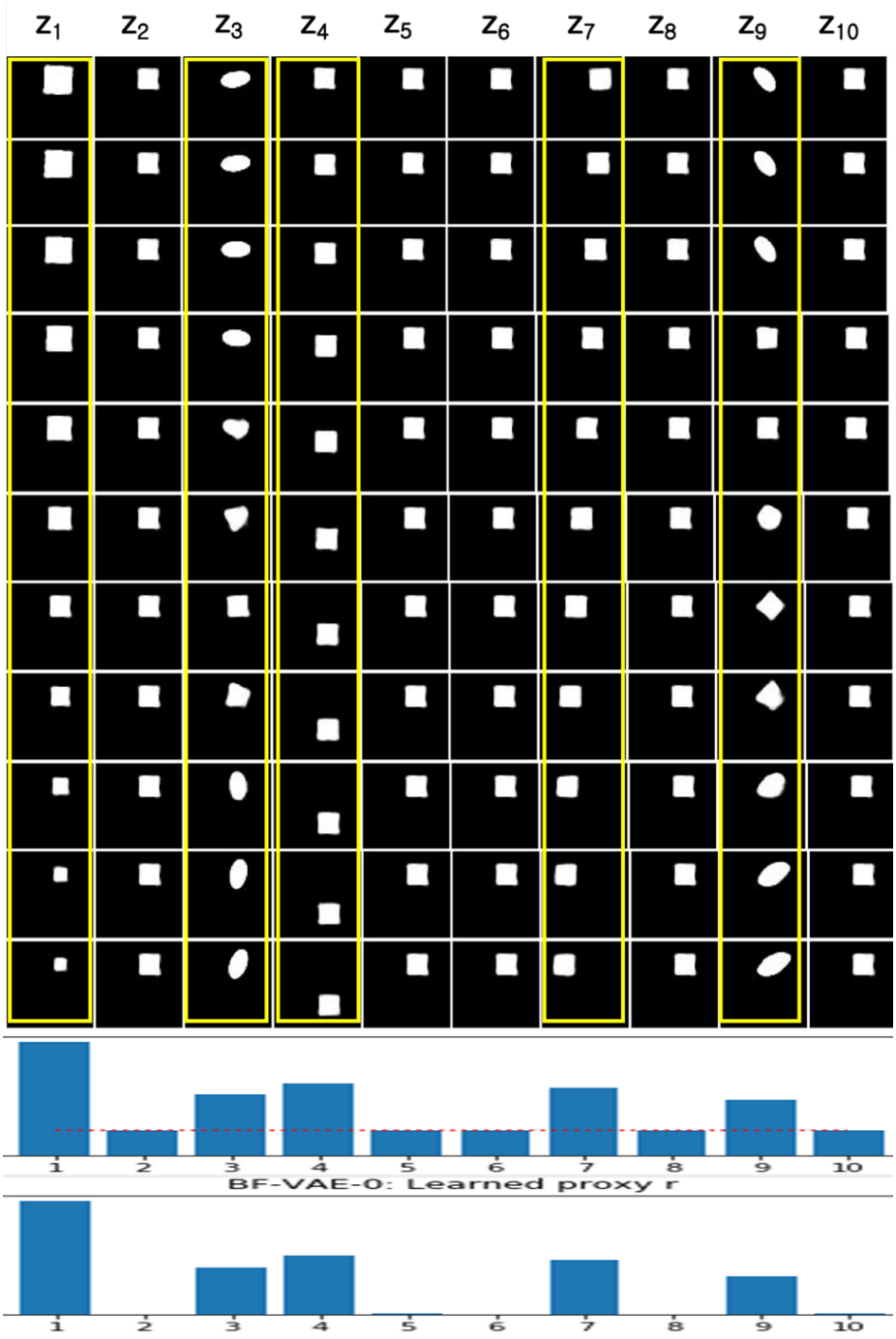

Figure 4. Latent space traversal in BF-VAE-0 on the `Sprites` dataset. The learned prior variances $\boldsymbol{\alpha}^{-1}$ (the value 1.0 depicted as the red dotted line) and the (proxy) relevance indicator $\hat{r}_j = |1 - \alpha_j^{-1}|$ are shown at the bottom. The five visually evident dimensions of variability (z_1, z_3, z_4, z_7, z_9) are highlighted within colored boxes, where each corresponds to: $z_1 =$ scale, $z_3 =$ shape and rotation, $z_4 = Y$-pos, $z_7 = X$-pos, and $z_9 =$ shape and rotation. The learned α_j for all these five dims are away from 1, as we anticipated, while $\hat{r}_j$ are high for those dimensions.

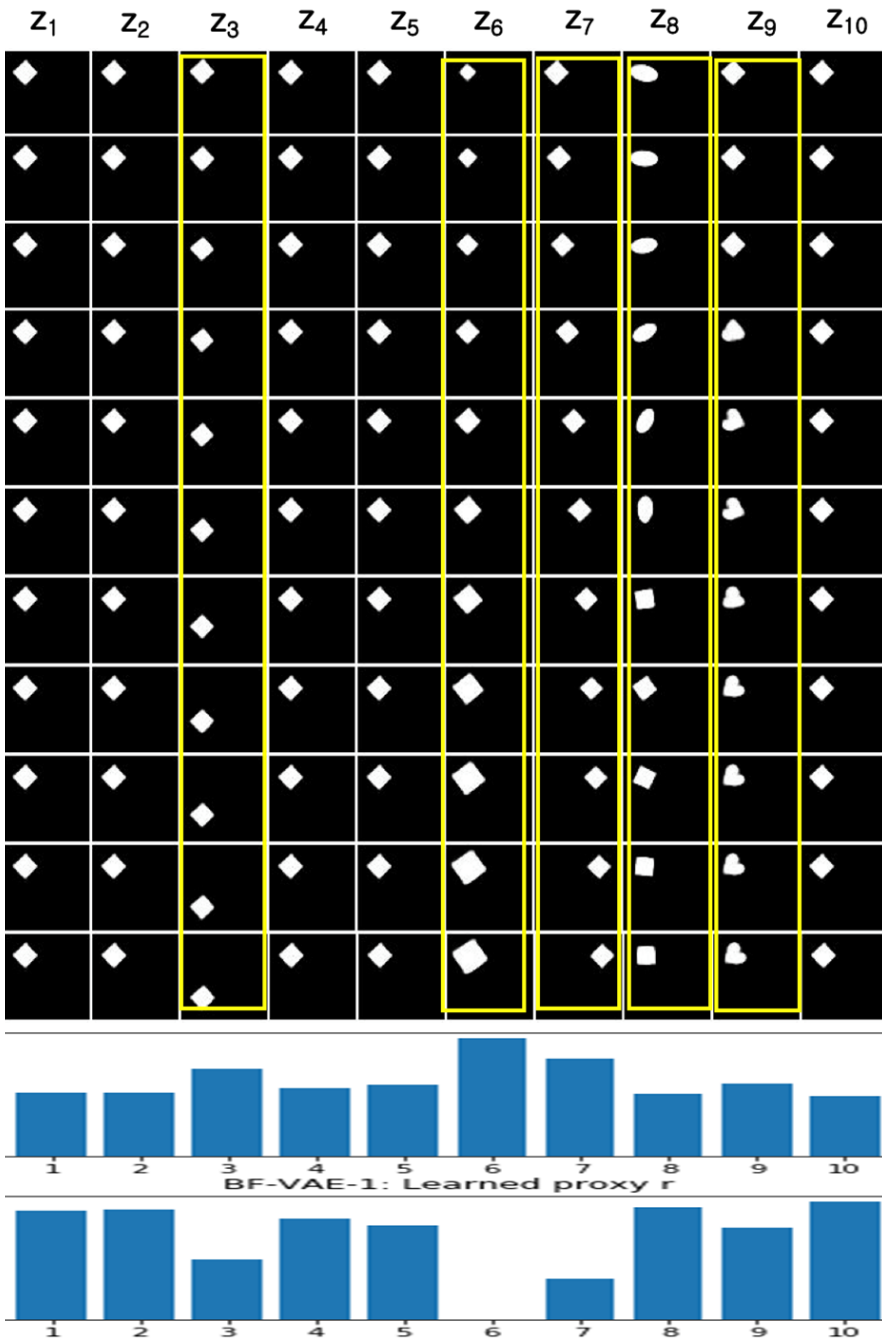

Figure 5. Latent space traversal in BF-VAE-1 on the `Sprites` dataset. The DOF ($2\hat{a}_j$) of the corrected prior $\overline{p}(z_j)$ and the (proxy) relevance indicator $\hat{r}_j$ (obtained by linearly transforming/interpolating the DOF into $[0, 1]$) are shown at the bottom. The five recovered, highlighted, dimensions match the ground-truth factors.

