Supplementary material for ICML 2021
An Identifiable Double VAE For Disentangled Representations

A. ELBO derivation for IDVAE

$$\begin{aligned}
\log p(\mathbf{x},\mathbf{u}) &= \log \int p(\mathbf{x},\mathbf{u},\mathbf{z}) d\mathbf{z} = \\
&= \log \int p(\mathbf{x}|\mathbf{u},\mathbf{z}) p(\mathbf{z}|\mathbf{u}) p(\mathbf{u}) d\mathbf{z} = \\
&= \log \int \frac{p(\mathbf{x}|\mathbf{u},\mathbf{z}) p(\mathbf{z}|\mathbf{u}) p(\mathbf{u})}{q(\mathbf{z}|\mathbf{x},\mathbf{u})} q(\mathbf{z}|\mathbf{x},\mathbf{u}) d\mathbf{z} \geq \mathcal{L}_{\text{IDVAE}} \\
&= \mathbb{E}_{q(\mathbf{z}|\mathbf{x},\mathbf{u})}[\log \frac{p(\mathbf{x}|\mathbf{u},\mathbf{z}) p(\mathbf{z}|\mathbf{u}) p(\mathbf{u})}{q(\mathbf{z}|\mathbf{x},\mathbf{u})}] = \\
&= \mathbb{E}_{q(\mathbf{z}|\mathbf{x},\mathbf{u})}[\log p(\mathbf{x}|\mathbf{u},\mathbf{z})] - KL(q(\mathbf{z}|\mathbf{x},\mathbf{u})||p(\mathbf{z}|\mathbf{u})) + \log p(\mathbf{u}),
\end{aligned} \tag{1}$$

where:

$$\begin{aligned}
\log p(\mathbf{u}) &= \log \int p(\mathbf{u},\mathbf{z}) d\mathbf{z} \geq \mathcal{L}_{\text{prior}} = \\
&= \mathbb{E}_{q(\mathbf{z}|\mathbf{u})}[\log p(\mathbf{u}|\mathbf{z})] - KL(q(\mathbf{z}|\mathbf{u})||p(\mathbf{z})).
\end{aligned} \tag{2}$$

B. ELBO derivation for SS-IDVAE

$$\begin{aligned}
\log p(\mathbf{x}) &= \log \int p(\mathbf{x},\mathbf{u},\mathbf{z}) d\mathbf{u} d\mathbf{z} = \\
&= \log \int p(\mathbf{x}|\mathbf{u},\mathbf{z}) p(\mathbf{z}|\mathbf{u}) p(\mathbf{u}) d\mathbf{u} d\mathbf{z} = \\
&= \log \int \frac{p(\mathbf{x}|\mathbf{u},\mathbf{z}) p(\mathbf{z}|\mathbf{u}) p(\mathbf{u})}{q(\mathbf{u},\mathbf{z}|\mathbf{x})} q(\mathbf{u},\mathbf{z}|\mathbf{x}) d\mathbf{u} d\mathbf{z} \geq \\
&\geq \mathbb{E}_{q(\mathbf{u},\mathbf{z}|\mathbf{x})}[\log \frac{p(\mathbf{x}|\mathbf{u},\mathbf{z}) p(\mathbf{z}|\mathbf{u}) p(\mathbf{u})}{q(\mathbf{u},\mathbf{z}|\mathbf{x})}] = \\
&= \mathbb{E}_{q(\mathbf{z}|\mathbf{x},\mathbf{u}) q(\mathbf{u}|\mathbf{x})}[\log \frac{p(\mathbf{x}|\mathbf{u},\mathbf{z}) p(\mathbf{z}|\mathbf{u}) p(\mathbf{u})}{q(\mathbf{z}|\mathbf{x},\mathbf{u}) q(\mathbf{u}|\mathbf{x})}] = \\
&= \mathbb{E}_{q(\mathbf{u}|\mathbf{x})}[\mathcal{L}_{\text{IDVAE}}] + \mathcal{H}(q(\mathbf{u}|\mathbf{x})).
\end{aligned} \tag{3}$$

Combining eqs. (1) to (3) we obtain $\mathcal{L}_{\text{SS-IDVAE}}$, where it is clear that we use the sum over the data samples instead of the expectation. As stated in the main paper, we also add the term – $\mathbb{E}_{(\mathbf{x},\mathbf{u})\sim p_l}[\log q(\mathbf{u}|\mathbf{x})]$ – such that it can learn also from labeled data.

C. Sketch of the proof of Theorem 1

In this section, we report a sketch of the proof of Theorem 1. Following the proof strategy of Khemakhem et al. (2020), the proof consists of three main steps.

