Supplementary matarial (github)

1 Latent space of trained models

In this section, we consider the properties of classifier's latent space for both the hand-crafted and learnable priors under different amount of training samples. Tables 1 and 2 show t-sne plots for the perflexion 30 for 100, 1000 and 60000 ("all") training labels of the MNIST dataset.

The first raw of Table 1 with the label "$\mathcal{D}_{c\hat{c}}$" corresponds to the classifier considered in section 2.1.1 of the Complementary materials. The latent space $\mathbf{a}$ of the fully supervised classifier with "all" labels demonstrates the perfect separability of classes. The classes are far away from each other and there are practically no outliers leading to the misclassification. The decrease of the number of labels in the supervised setup, see the columns 1000 and 100, leads to a visible degradation of separability between the classes.

The regularization of class label space by the regularizer $\mathcal{D}_c$ or by the hand-crafted latent space regularizer $\mathcal{D}_a$ shown in raws "$\mathcal{D}_{c\hat{c}} + \alpha_c \mathcal{D}_c$" considered in section 2.1.2 and "$\mathcal{D}_{c\hat{c}} + \alpha_a \mathcal{D}_a$" considered in section 2.1.3 for the small number of training samples equal 100 does not significantly enhance the class separability with respect to "$\mathcal{D}_{c\hat{c}}$".

At the same time, the joint usage of the above regularizers according to the model "$\mathcal{D}_{c\hat{c}} + \alpha_c \mathcal{D}_c + \alpha_a \mathcal{D}_a$" according to the model 2.1.4 leads to the better separability of classes for 100 labels in comparison with the previous cases. At the same time, the addition of these regularizers does not have any impact on the latent space for "all" label case.

The introduction of learnable regularization of latent space along with the class label regularization according to the model"$\mathcal{D}_{c\hat{c}} + \mathcal{D}_c + \mathcal{D}_z + \mathcal{D}_{x\hat{x}} + \alpha_x \mathcal{D}_x$" considered in section 2.2.2 enhances the class separability in the latent space of classifier for 100 label case that is also very close to the fully supervised case.

For the comparison reasons, we also visualize the latent space of the auto-encoder for the above model in Table 2.

2 Implementation details

In this section, we present the implementation details for each considered architecture.

2.1 Classification based on hand-crafted priors

2.1.1 Supervised training without latent space regularization (baseline)

The baseline architecture is based on the cross-entropy term $\mathcal{D}_{c\hat{c}}$ (6) in the main part of paper and depicted in Figure 1:

$$\mathcal{L}^{\mathrm{HCP}}_{\mathrm{S-NoReg}}(\boldsymbol{\theta}_c, \boldsymbol{\phi}_a) = \mathcal{D}_{c\hat{c}}. \tag{18}$$

The parameters of encoder and decoder are shown in Table 3.

The performance of semi-supervised classifier with and without batch normalization is shown in Table 5 and corresponds to the parameter $\alpha_c = 0$.

2.1.2 Semi-supervised training without latent space regularization

This model is based on terms $\mathcal{D}_{c\hat{c}}$ and $\mathcal{D}_c$ in (7) in the main part of paper and schematically shown in Figure 2:

$$\mathcal{L}^{\mathrm{HCP}}_{\mathrm{SS-NoReg}}(\boldsymbol{\theta}_c, \boldsymbol{\phi}_a) = \mathcal{D}_{c\hat{c}} + \alpha_c \mathcal{D}_c. \tag{19}$$

The parameters of encoder, decoder and discriminator are shown in Table 4. The KL-divergence term $\mathcal{D}_c$ is implemented in a form of density ratio estimator (DRE). In the considered practical implementation, the parameter α_c controls the trade-off between the cross-entropy and class discriminator terms. The discriminator $\mathcal{D}_c$ is trained in an adversarial way based on samples generated by the decoder and from targeted distribution.