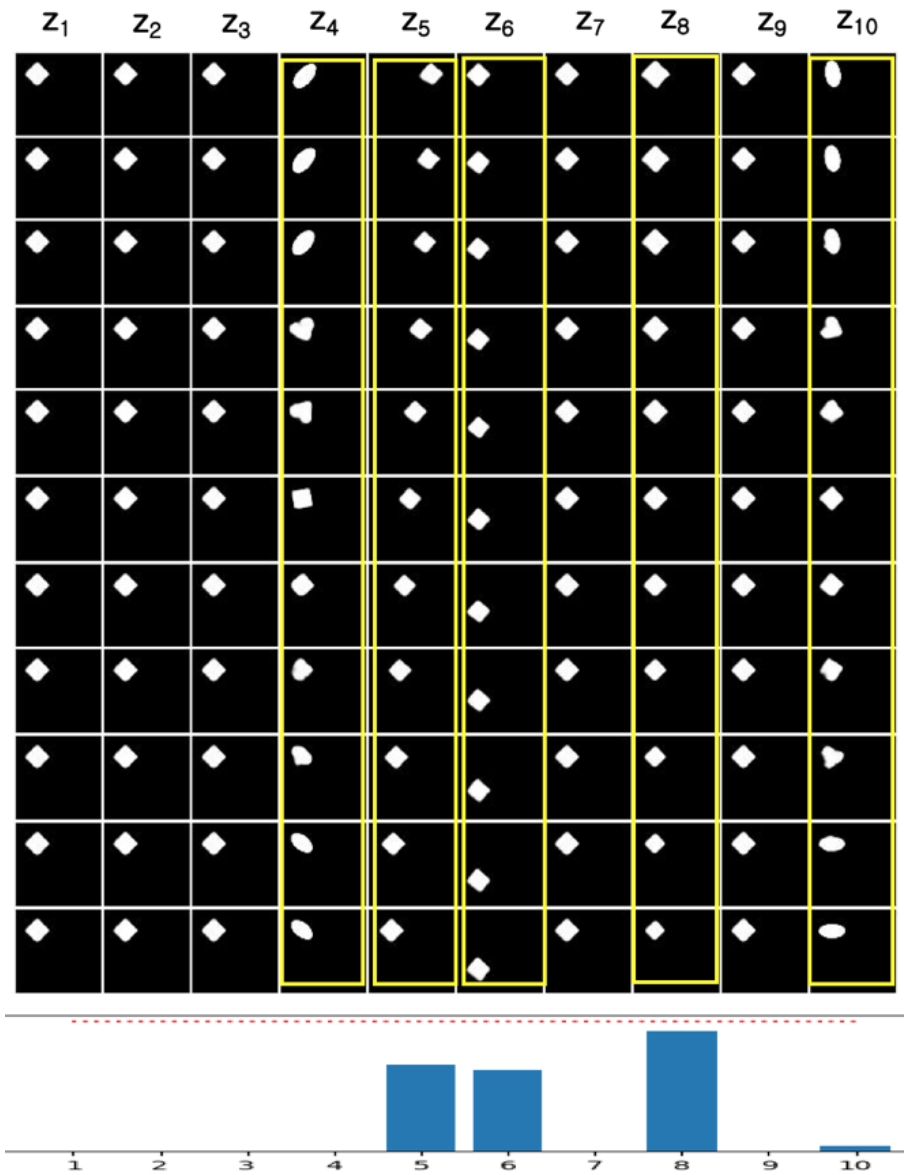

Figure 6. Latent space traversal in BF-VAE-2 on the `Sprites` dataset. Again the five factors are nearly correctly identified except for $j = 4$, corresponding to the high values in the indicator variables r_j's.

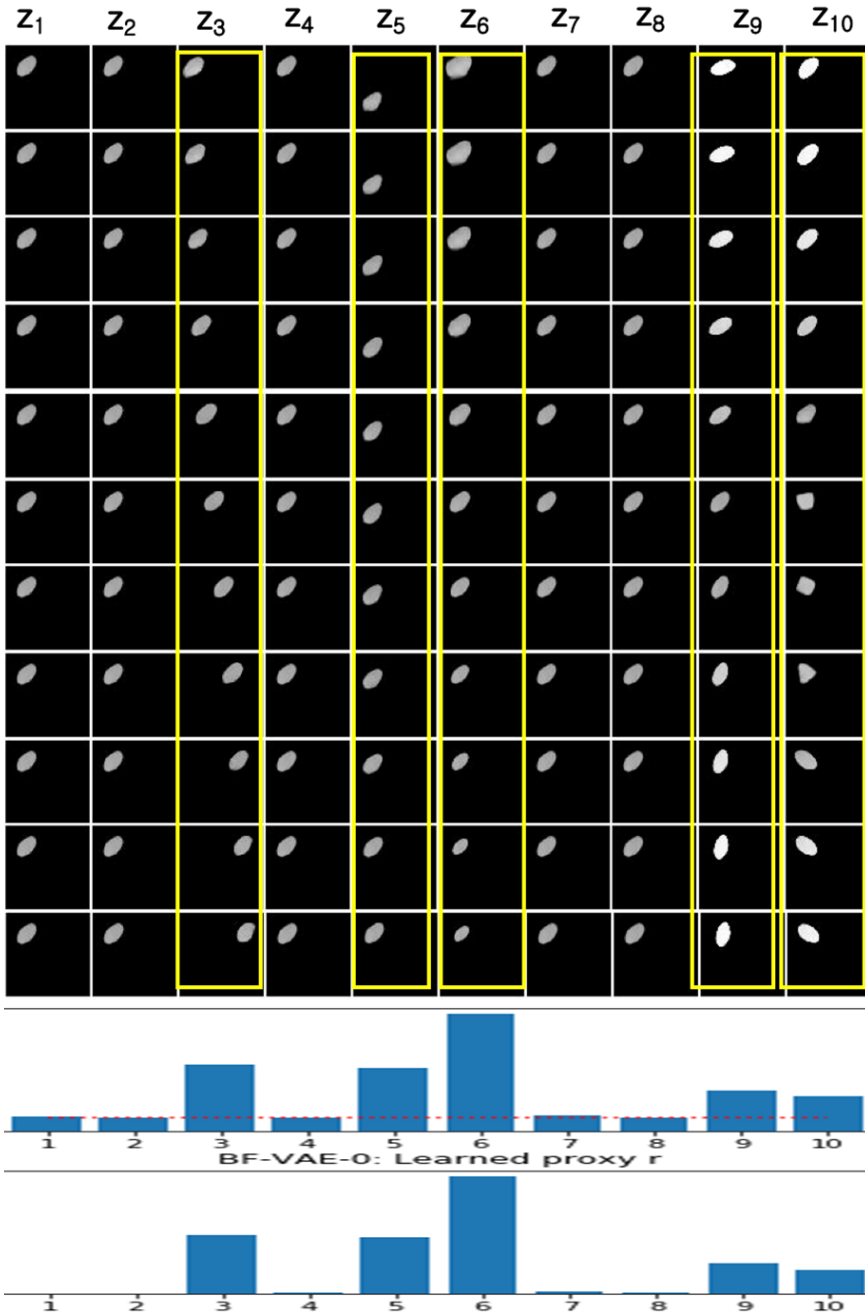

Figure 7. Latent space traversal in BF-VAE-0 on the `C-Spr` dataset. The learned prior variances $\boldsymbol{\alpha}^{-1}$ (the value 1.0 depicted as the red dotted line) and the (proxy) relevance indicator $\hat{r}_j = |1 - \alpha_j^{-1}|$ are shown at the bottom. The five visually evident dimensions of variability (z_3, z_5, z_6, z_9, z_{10}) are highlighted within colored boxes, where each exactly matches one of the five ground-truth factors ($z_3 = X$-pos, $z_5 = Y$-pos, $z_6 =$ scale, $z_9 =$ rotation, and $z_{10} =$ shape and rotation). The learned α_j for all these five dims are away from 1, while $\hat{r}_j$ are high for those dimensions.