In the first step, we use assumption (i) to demonstrate that observed data distributions are equal to noiseless distributions. Supposing to have two sets of parameters $(\mathbf{f}, \mathbf{T}, \boldsymbol{\eta})$ and $(\tilde{\mathbf{f}}, \tilde{\mathbf{T}}, \tilde{\boldsymbol{\eta}})$, with a change of variable $\bar{\mathbf{x}} = \mathbf{f}(\mathbf{z}) = \tilde{\mathbf{f}}(\mathbf{z})$, we show that:

$$\tilde{p}_{\mathbf{T},\boldsymbol{\eta},\mathbf{f},\mathbf{u}}(\mathbf{x}) = \tilde{p}_{\tilde{\mathbf{T}},\tilde{\boldsymbol{\eta}},\tilde{\mathbf{f}},\tilde{\mathbf{u}}}(\mathbf{x}), \tag{4}$$

where:

$$\tilde{p}_{\mathbf{T},\boldsymbol{\eta},\mathbf{f},\mathbf{u}}(\mathbf{x}) = p_{\mathbf{T},\boldsymbol{\eta}}(\mathbf{f}^{-1}(\mathbf{x})|\mathbf{u})|detJ_{\mathbf{f}^{-1}}(\mathbf{x})|\mathbb{1}_{\mathcal{X}}(\mathbf{x}) \tag{5}$$

In the second step, we use assumption (iv) to remove all the terms that are a function of $\mathbf{x}$ or $\mathbf{u}$. By substituting $p_{\mathbf{T},\boldsymbol{\eta}}$ with its exponential conditionally factorial form, taking the log of both sides of eq. (5), we obtain $dk + 1$ equations. Then:

$$\mathbf{T}(\mathbf{f}^{-1}(\mathbf{x})) = \mathbf{A}\mathbf{T}'(\mathbf{f}'^{-1}(\mathbf{x})) + \mathbf{c}. \tag{6}$$

In the last step, assumptions (i) and (iii) are used to show that the linear transformation is invertible and so $(\mathbf{f}, \mathbf{T}, \boldsymbol{\eta}) \sim (\tilde{\mathbf{f}}, \tilde{\mathbf{T}}, \tilde{\boldsymbol{\eta}})$. This concludes the proof.

For a full derivation of the proof, we point the reader to section $\mathbf{B}$ of the supplement in Khemakhem et al. (2020), which holds also for our variant of the theorem.

D. Model architectures, parameters and hyperparameters

All the selected methods (including the semi-supervised variants) share the same convolutional architecture. The conditional prior in IVAE is a MLP network, in IDVAE we use a simple MLP VAE, both with leaky ReLU activation functions. The ground-truth factor learner implementing $q_{\boldsymbol{\zeta}}(\mathbf{u}|\mathbf{x})$ in SS-IDVAE and SS-IVAE is a convolutional neural network.

Encoder	**Decoder**
Input: $64 \times 64 \times$ number of channels	Input: $\mathbb{R}^d$, where d is the number of ground-truth factors
4×4 conv, 32 ReLU, stride 2	FC, 256 ReLU
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*, FC $2 \times d$	4×4 upconv, 32 ReLU, stride 2
	4×4 upconv, number of channels, stride 2

Table 1. Main Encoder-Decoder architecture. In IVAE and IDVAE, we give $\mathbf{u}$ as input to the fully connected layer of the Encoder which size becomes $256 + d$.

Conditional Prior Encoder	**Conditional Prior Decoder**
FC, 1000 leaky ReLU	FC, 1000 leaky ReLU
FC, 1000 leaky ReLU	FC, 1000 leaky ReLU
FC, 1000 leaky ReLU	FC, 1000 leaky ReLU
FC $2 \times d$	FC d

Table 2. IDVAE Conditional Prior Encoder-Decoder architecture. IVAE uses the encoder only.

Ground-truth Factor Learner
Input: $64 \times 64 \times$ number of channels. d is the number of ground-truth factors.
4×4 conv, 32 ReLU, stride 2
4×4 conv, 32 ReLU, stride 2
4×4 conv, 64 ReLU, stride 2
4×4 conv, 64 ReLU, stride 2
FC 256, FC $2 \times d$

Table 3. Ground-truth factor learner implementing $q_\varsigma(\mathbf{u}|\mathbf{x})$ in SS-IDVAE and SS-IVAE.

Parameters	**Values**
batch_size	64
optimizer	Adam
Adam: beta1	0.9
Adam: beta2	0.999
Adam: epsilon	1e-8
Adam: learning_rate	1e-4
training_steps	300'000

Table 4. Common hyperparameters to each of the considered methods.

E. Implementation of disentanglement metrics

Beta score The idea behind the beta score (Higgins et al., 2017) is to fix a random ground-truth factor and sample two mini batches of observations from the corresponding generative model. The encoder is then used to obtain a learned representation from the observations (with a ground-truth factor in common). The dimension-wise absolute difference between the two representation is computed and a simple linear classifier C is used to predict the corresponding ground-truth factor. This is repeated $batch\_size$ times and the accuracy of the predictor is the disentanglement metric score.