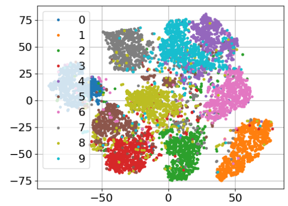
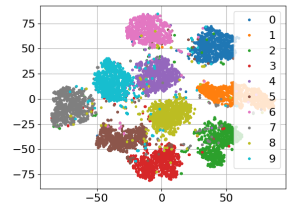
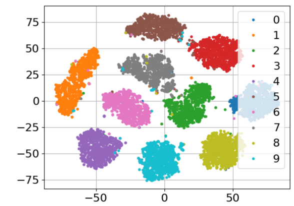
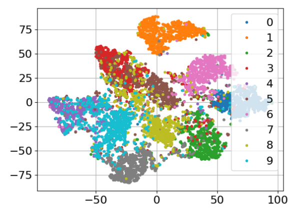
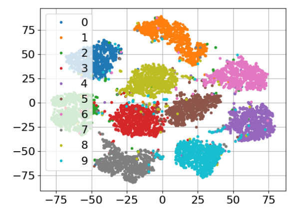
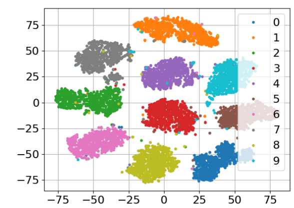
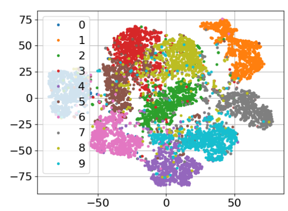
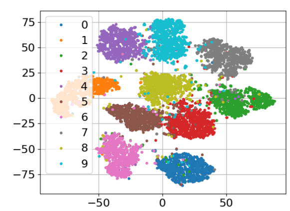
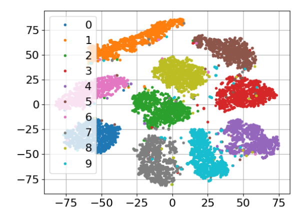
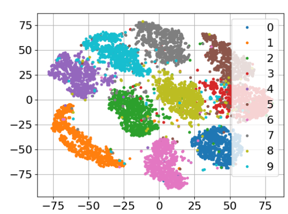
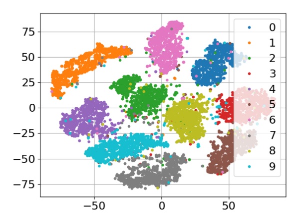
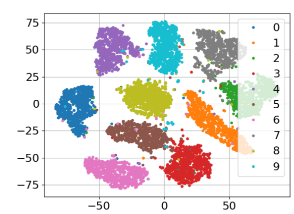
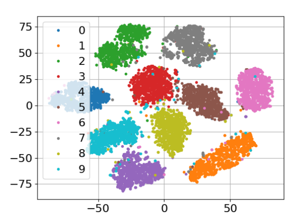
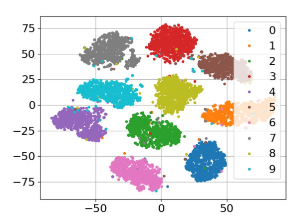
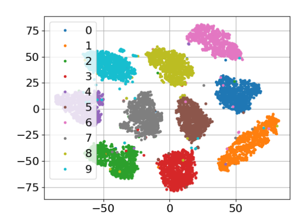
	MNIST (100)	**MIST** (1000)	**MNIST** (all)
$\mathcal{D}_{\mathrm{c}\hat{\mathrm{c}}}$			
$\mathcal{D}_{\mathrm{c}\hat{\mathrm{c}}} + \alpha_{\mathrm{c}}\mathcal{D}_{\mathrm{c}}$			
$\mathcal{D}_{\mathrm{c}\hat{\mathrm{c}}} + \alpha_{\mathrm{a}}\mathcal{D}_{\mathrm{a}}$			
$\mathcal{D}_{\mathrm{c}\hat{\mathrm{c}}} + \alpha_{\mathrm{a}}\mathcal{D}_{\mathrm{a}} + \alpha_{\mathrm{c}}\mathcal{D}_{\mathrm{c}}$			
$\mathcal{D}_{\mathrm{c}\hat{\mathrm{c}}} + \mathcal{D}_{\mathrm{c}} + \mathcal{D}_{\mathrm{z}} + \mathcal{D}_{\mathrm{x}\hat{\mathrm{x}}} + \alpha_{\mathrm{x}}\mathcal{D}_{\mathrm{x}}$			

Table 1: Latent space $\mathbf{a}$ of classifier.