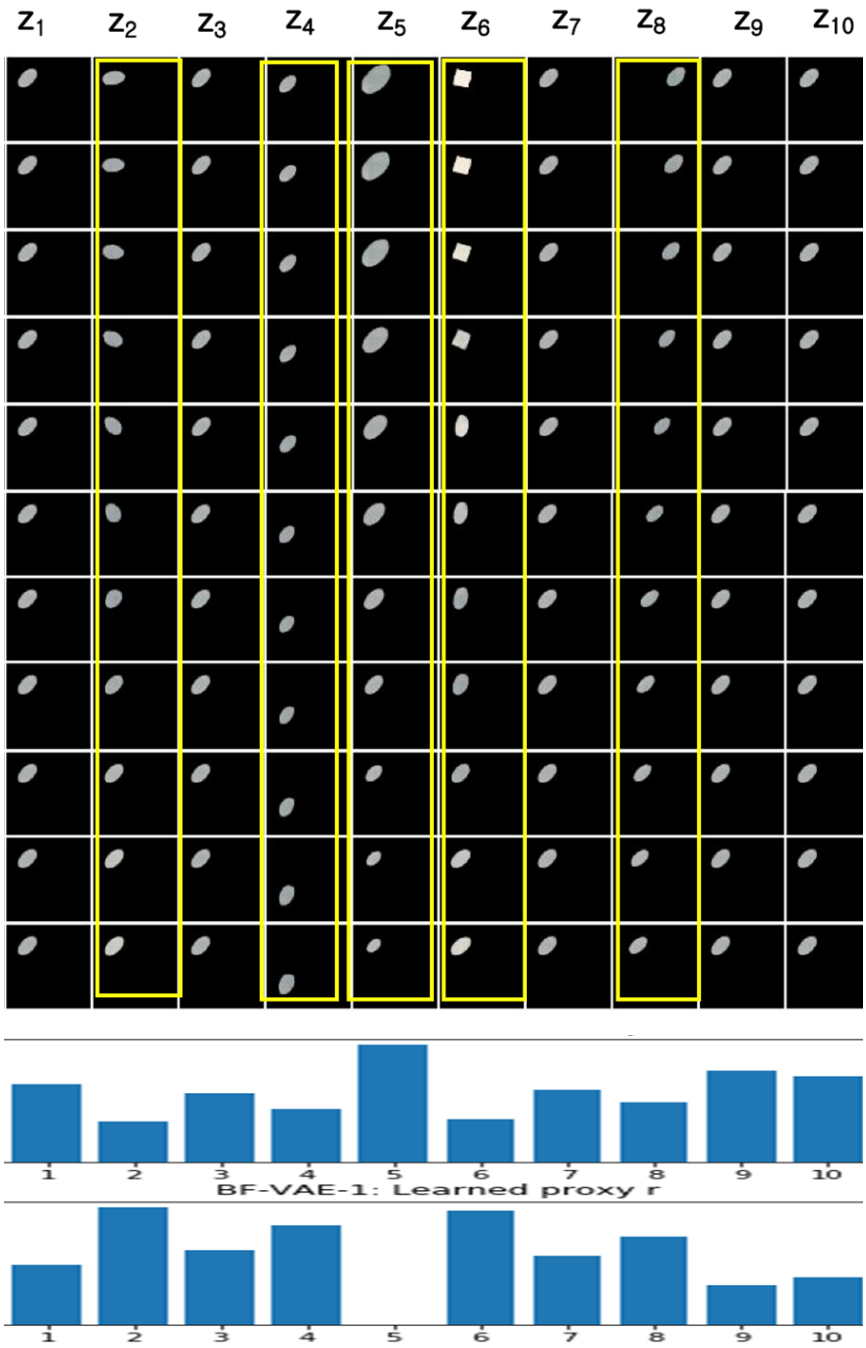

Figure 8. Latent space traversal in BF-VAE-1 on the `C-Spr` dataset. The DOF ($2\hat{a}_j$) of the corrected prior $\overline{p}(z_j)$ and the (proxy) relevance indicator $\hat{r}_j$ (obtained by linearly transforming/interpolating the DOF into $[0, 1]$) are shown at the bottom. The five recovered, highlighted, dimensions match the ground-truth factors, and their $\overline{p}(z_j)$'s also have relatively small DOFs and large proxy indicator values, except for $j = 5$.

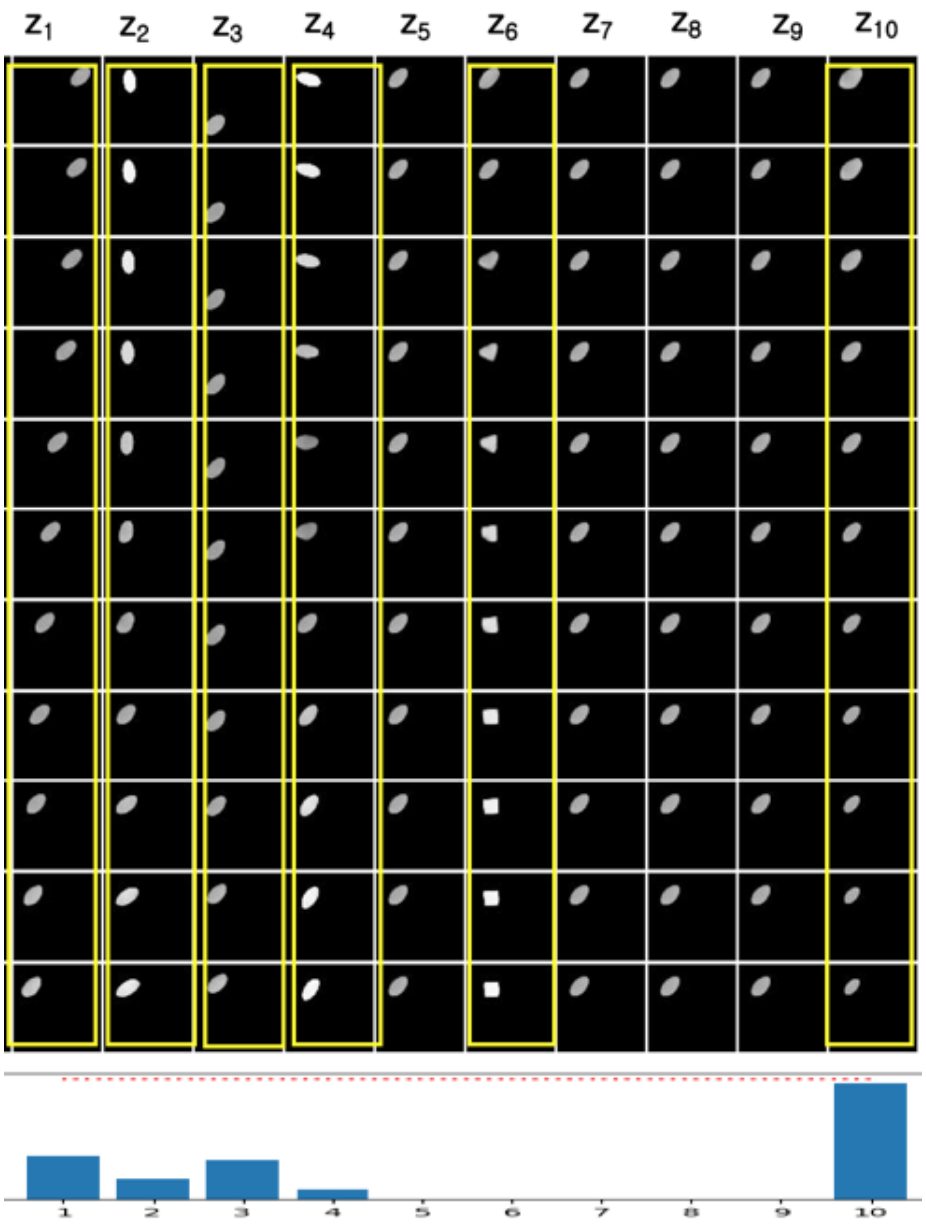

Figure 9. Latent space traversal in BF-VAE-2 on the `C-Spr` dataset. Again the five factors are nearly correctly identified, corresponding to the high values in the indicator variables r_j's, except for $j = 6$.