MIG - Mutual Information Gap The mutual information gap (MIG) (Chen et al., 2018) is computed as the average, normalized difference between the highest and second highest mutual information of each ground-truth factor with the dimensions of the learned representation. As done in Locatello et al. (2019), we consider the mean representation. and compute the discrete mutual information by binning each dimension of the mean learned representation into n_bins bins.

Modularity and Explicitness A representation is modular if each dimension depends on at most one ground-truth factor. Ridgeway and Mozer (2018) propose to measure the Modularity as the average normalized squared difference of the mutual information of the factor of variations with the highest and second-highest mutual information with a dimension of the learned representation. A representation is explicit if it is easy to predict a factor of variation. To compute the explicitness, they train a one-versus-rest logistic regression classifier to predict the ground-truth factor of variation and measeure its ROC-AUC. In the current implementation, observations are discretized into n_bins bins.

SAP - Separated Attribute Predictability According to Kumar et al. (2018), the Separated Attribute Predictability (SAP) score is computed from a score matrix where each entry is the linear regression or classification score (in case of discrete factors) of predicting a given ground-truth factors with a given dimension of the learned representation. The (SAP) score is the average difference of the prediction error of the two most predictive learned dimensions for each factor. As done in (Locatello et al., 2019), we use a linear SVM as classifier.

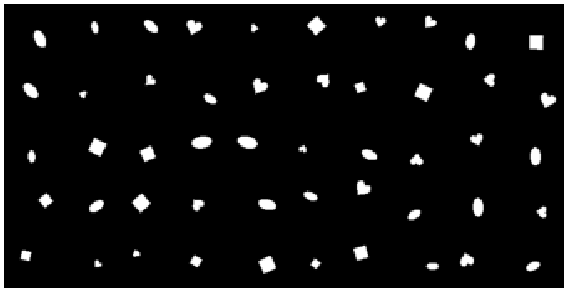
As explained in the main paper, the implementation of the selected disentanglement evaluation metrics is based on Locatello et al. (2019). We report the main parameters in table 5.

Disentanglement metrics	Parameters
Beta score	train_size=10'000, test_size=5'000, batch_size=64, predictor=logistic_regression
MIG	train_size=10'000, n_bins=20
Modularity and Explicitness	train_size=10'000, test_size=5'000, batch_size=16, n_bins=20
SAP score	train_size=10'000, test_size=5'000, batch_size=16, predictor=linearSVM, C=0.01

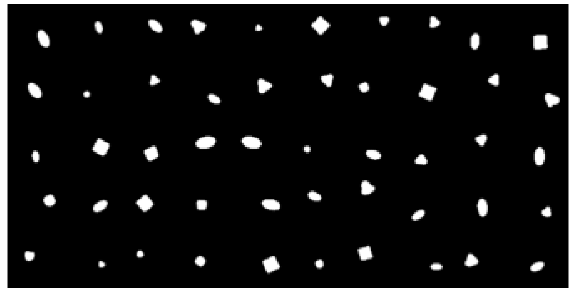
Table 5. Disentanglement metrics and their parameters.

F. Full experiments

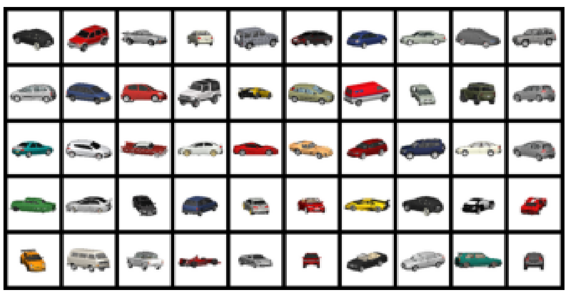
In this section, we report the full set of experiments, including reconstructions and latent traversals.



(a) DSPRITES: original observations.

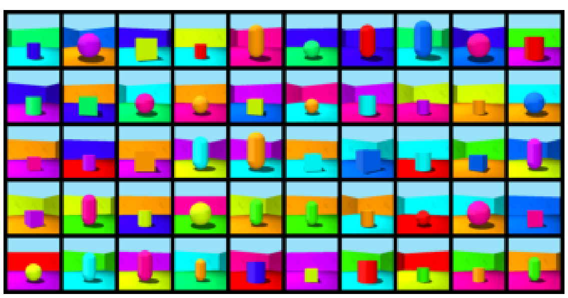


(b) DSPRITES: reconstructions by IDVAE.