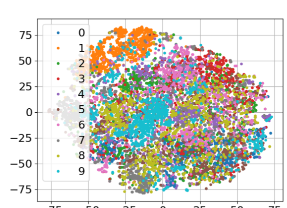
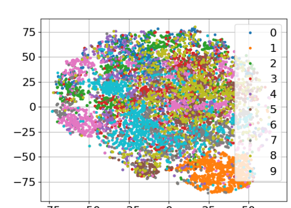
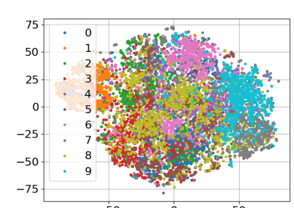
	MNIST (100)	**MIST** (1000)	**MNIST** (all)
$\mathcal{D}_{\mathrm{c}\hat{\mathrm{c}}} + \mathcal{D}_{\mathrm{c}} + \mathcal{D}_{\mathrm{z}} + \mathcal{D}_{\mathrm{x}\hat{\mathrm{x}}} + \alpha_{\mathrm{x}}\mathcal{D}_{\mathrm{x}}$			

Table 2: Latent space $\mathbf{z}$ of auto-encoder.

The performance of semi-supervised classifier with and without batch normalization is shown in Table 5.

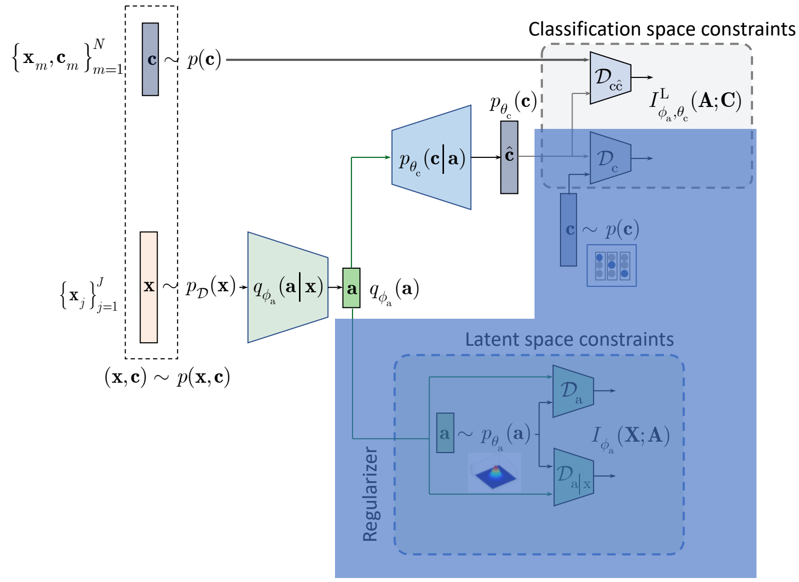


Figure 1: Baseline classifier based on $\mathcal{D}_{c\hat{c}}$. The blue shadowed regions are not used.

Encoder	
Size	Layer
$28 \times 28 \times 1$	Input
$14 \times 14 \times 32$	Conv2D, LeakyReLU
$7 \times 7 \times 64$	Conv2D, LeakyReLU
$4 \times 4 \times 128$	Conv2D, LeakyReLU
2048	Flatten
1024	FC

Decoder	
Size	Layer
1024	Input
500	FC, ReLU
10	FC, Softmax

Table 3: The network parameters of baseline classifier trained on $\mathcal{D}_{c\hat{c}}$. The encoder is trained with and without batch normalization (BN) after Conv2D layers.

Encoder	
Size	Layer
$28 \times 28 \times 1$	Input
$14 \times 14 \times 32$	Conv2D, LeakyReLU
$7 \times 7 \times 64$	Conv2D, LeakyReLU
$4 \times 4 \times 128$	Conv2D, LeakyReLU
2048	Flatten
1024	FC, ReLU
500	FC, ReLU
10	FC, Softmax

Decoder	
Size	Layer
1024	Input
500	FC, ReLU
10	FC, Softmax

$\mathcal{D}_c$	
Size	Layer
10	Input
500	FC, ReLU
500	FC, ReLU
1	FC, Sigmoid

Table 4: The network parameters of semi-supervised classifier are trained on $\mathcal{D}_{c\hat{c}}$ and $\mathcal{D}_c$. The encoder is trained with and without batch normalization (BN) after Conv2D layers.