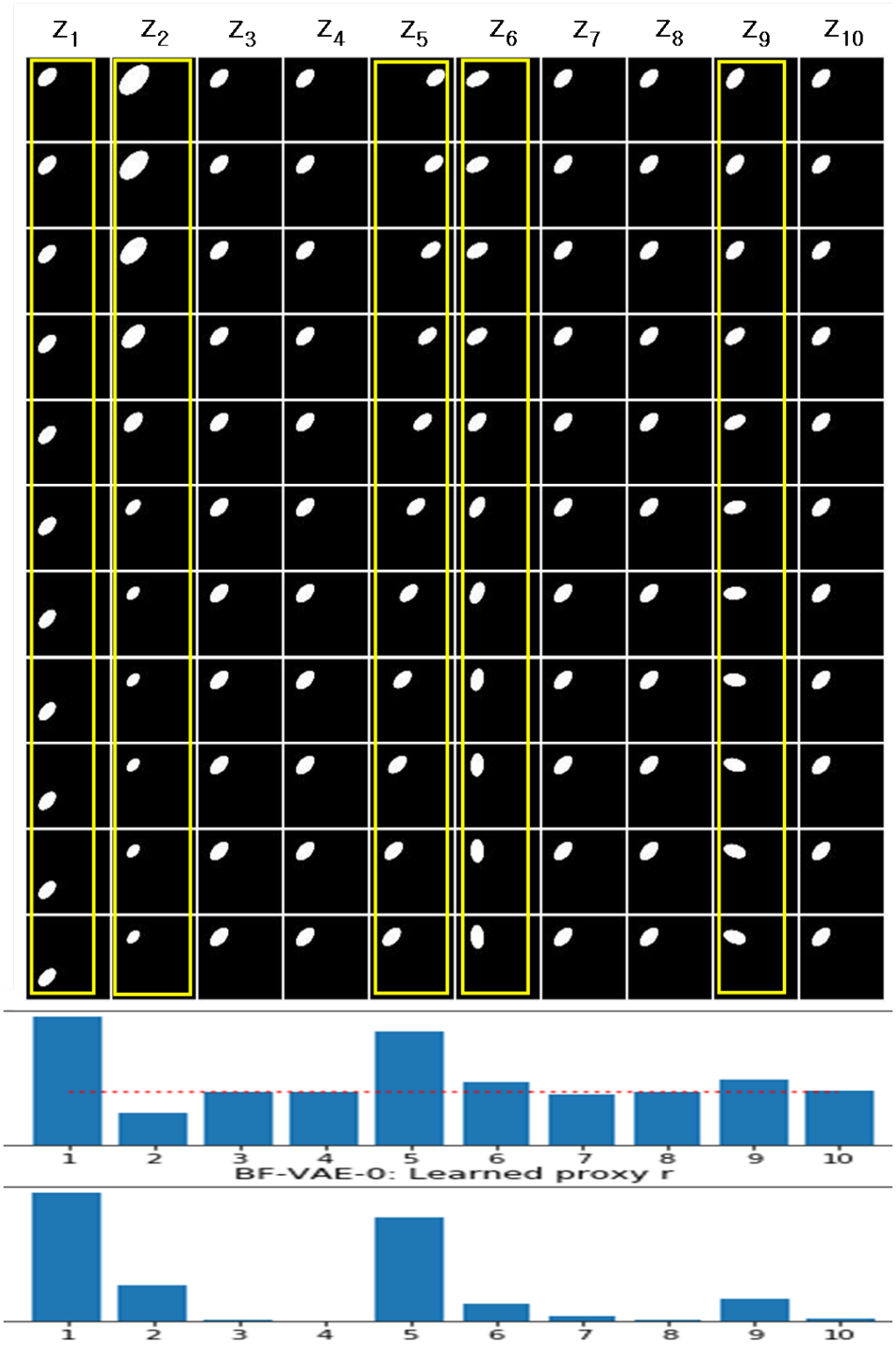

Figure 10. Latent space traversal in BF-VAE-0 on the `O-Spr` dataset. The learned prior variances $\boldsymbol{\alpha}^{-1}$ (the value 1.0 depicted as the red dotted line) and the (proxy) relevance indicator $\hat{r}_j = |1 - \alpha_j^{-1}|$ are shown at the bottom. Those five highlighted dimensions of major variability (z_1, z_2, z_5, z_6, z_9), match the four ground-truth factors (scale, X-, Y-pos, rotation), while the rotation is spread across z_6 and z_9. These factors also exactly correspond to the learned α_j's that are distant from 1, as we anticipated, while $\hat{r}_j$ are high for those dimensions.

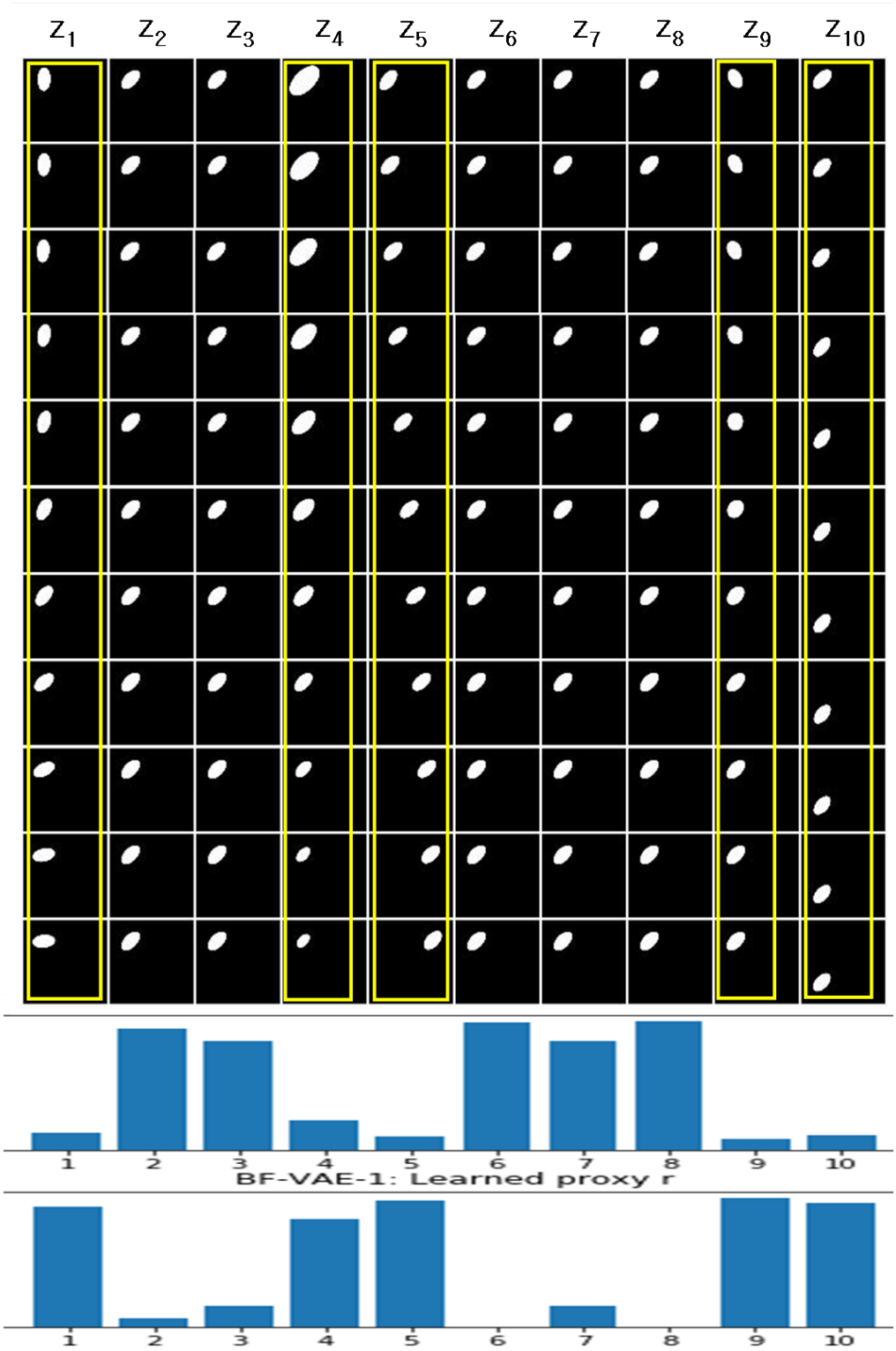

Figure 11. Latent space traversal in BF-VAE-1 on the `O-Spr` dataset. The DOF ($2\hat{a}_j$) of the corrected prior $\overline{p}(z_j)$ and the (proxy) relevance indicator $\hat{r}_j$ (obtained by linearly transforming/interpolating the DOF into $[0, 1]$) are shown at the bottom. Similar to BF-VAE-0, it identifies five variables with the rotation spread across z_1 and z_9. These relevant variables, as expected, have small DOFs in $\overline{p}(z_j)$'s, while the proxy indicator $\hat{r}_j$ for these dimensions are high (away from 0).