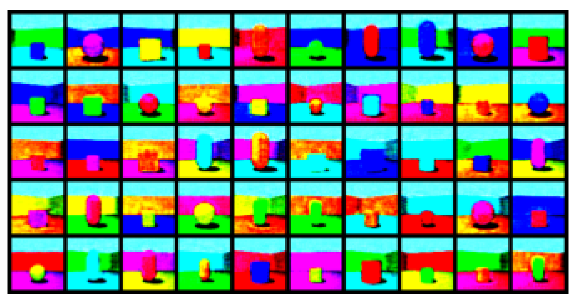


(c) CARS3D: original observations.

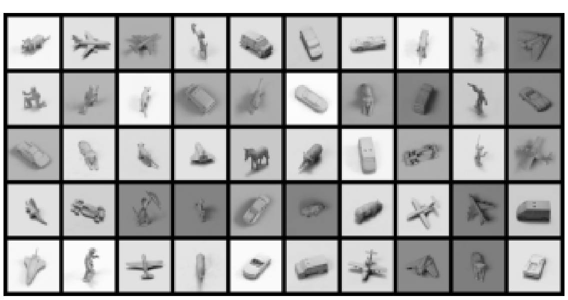
(d) CARS3D: reconstructions by IDVAE.



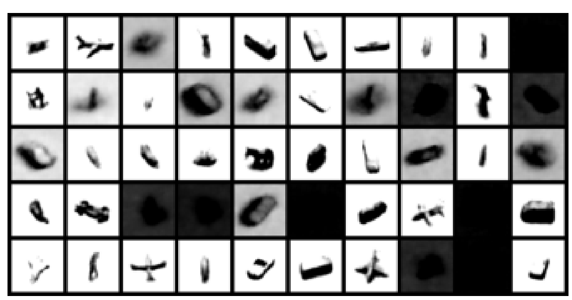
(e) SHAPES3D: original observations.



(f) SHAPES3D: reconstructions by IDVAE.

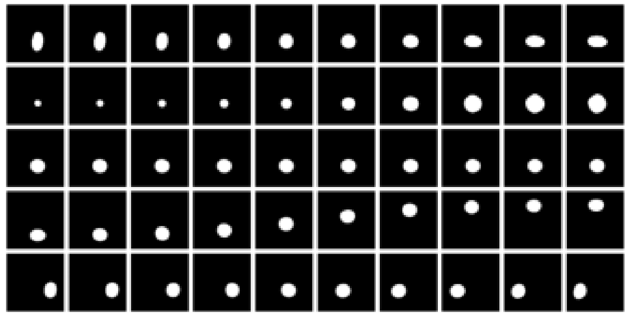


(g) SMALLNORB: original observations.

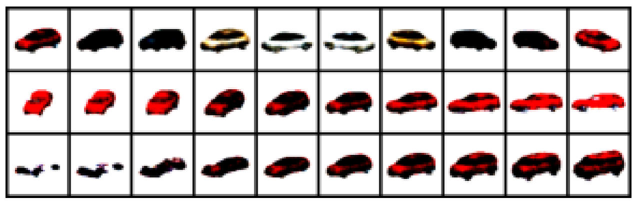


(h) SMALLNORB: reconstructions by IDVAE.

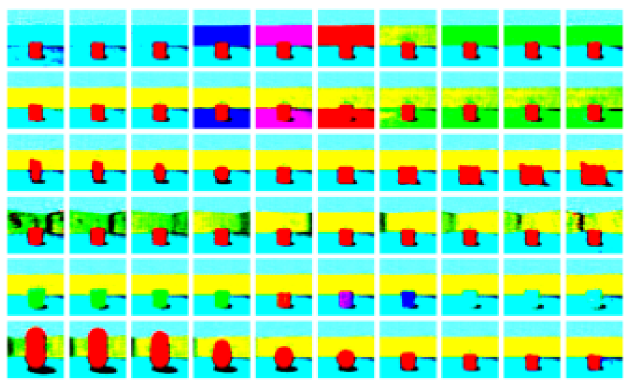
Figure 1. Original observations vs IDVAE reconstructions.



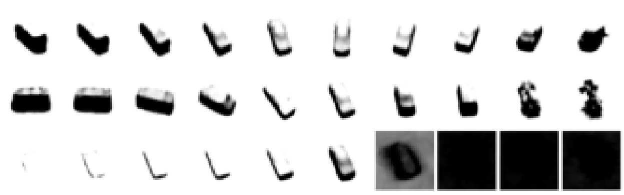
(a) DSPRITES.



(b) CARS3D.



(c) SHAPES3D.



(d) SMALLNORB.

Figure 2. IDVAE latent traversals. Each row corresponds to a dimension of $\mathbf{z}$, that we vary in the range $[-3, 3]$. We can see that, in some cases, changing a dimension can affect multiple ground-truth factors, meaning that IDVAE has not obtained full disentanglement. (a) From top to bottom: orientation, scale, shape(?), posY, posX. (b) From top to bottom: azimuth, elevation, object type. (c) From top to bottom: wall color, floor color, object type, azimuth, object color, object size. (d) azimuth, elevation, lighting, category.