2.1.3 Supervised training with latent space regularization

This model is based on the cross-entropy term $\mathcal{D}_{c\hat{c}}$ and either term $\mathcal{D}_{a|x}$ or $\mathcal{D}_a$ or jointly $\mathcal{D}_{a|x}$ and $\mathcal{D}_a$ as defined by (8) in the main part of paper. In our implementation, we consider the regularization based on the adversarial term $\mathcal{D}_a$ similar to AAE due to the flexibility of imposing different priors on the latent space distribution. The implemented system is shown in Figure 3 and the training is based on:

$$\mathcal{L}^{\text{HCP}}_{\text{S-Reg}}(\boldsymbol{\theta}_c, \boldsymbol{\phi}_a) = \mathcal{D}_{c\hat{c}} + \alpha_a \mathcal{D}_a, \tag{20}$$

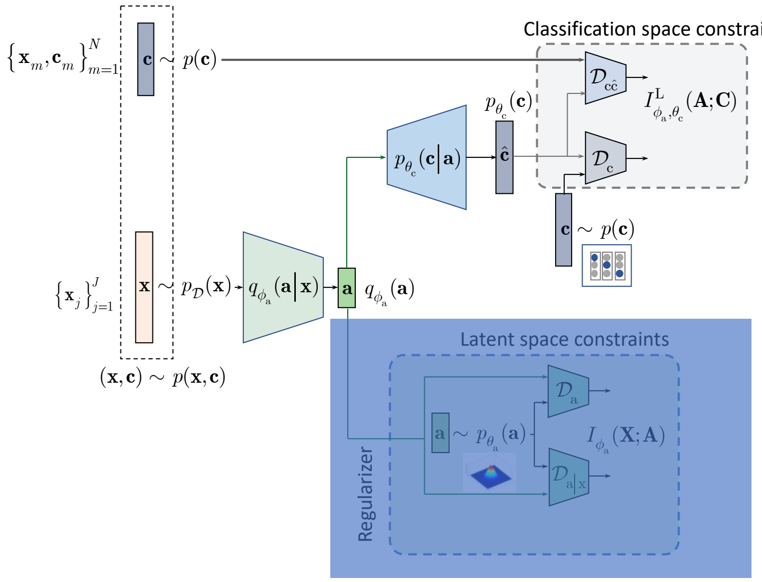


Figure 2: Adversarial semi-supervised classifier based on the cross-entropy $\mathcal{D}_{\mathrm{c}\hat{\mathrm{c}}}$ and categorical class discriminator $\mathcal{D}_{\mathrm{c}}$. No latent space regularization is applied. The blue shadowed regions are not used.

Encoder model	α_{c}	runs 1	runs 2	runs 3	mean	std
		MNIST 100				
without BN	0	26.56	26.41	28.04	26.95	0.96
	0.005	20.44	21.93	18.98	20.45	1.48
	0.0005	18.55	20.43	20.59	**19.86**	1.14
	1	19.23	22.42	20.57	20.74	1.60
with BN	0	29.37	29.27	30.62	29.75	0.75
	0.005	27.97	28.02	26.27	27.42	1.00
	0.0005	29.99	23.70	24.47	**24.72**	1.17
	1	27.78	31.98	35.88	31.88	4.05
		MNIST 1000				
without BN	0	7.74	6.99	6.97	7.23	0.44
	0.005	5.62	6.06	5.60	**5.76**	0.26
	0.0005	6.30	6.12	6.02	6.15	0.14
	1	5.99	6.27	6.28	6.18	0.16
with BN	0	7.45	6.95	7.52	7.31	0.31
	0.005	5.57	5.08	5.22	**5.29**	0.25
	0.0005	5.60	6.05	6.22	5.96	0.32
	1	6.05	6.41	5.82	6.09	0.30
		MNIST all				
without BN	0	0.83	0.83	0.74	**0.80**	0.05
	0.005	0.83	0.82	0.88	0.84	0.03
	0.0005	0.86	0.92	0.82	0.87	0.05
	1	0.72	0.85	0.87	0.81	0.08
with BN	0	0.73	0.67	0.79	0.73	0.06
	0.005	0.72	0.73	0.70	0.72	0.02
	0.0005	0.75	0.77	0.72	0.75	0.03
	1	0.67	0.68	0.73	**0.69**	0.03

Table 5: The performance of classifier based on $\mathcal{D}_{\mathrm{c}\hat{\mathrm{c}}} + \alpha_{\mathrm{c}}\mathcal{D}_{\mathrm{c}}$ for the encoder with and without batch normalization as a function of Lagrangian multiplier α_{c} and the number of labelled examples.

where α_{a} is a regularization parameter controlling a trade-off between the cross-entropy term and latent space regularization term. We have replaced the Lagrangians above with respect to (8) in the main part of paper and used it in front of $\mathcal{D}_{\mathrm{a}}$ in contrast to the original formulation (8). It is done to keep the term $\mathcal{D}_{\mathrm{c}\hat{\mathrm{c}}}$ without a multiplier as the reference to the baseline classifier.