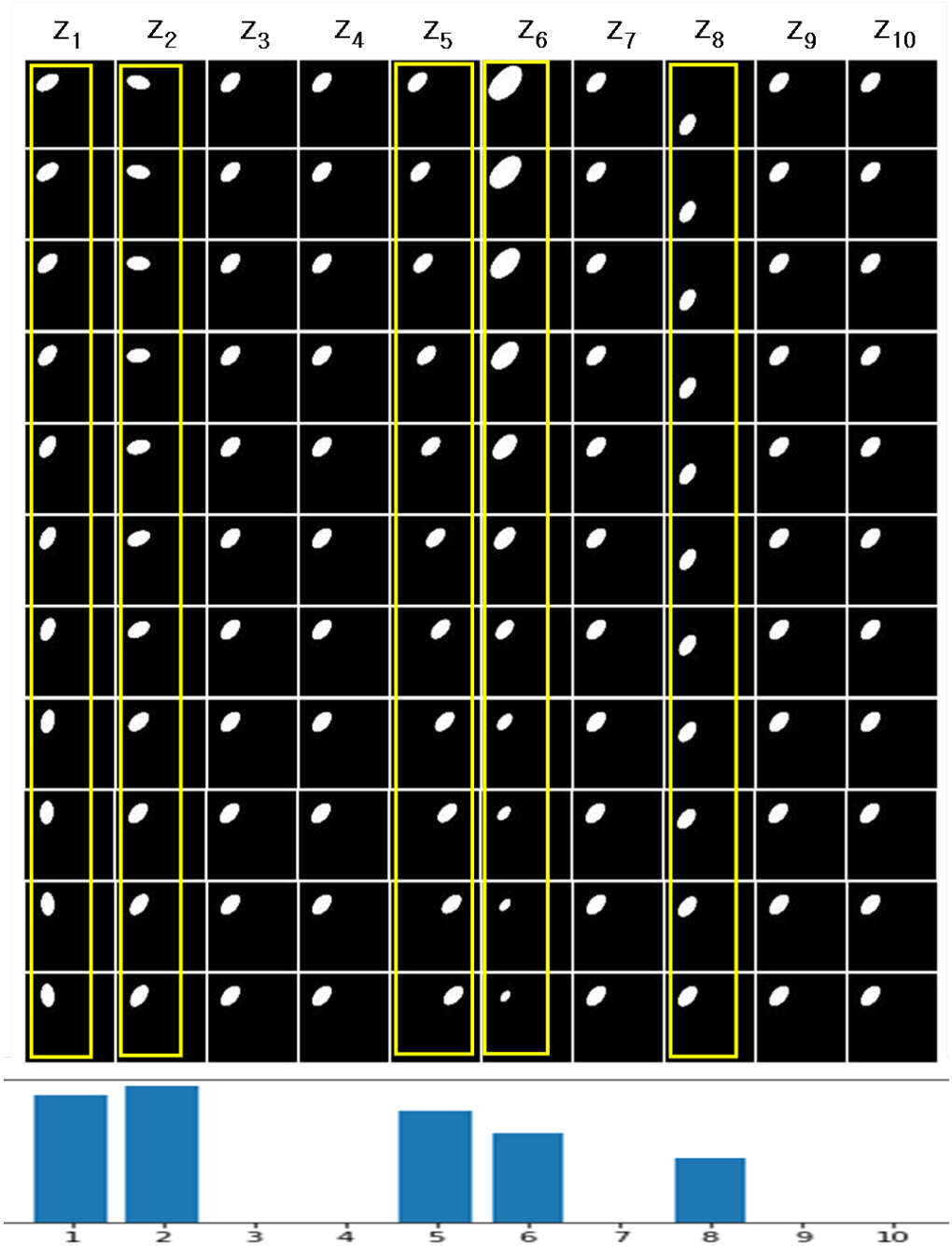

Figure 12. Latent space traversal in BF-VAE-2 on the `O-Spr` dataset. The learned relevance vector $\mathbf{r}$ is shown at the bottom. The learned $\mathbf{r}$ accurately indicates the relevant dimensions.

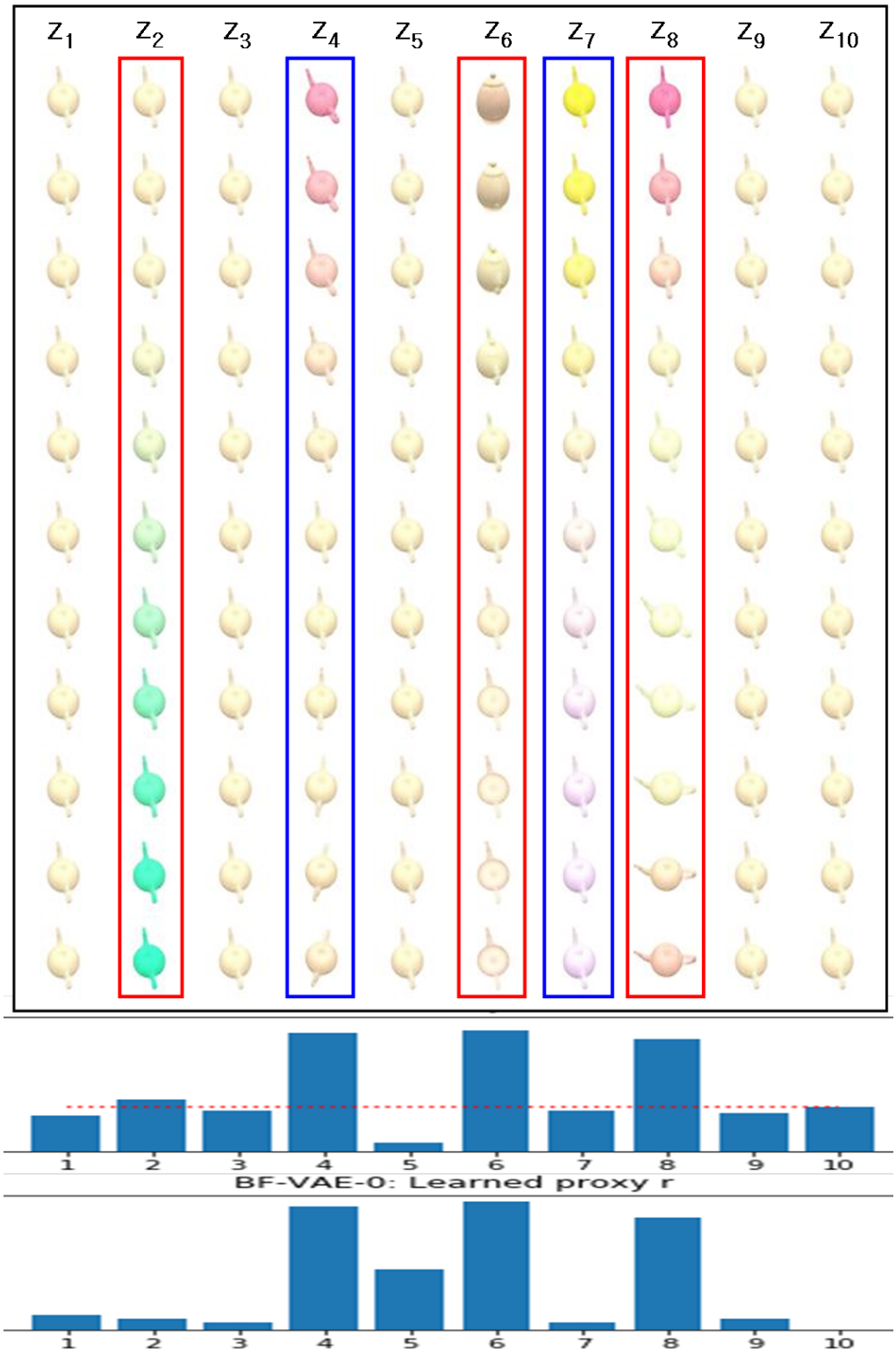

Figure 13. Latent space traversal in BF-VAE-0 on the `Teapots` dataset. The learned prior variances $\boldsymbol{\alpha}^{-1}$ (the value 1.0 depicted as the red dotted line) and the (proxy) relevance indicator $\hat{r}_j = |1 - \alpha_j^{-1}|$ are shown at the bottom. The five variables that explain the major variability in images, (z_2, z_4, z_6, z_7, z_8), do not perfectly match the true factors one by one, and two or more factors are entangled in some variables (e.g., z_8 explains both color R and azimuth). Note that a similar failure was also observed in [1] with complex ResNet models.