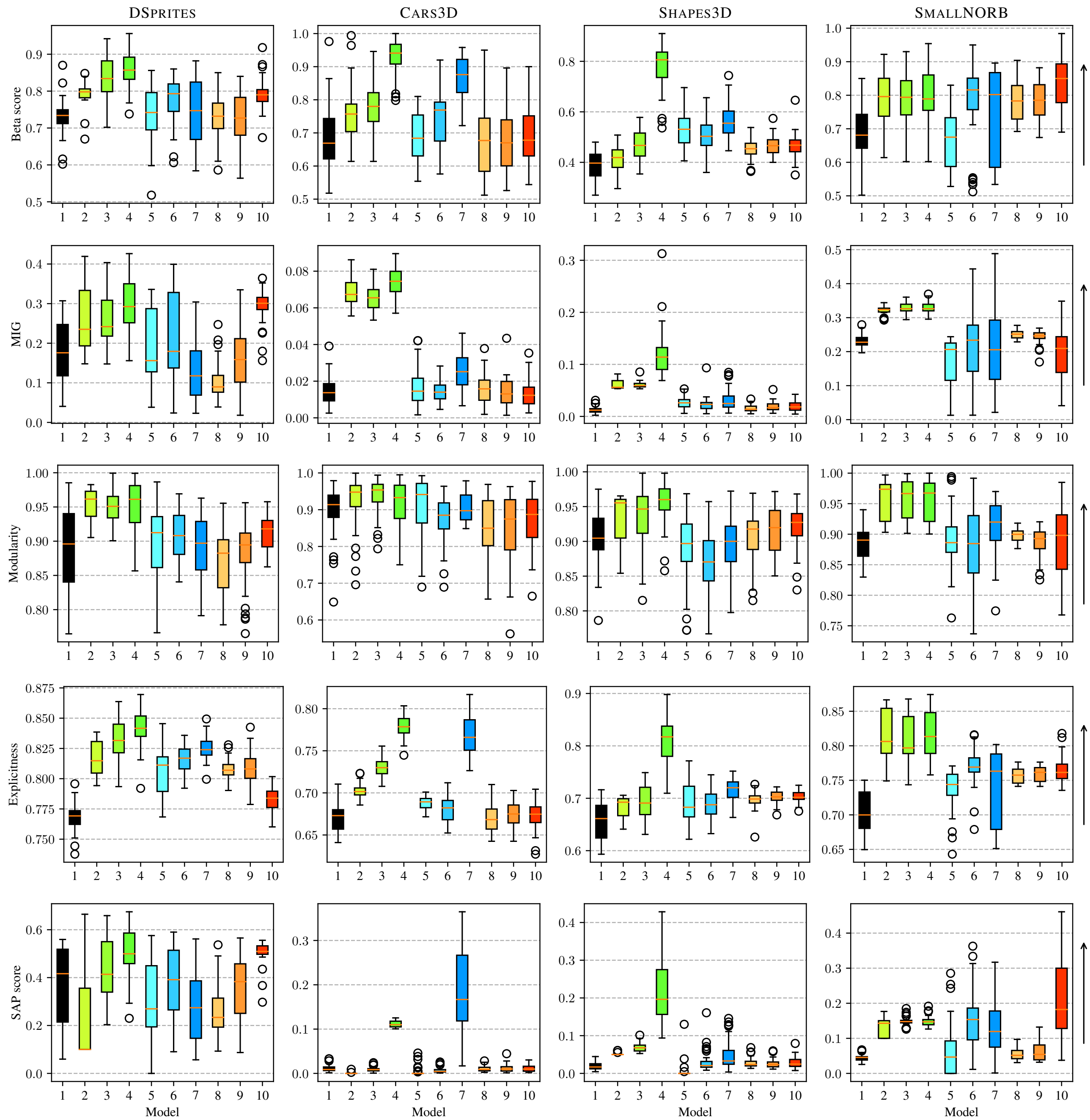


Figure 3. Beta score, MIG, Modularity, Explicitness, and SAP (the higher the better). 1=β-VAE, 2=SS-IDVAE (1%), 3=SS-IDVAE (10%), 4=IDVAE, 5=SS-IVAE (1%), 6=SS-IVAE (10%), 7=IVAE, 8=SS-FULLVAE (1%), 9=SS-FULLVAE (10%), 10=FULLVAE. Percentage of labeled samples in parenthesis.