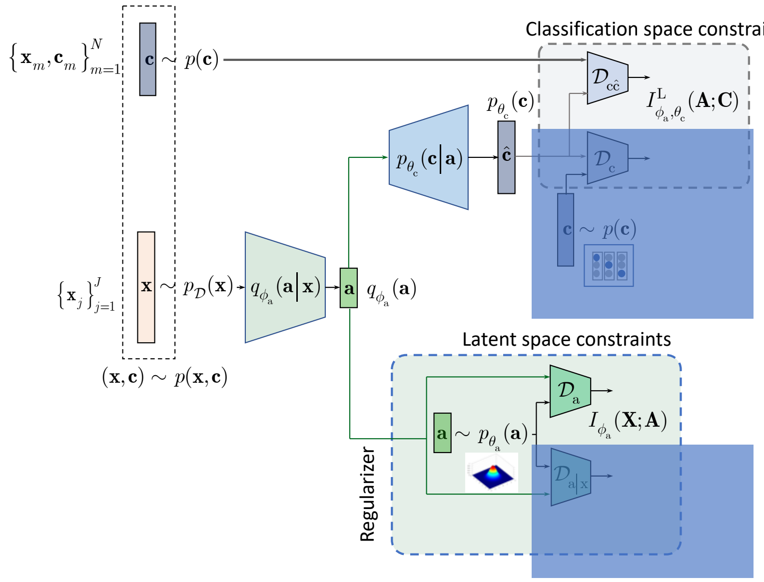


Figure 3: Supervised classifier based on the cross-entropy $\mathcal{D}_{\mathrm{c}\hat{\mathrm{c}}}$ and latent space regularization $\mathcal{D}_{\mathrm{a}}$. The blue shadowed parts are not used.

Encoder	
Size	Layer
$28 \times 28 \times 1$	Input
$14 \times 14 \times 32$	Conv2D, LeakyReLU
$7 \times 7 \times 64$	Conv2D, LeakyReLU
$4 \times 4 \times 128$	Conv2D, LeakyReLU
2048	Flatten
1024	FC, ReLU
500	FC, ReLU
10	FC, Softmax

Decoder	
Size	Layer
1024	Input
500	FC, ReLU
10	FC, Softmax

$\mathcal{D}_{\mathrm{a}}$	
Size	Layer
1024	Input
500	FC, ReLU
500	FC, ReLU
1	FC, Sigmoid

Table 6: The network parameters of supervised classifier are trained on $\mathcal{D}_{\mathrm{c}\hat{\mathrm{c}}}$ and $\mathcal{D}_{\mathrm{a}}$. The encoder is trained with and without batch normalization (BN) after Conv2D layers. $\mathcal{D}_{\mathrm{a}}$ is trained in the adversarial way.

The parameters of encoder, decoder and discriminator are summarized in Table 6. The performance of this classifier without and with batch normalization is shown in Table 7.

2.1.4 Semi-supervised training with latent space regularization

This model is based on the cross-entropy term $\mathcal{D}_{\mathrm{c}\hat{\mathrm{c}}}$ and either term $\mathcal{D}_{\mathrm{a|x}}$ or $\mathcal{D}_{\mathrm{a}}$ or jointly $\mathcal{D}_{\mathrm{a|x}}$ and $\mathcal{D}_{\mathrm{a}}$ and the label class regularizer $\mathcal{D}_{\mathrm{c}}$ as defined by (9) in the main part of paper. In our implementation, we consider the regularization based on the adversarial term $\mathcal{D}_{\mathrm{a}}$ only as shown in Figure 4. The training is based on:

$$\mathcal{L}^{\mathrm{HCP}}_{\mathrm{S-Reg}}(\boldsymbol{\theta}_{\mathrm{c}}, \boldsymbol{\phi}_{\mathrm{a}}) = \mathcal{D}_{\mathrm{c}\hat{\mathrm{c}}} + \alpha_{\mathrm{c}}\mathcal{D}_{\mathrm{c}} + \alpha_{\mathrm{a}}\mathcal{D}_{\mathrm{a}}. \tag{21}$$

The parameters of encoder, decoder and both discriminators are shown in Table 8.