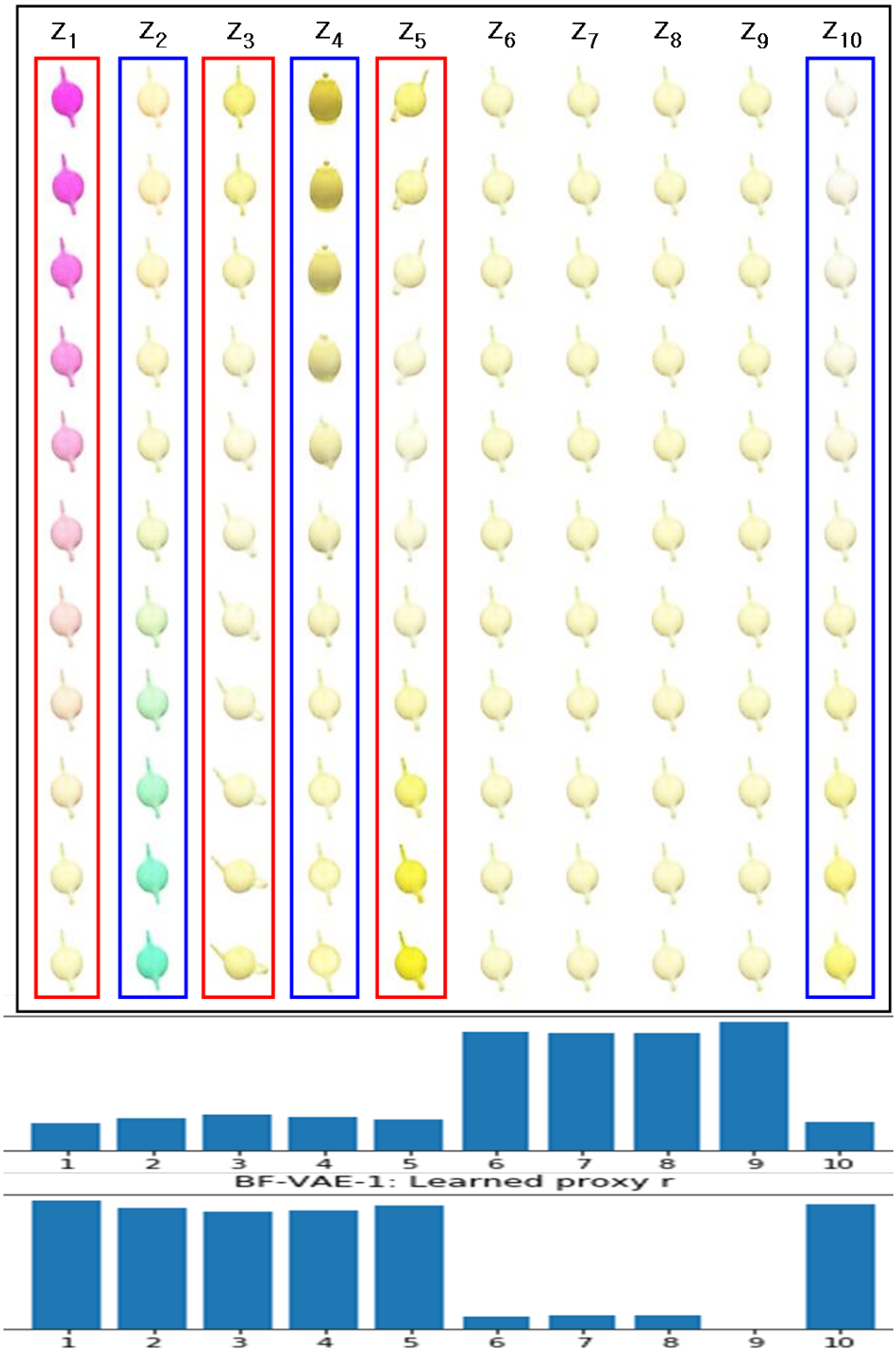

Figure 14. Latent space traversal in BF-VAE-1 on the `Teapots` dataset. The DOF ($2\hat{a}_j$) of the corrected prior $\overline{p}(z_j)$ and the (proxy) relevance indicator $\hat{r}_j$ (obtained by linearly transforming/interpolating the DOF into $[0, 1]$) are shown at the bottom. Overall similar behaviors as BF-VAE-0, but the DOFs of the corrected prior are small for those relevant dimensions correctly identifying the dimensions of major variability, while the proxy indicator $\hat{r}_j$ for these dimensions are high (away from 0).

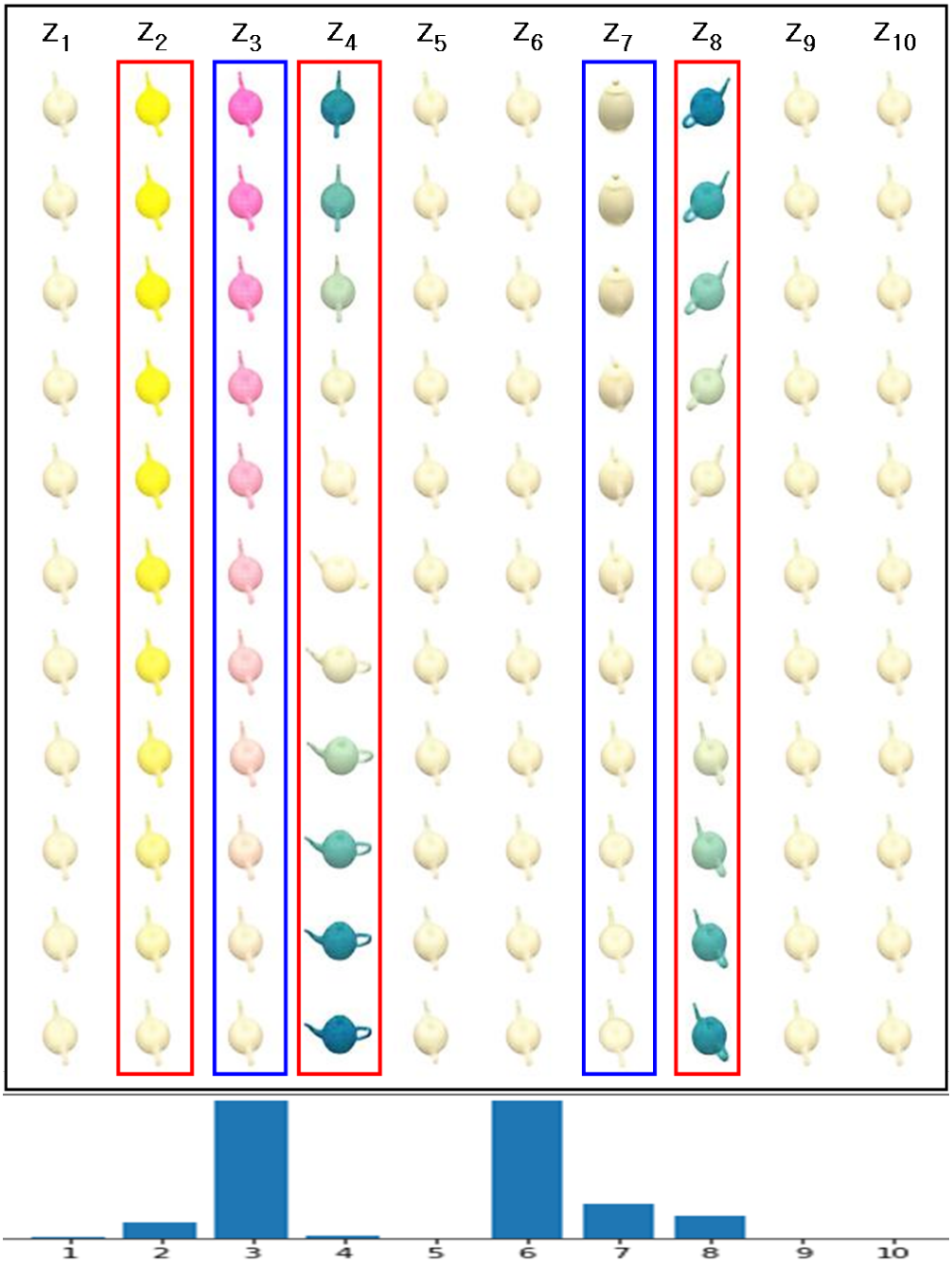

Figure 15. Latent space traversal in BF-VAE-2 on the `Teapots` dataset. The learned relevance vector $\mathbf{r}$ is shown at the bottom. The learned $\mathbf{r}$ accurately indicates the relevant dimensions for majority of dimensions, with the exception of z_6. Upon careful examination, this was related to the choice of the traversal range, which was significantly different for this factor from the range of other factors.