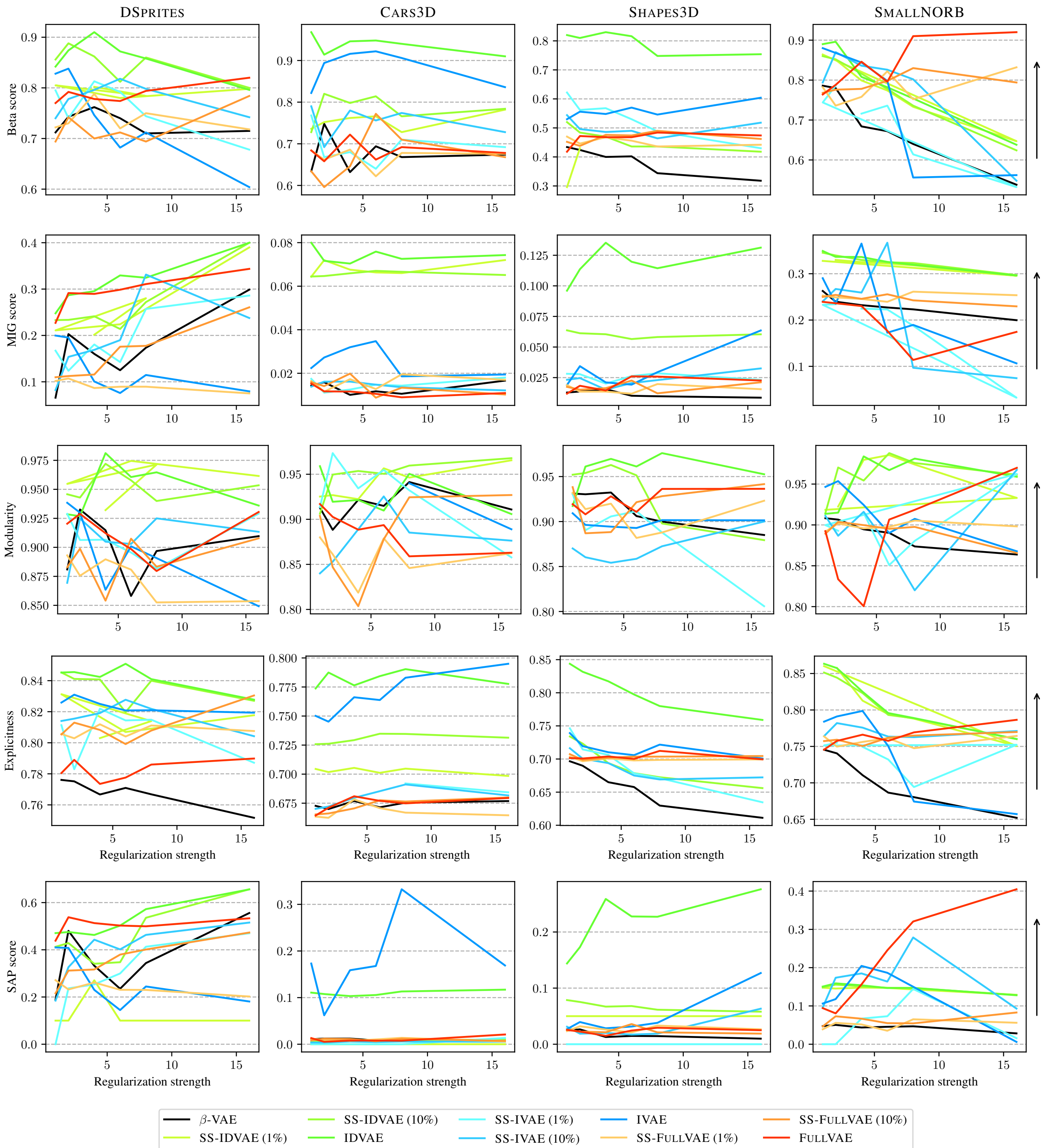


Figure 4. Beta score, MIG, modularity, explicitness and SAP median (the higher the better) as a function of the regularization strength, for each method on DSprites, Cars3D, Shapes3D, SmallNORB.

		DSprites	Cars3D	Shapes3D	SmallNORB
β-VAE	median	0.73	0.67	0.40	0.68
	mean	0.73	0.68	0.39	0.69
	stdev	0.06	0.10	0.06	0.09
SS-IDVAE (1%)	median	0.80	0.76	0.42	0.80
	mean	0.79	0.76	0.41	0.79
	stdev	0.04	0.08	0.08	0.08
SS-IDVAE (10%)	median	0.83	0.78	0.47	0.79
	mean	0.84	0.78	0.47	0.78
	stdev	0.06	0.07	0.06	0.09
IDVAE	median	0.86	0.94	0.81	0.79
	mean	0.86	0.93	0.78	0.79
	stdev	0.05	0.06	0.09	0.09
SS-IVAE (1%)	median	0.74	0.68	0.53	0.68
	mean	0.74	0.69	0.53	0.66
	stdev	0.07	0.07	0.07	0.09
SS-IVAE (10%)	median	0.79	0.77	0.50	0.82
	mean	0.78	0.74	0.51	0.78
	stdev	0.06	0.08	0.07	0.11
IVAE	median	0.75	0.88	0.56	0.8
	mean	0.74	0.87	0.56	0.74
	stdev	0.09	0.06	0.07	0.14
SS-FullVAE (1%)	median	0.73	0.68	0.46	0.78
	mean	0.73	0.67	0.46	0.78
	stdev	0.06	0.12	0.05	0.05
SS-FullVAE (10%)	median	0.72	0.67	0.47	0.79
	mean	0.73	0.67	0.47	0.79
	stdev	0.07	0.10	0.04	0.05
FullVAE	median	0.79	0.68	0.47	0.85
	mean	0.79	0.70	0.47	0.84
	stdev	0.04	0.09	0.05	0.07