The performance of this classifier without and with batch normalization is shown in Table 9.

Encoder model	α_a	runs 1	2	3	mean	std
		MNIST 100				
without BN	0	26.79	27.26	27.39	**27.15**	0.32
	0.005	28.05	25.95	30.72	28.24	2.39
	0.0005	26.67	27.69	28.46	27.61	0.89
	1	33.42	33.05	34.81	33.76	0.92
with BN	0	30.37	29.32	29.82	**29.83**	0.52
	0.005	28.02	31.49	30.80	30.11	1.84
	0.0005	34.54	31.92	29.82	31.09	2.36
	1	34.43	44.35	44.25	41.01	5.70
		MNIST 1000				
without BN	0	7.16	8.12	7.55	7.61	0.48
	0.005	7.02	6.34	6.59	6.65	0.34
	0.0005	6.73	6.34	6.82	**6.63**	0.26
	1	9.49	9.93	10.56	9.99	0.54
with BN	0	7.39	7.83	7.92	**7.72**	0.28
	0.005	7.94	7.15	8.53	7.88	0.69
	0.0005	8.00	9.62	9.51	9.05	0.91
	1	15.79	14.88	13.71	14.79	1.04
		MNIST all				
without BN	0	0.76	0.70	0.81	**0.76**	0.06
	0.005	1.07	1.03	1.13	1.08	0.05
	0.0005	0.84	0.78	0.89	0.84	0.06
	1	4.78	7.24	4.71	5.58	1.44
with BN	0	0.68	0.68	0.69	**0.68**	0.01
	0.005	0.90	0.81	1.12	0.94	0.16
	0.0005	0.87	0.80	0.89	0.85	0.05
	1	2.37	3.61	4.35	3.44	1.00

Table 7: . The performance of classifier based on $\mathcal{D}_{c\hat{c}} + \alpha_a \mathcal{D}_a$ for the encoder with and without batch normalization as a function of Lagrangian multiplier.

Encoder	
Size	Layer
$28 \times 28 \times 1$	Input
$14 \times 14 \times 32$	Conv2D, LeakyReLU
$7 \times 7 \times 64$	Conv2D, LeakyReLU
$4 \times 4 \times 128$	Conv2D, LeakyReLU
2048	Flatten
1024	FC, ReLU
500	FC, ReLU
10	FC, Softmax

Decoder	
Size	Layer
1024	Input
500	FC, ReLU
10	FC, Softmax

$\mathcal{D}_c$	
Size	Layer
10	Input
500	FC, ReLU
500	FC, ReLU
1	FC, Sigmoid

$\mathcal{D}_a$	
Size	Layer
1024	Input
500	FC, ReLU
500	FC, ReLU
1	FC, Sigmoid

Table 8: The network parameters of supervised classifier are trained on $\mathcal{D}_{c\hat{c}}$, $\mathcal{D}_a$ and $\mathcal{D}_c$. The encoder is trained with and without batch normalization (BN) after Conv2D layers. $\mathcal{D}_a$ and $\mathcal{D}_c$ are trained in the adversarial way.

2.2 Classification based on learnable priors

2.2.1 Semi-supervised training with latent space regularization

This model is based on the cross-entropy term $\mathcal{D}_{c\hat{c}}$, the MSE term representing $\mathcal{D}_{x\hat{x}}$, the label class regularizer $\mathcal{D}_c$ and either term $\mathcal{D}_{z|x}$ or $\mathcal{D}_z$ or jointly $\mathcal{D}_{z|x}$ and $\mathcal{D}_z$ as defined by (15) in the main part of paper. In our implementation, we consider the regularization of the latent space based on the adversarial term $\mathcal{D}_z$ only to compare it with the vanila AAE as shown in Figure 5. The encoder is also