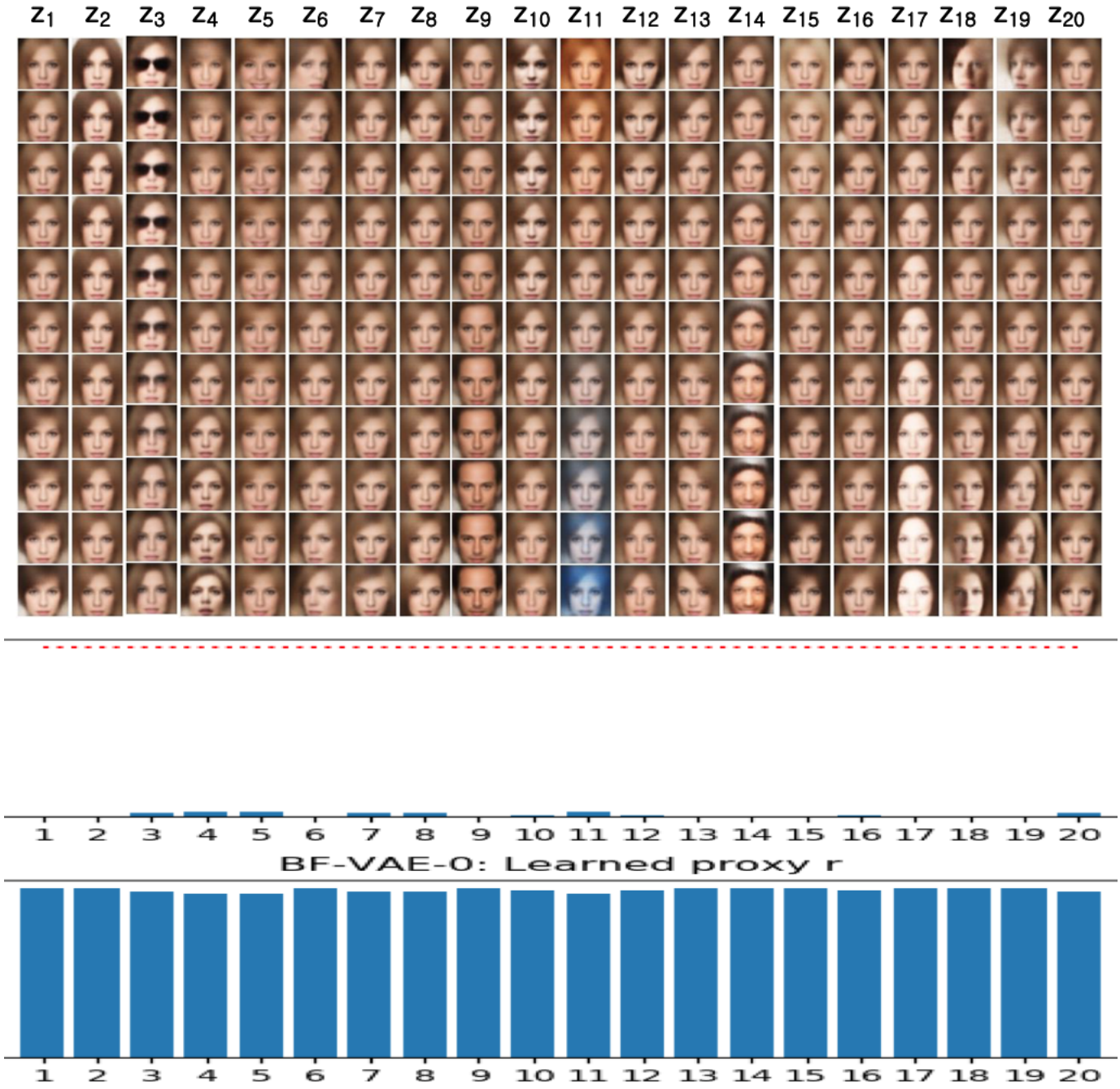

Figure 16. Latent traversal in BF-VAE-0 on `Celeb-A`, where the learned prior variances $\boldsymbol{\alpha}^{-1}$ shown at the bottom (the value 1.0 depicted as the red dotted line). The learned α_j for all j's are away from 1, and the model identifies many relevant variables (in fact, all dimensions affect image observation visually). Some latent dimensions visually evidently correspond to specific factors, e.g. z_1 = frontal hair, z_3 = sunglasses, z_4 = elevation, z_5 = smiling, z_6/z_{19} = azimuth, z_9 = gender, z_{11} = hue, z_{15} = hair color, z_{17} = brightness, and z_{18} = lighting (azimuth).

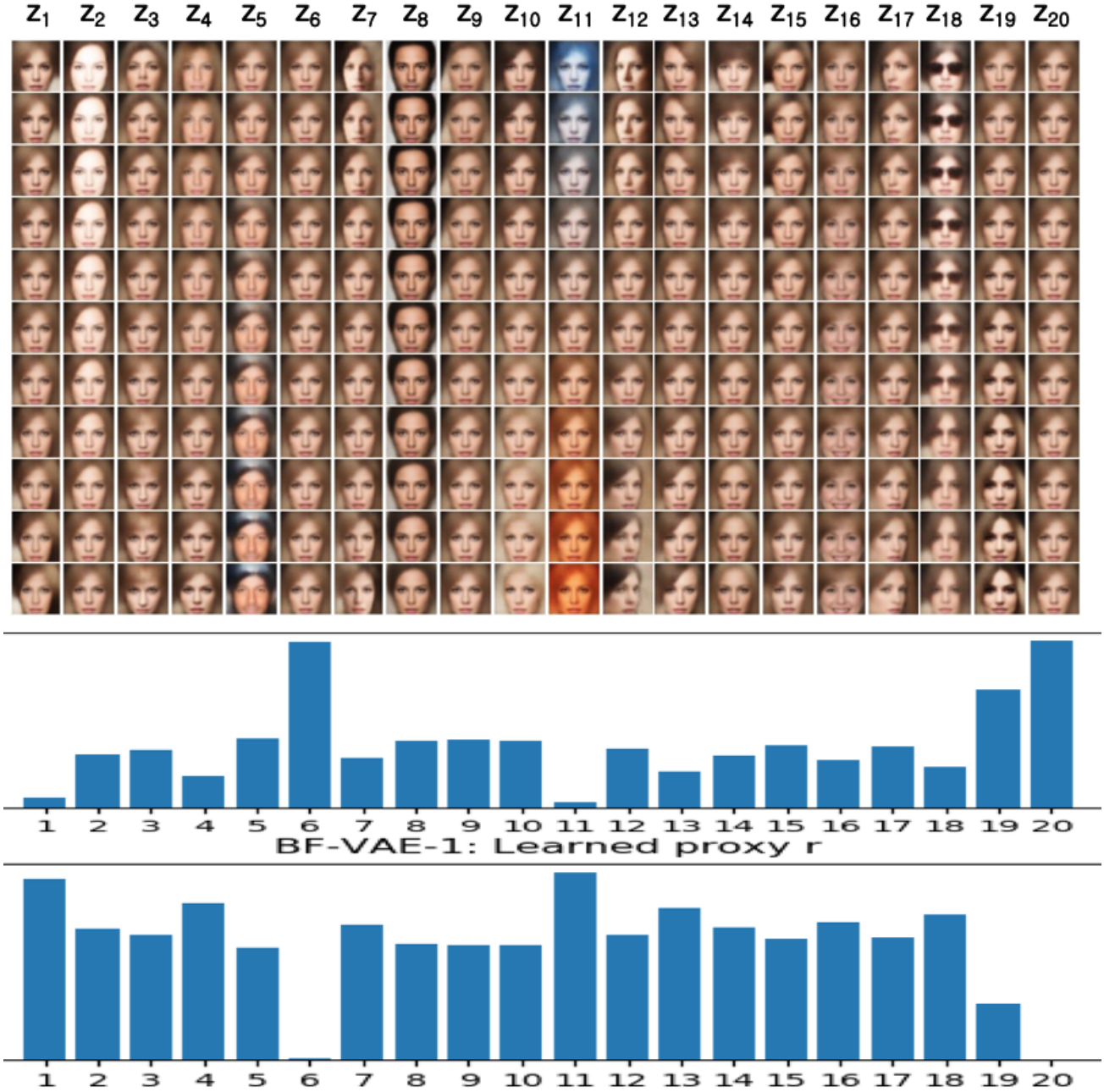

Figure 17. Latent traversal in BF-VAE-1 on `Celeb-A`, where the DOF ($2\hat{a}_j$) of the corrected prior $\overline{p}(z_j)$ shown at the bottom. Except dimensions z_6 and z_{20}, all dimensions correspond to some meaningful factors visually, and the DOFs are mostly relatively small, and most of the proxy indicators are active. Some latent dimensions visually evidently correspond to specific factors, e.g. z_2 = brightness, z_3 = elevation, z_7 = lighting (azimuth), z_8 = gender, z_{10} = hair color, z_{11} = hue, z_{12} = azimuth, z_{14} = frontal hair, z_{16} = smiling, z_{18} = sunglasses.