Table 6. Beta score median, mean and standard deviation (stdev) for all the tested methods and datasets (the higher the better).

		DSPRITES	CARS3D	SHAPES3D	SMALLNORB
β-VAE	median	0.18	0.01	0.01	0.23
	mean	0.18	0.01	0.01	0.23
	stdev	0.08	0.01	0.01	0.02
SS-IDVAE (1%)	median	0.24	0.07	0.06	0.32
	mean	0.26	0.07	0.06	0.32
	stdev	0.09	0.01	0.01	0.01
SS-IDVAE (10%)	median	0.24	0.07	0.06	0.33
	mean	0.27	0.07	0.06	0.33
	stdev	0.07	0.01	0.01	0.02
IDVAE	median	0.29	0.07	0.11	0.33
	mean	0.30	0.07	0.12	0.33
	stdev	0.07	0.01	0.04	0.02
SS-IVAE (1%)	median	0.16	0.01	0.03	0.21
	mean	0.19	0.02	0.03	0.17
	stdev	0.09	0.01	0.01	0.08
SS-IVAE (10%)	median	0.18	0.01	0.02	0.23
	mean	0.21	0.01	0.02	0.22
	stdev	0.10	0.01	0.01	0.11
IVAE	median	0.12	0.03	0.02	0.21
	mean	0.14	0.03	0.03	0.22
	stdev	0.08	0.01	0.02	0.12
SS-FULLVAE (1%)	median	0.08	0.01	0.01	0.25
	mean	0.09	0.01	0.01	0.25
	stdev	0.05	0.01	0.01	0.02
SS-FULLVAE (10%)	median	0.16	0.01	0.02	0.25
	mean	0.16	0.01	0.02	0.25
	stdev	0.09	0.01	0.01	0.02
FULLVAE	median	0.30	0.01	0.02	0.21
	mean	0.29	0.01	0.02	0.20
	stdev	0.04	0.01	0.01	0.07

Table 7. MIG median, mean and standard deviation (stdev) for all the tested methods and datasets (the higher the better).

		DSPRITES	CARS3D	SHAPES3D	SMALLNORB
β-VAE	median	0.90	0.91	0.90	0.89
	mean	0.89	0.90	0.91	0.88
	stdev	0.06	0.07	0.03	0.03
SS-IDVAE (1%)	median	0.95	0.95	0.96	0.97
	mean	0.95	0.92	0.92	0.96
	stdev	0.02	0.07	0.05	0.03
SS-IDVAE (10%)	median	0.95	0.95	0.95	0.97
	mean	0.95	0.94	0.93	0.96
	stdev	0.03	0.05	0.04	0.03
IDVAE	median	0.96	0.93	0.96	0.97
	mean	0.95	0.91	0.95	0.96
	stdev	0.03	0.07	0.03	0.03
SS-IVAE (1%)	median	0.91	0.94	0.90	0.89
	mean	0.90	0.92	0.89	0.89
	stdev	0.06	0.07	0.05	0.05
SS-IVAE (10%)	median	0.91	0.89	0.87	0.88
	mean	0.91	0.87	0.87	0.88
	stdev	0.03	0.06	0.04	0.07
IVAE	median	0.90	0.90	0.90	0.92
	mean	0.89	0.91	0.90	0.91
	stdev	0.04	0.04	0.04	0.04
SS-FULLVAE (1%)	median	0.88	0.84	0.92	0.90
	mean	0.87	0.85	0.91	0.90
	stdev	0.04	0.08	0.05	0.01
SS-FULLVAE (10%)	median	0.89	0.88	0.92	0.89
	mean	0.88	0.87	0.92	0.89
	stdev	0.03	0.08	0.04	0.03
FULLVAE	median	0.92	0.89	0.93	0.90
	mean	0.91	0.87	0.92	0.89
	stdev	0.02	0.07	0.03	0.06

Table 8. Modularity median, mean and standard deviation (stdev) for all the tested methods and datasets (the higher the better).