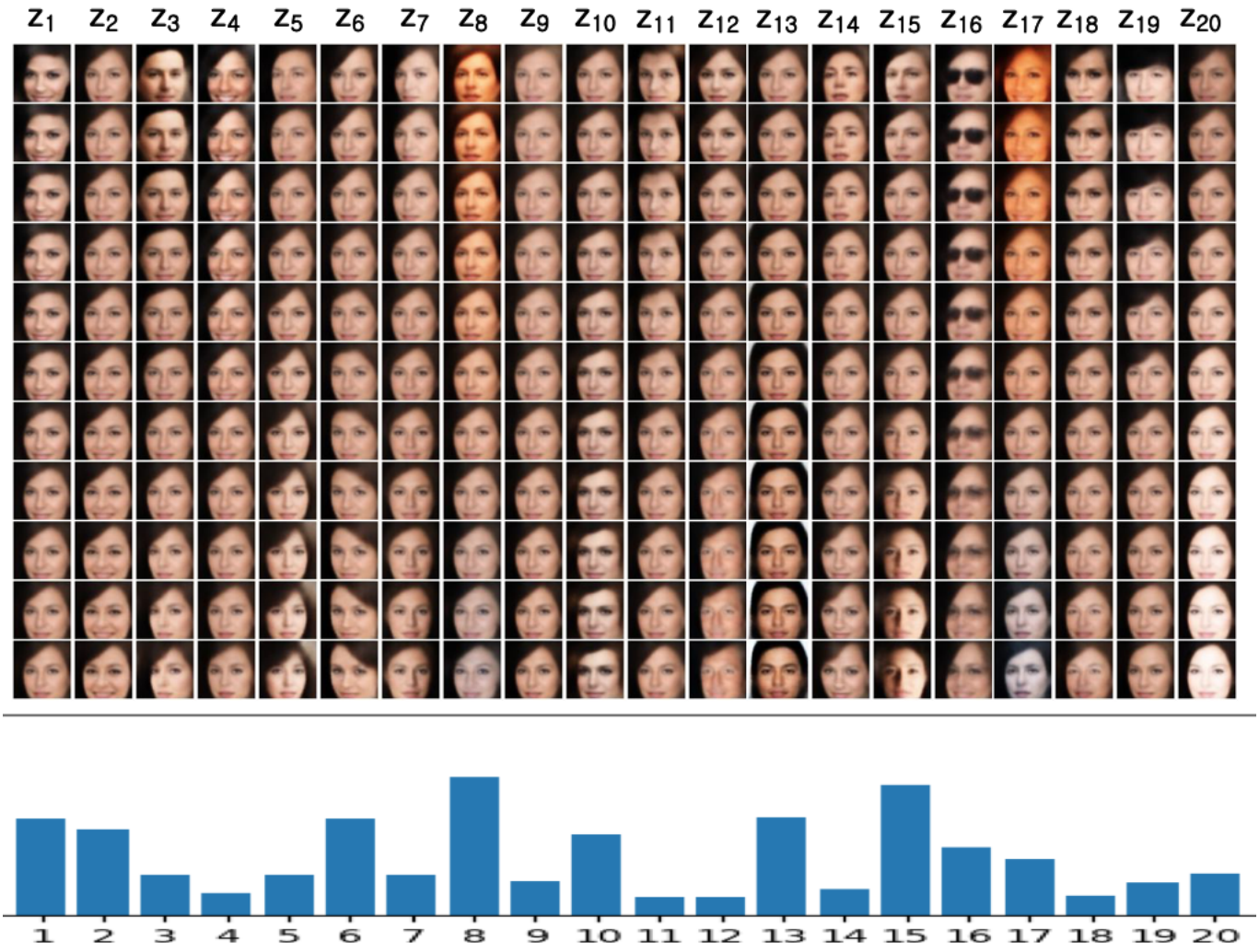

Figure 18. Latent traversal in BF-VAE-2 on `Celeb-A`, where the learned relevance vector $\mathbf{r}$ shown at the bottom. The model is able to detect many relevant variables. Some latent dimensions visually evidently correspond to specific factors, e.g. z_2/z_4 = smiling, z_5/z_{15} = azimuth, z_8/z_{17} = hue, z_9 = baldness, z_{12}/z_{13} = gender, z_{14} = elevation, z_{16} = sunglasses, z_{19} = frontal hair, and z_{20} = brightness.

Figure 19. Latent space traversal in BF-VAE-2 on `Celeb-A`. We train two BF-VAE-2 models with two different η values ($\eta = \eta_S = \eta_H$ large and small). **(Left panel: strong factors)** contains latent traversal results with four latent variables (two subjects for each) that are detected (according to high r_j) by both η small and large models. They correspond to (from top to bottom): gender, frontal hair, azimuth, and brightness, which are considered as strong/major factors. **(Right panel: weak factors)** shows traversal with four other latent variables that are detected (according to high r_j) only by the small η model. They correspond to: smiling, sunglasses, elevation, and baldness, which are considered as weak/minor factors.

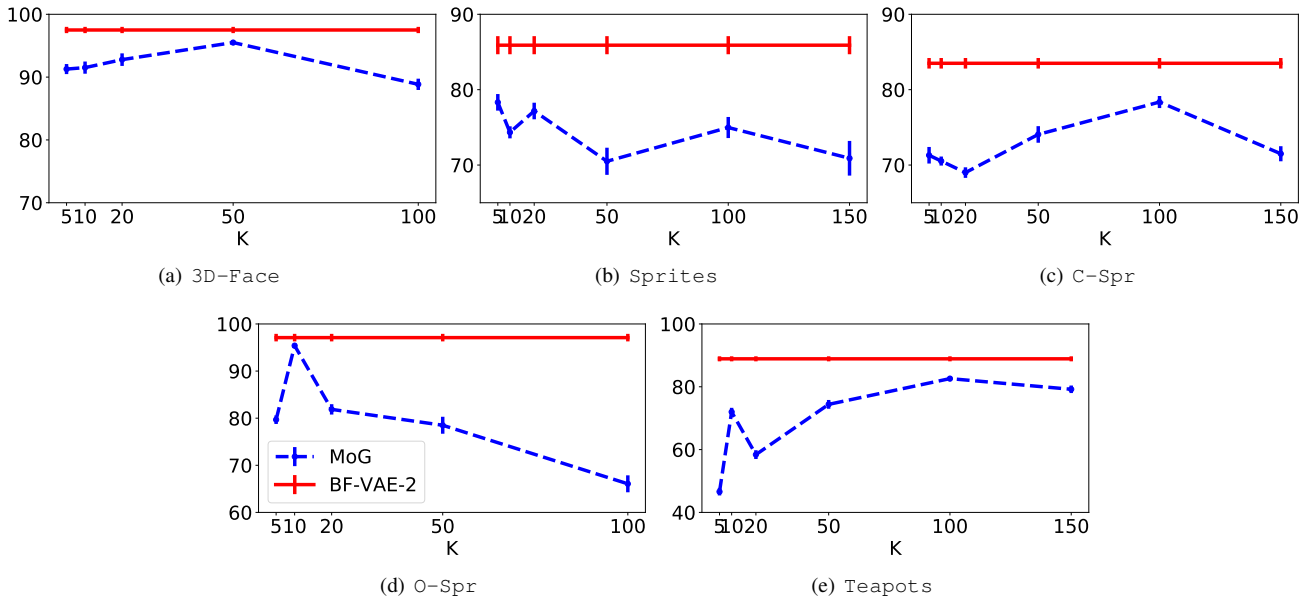

Figure 20. Disentanglement performance (Metric II) of F-VAE with MoG prior (Blue/Dashed) with different mixture orders (K) vs. BF-VAE-2 (Red/Solid). Overall, the MoG overfits as K increases for all datasets.