		DSPRITES	CARS3D	SHAPES3D	SMALLNORB
β-VAE	median	0.77	0.67	0.66	0.70
	mean	0.77	0.67	0.66	0.70
	stdev	0.01	0.02	0.03	0.03
SS-IDVAE (1%)	median	0.81	0.70	0.69	0.81
	mean	0.82	0.70	0.68	0.81
	stdev	0.02	0.01	0.03	0.04
SS-IDVAE (10%)	median	0.83	0.73	0.69	0.80
	mean	0.83	0.73	0.69	0.81
	stdev	0.02	0.01	0.03	0.04
IDVAE	median	0.84	0.78	0.82	0.81
	mean	0.84	0.78	0.81	0.81
	stdev	0.01	0.01	0.04	0.04
SS-IVAE (1%)	median	0.81	0.69	0.68	0.74
	mean	0.81	0.69	0.69	0.74
	stdev	0.02	0.01	0.04	0.03
SS-IVAE (10%)	median	0.82	0.68	0.69	0.77
	mean	0.82	0.68	0.69	0.77
	stdev	0.01	0.01	0.02	0.03
IVAE	median	0.82	0.77	0.72	0.76
	mean	0.83	0.77	0.72	0.74
	stdev	0.01	0.02	0.02	0.06
SS-FULLVAE (1%)	median	0.80	0.66	0.70	0.76
	mean	0.80	0.66	0.70	0.76
	stdev	0.01	0.02	0.01	0.01
SS-FULLVAE (10%)	median	0.80	0.67	0.70	0.76
	mean	0.80	0.67	0.70	0.76
	stdev	0.02	0.02	0.01	0.02
FULLVAE	median	0.78	0.67	0.70	0.76
	mean	0.78	0.67	0.70	0.76
	stdev	0.01	0.02	0.01	0.02

Table 9. Explicitness median, mean and standard deviation (stdev) for all the tested methods and datasets (the higher the better).

		DSPRITES	CARS3D	SHAPES3D	SMALLNORB
β-VAE	median	0.42	0.01	0.02	0.05
	mean	0.37	0.01	0.02	0.04
	stdev	0.16	0.01	0.01	0.01
SS-IDVAE (1%)	median	0.10	0.00	0.05	0.14
	mean	0.23	0.00	0.05	0.13
	stdev	0.19	0.00	0.00	0.02
SS-IDVAE (10%)	median	0.41	0.01	0.07	0.15
	mean	0.46	0.01	0.07	0.15
	stdev	0.13	0.01	0.01	0.01
IDVAE	median	0.50	0.11	0.20	0.15
	mean	0.51	0.11	0.22	0.15
	stdev	0.11	0.01	0.09	0.01
SS-IVAE (1%)	median	0.27	0.00	0.00	0.05
	mean	0.29	0.00	0.00	0.06
	stdev	0.19	0.01	0.02	0.07
SS-IVAE (10%)	median	0.39	0.00	0.02	0.15
	mean	0.39	0.01	0.03	0.16
	stdev	0.14	0.01	0.03	0.09
IVAE	median	0.27	0.17	0.03	0.12
	mean	0.28	0.19	0.05	0.12
	stdev	0.14	0.10	0.04	0.08
SS-FULLVAE (1%)	median	0.25	0.01	0.02	0.05
	mean	0.26	0.01	0.02	0.05
	stdev	0.11	0.01	0.01	0.02
SS-FULLVAE (10%)	median	0.39	0.01	0.02	0.05
	mean	0.38	0.01	0.02	0.06
	stdev	0.13	0.01	0.01	0.03
FULLVAE	median	0.51	0.01	0.02	0.18
	mean	0.50	0.01	0.03	0.20
	stdev	0.06	0.01	0.01	0.11

Table 10. SAP score median, mean and standard deviation (stdev) for all the tested methods and datasets (the higher the better).

REFERENCES

T. Q. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Proc. of the 31st Int. Conf. on Neural Inf. Proc. Sys.*, NeurIPS, 2018.

I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *Proc. of the 5th Int. Conf. on Learn. Repr.*, ICLR, 2017.

I. Khemakhem, D. P. Kingma., R. P. Mont, and A. Hyvärinen. Variational autoencoders and nonlinear ica: A unifying framework. In *Proc. of the 23rd Int. Conf. on Artif. Intel. and Stat.*, AISTATS, 2020.

A. Kumar, P. Sattigeri, and A. Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *Proc. of the 6th Int. Conf. on Learn. Repr.*, ICLR, 2018.

F. Locatello, S. Bauer, M. Lucic, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proc. of the 36th Int. Conf. on Mach. Learn.*, ICML, 2019.

K. Ridgeway and M. C. Mozer. Learning deep disentangled embeddings with the f-statistic loss. In *Proc. of the 31st Int. Conf. on Neural Inf. Proc. Sys.*, NeurIPS, 2018.