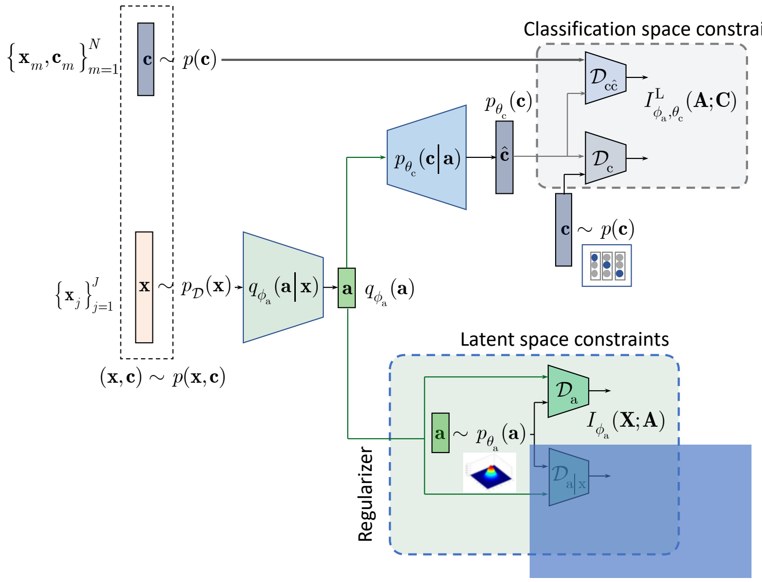


Figure 4: Supervised classifier based on the cross-entropy $\mathcal{D}_{\mathrm{c}\hat{\mathrm{c}}}$ and latent space regularization $\mathcal{D}_{\mathrm{a}}$. The blue shadowed parts are not used.

Encoder model	α_{a}	α_{c}	runs 1	runs 2	runs 3	mean	std
MNIST 100							
without BN	0.005	0.005	21.39	18.12	18.34	19.28	1.83
	0.0005	0.0005	15.33	22.36	13.80	17.16	4.56
	0.005	0.0005	25.66	26.25	28.81	26.91	1.67
	0.0005	0.005	9.82	13.44	13.06	**12.11**	1.99
with BN	0.005	0.005	23.45	21.19	28.87	24.50	3.94
	0.0005	0.0005	28.57	19.06	26.37	24.67	4.98
	0.005	0.0005	26.18	26.18	25.49	25.95	0.40
	0.0005	0.005	8.96	13.82	14.76	**12.52**	3.11
MNIST 1000							
without BN	0.005	0.005	3.91	4.21	3.70	3.94	0.26
	0.0005	0.0005	3.54	3.72	3.54	3.60	0.10
	0.005	0.0005	6.19	5.80	7.31	6.43	0.78
	0.0005	0.005	2.80	2.82	2.83	**2.82**	0.02
with BN	0.005	0.005	3.30	2.94	2.93	3.06	0.21
	0.0005	0.0005	2.80	2.53	2.50	2.61	0.17
	0.005	0.0005	3.51	3.75	4.12	3.79	0.31
	0.0005	0.005	2.58	2.27	2.24	**2.37**	0.19
MNIST all							
without BN	0.005	0.005	1.04	1.07	1.07	1.06	0.02
	0.0005	0.0005	0.86	0.90	0.88	**0.88**	0.02
	0.005	0.0005	1.08	0.92	1.09	1.03	0.10
	0.0005	0.005	0.85	0.93	0.93	0.90	0.05
with BN	0.005	0.005	1.10	1.01	0.93	1.01	0.09
	0.0005	0.0005	0.84	0.88	0.83	0.85	0.03
	0.005	0.0005	1.10	1.12	0.93	1.05	0.10
	0.0005	0.005	0.76	0.82	0.79	**0.79**	0.03

Table 9: The performance of classifier based on $\mathcal{D}_{\mathrm{c}\hat{\mathrm{c}}} + \alpha_{\mathrm{a}}\mathcal{D}_{\mathrm{a}} + \alpha_{\mathrm{c}}\mathcal{D}_{\mathrm{c}}$ for the encoder with and without batch normalization.

not conditioned on $\mathbf{c}$ as in the original semi-supervised AAE. Thus, the tested system is based on:

$$\mathcal{L}^{\mathrm{LP}}_{\mathrm{SS-AAE}}(\boldsymbol{\theta}_{\mathrm{c}}, \boldsymbol{\theta}_{\mathrm{x}}, \boldsymbol{\phi}_{\mathrm{a}}, \boldsymbol{\phi}_{\mathrm{z}}) = \mathcal{D}_{\mathrm{z}} + \beta_{\mathrm{x}}\mathcal{D}_{\mathrm{x}\hat{\mathrm{x}}} + \beta_{\mathrm{c}}\mathcal{D}_{\mathrm{c}\hat{\mathrm{c}}} + \beta_{\mathrm{c}}\mathcal{D}_{\mathrm{c}}. \tag{22}$$

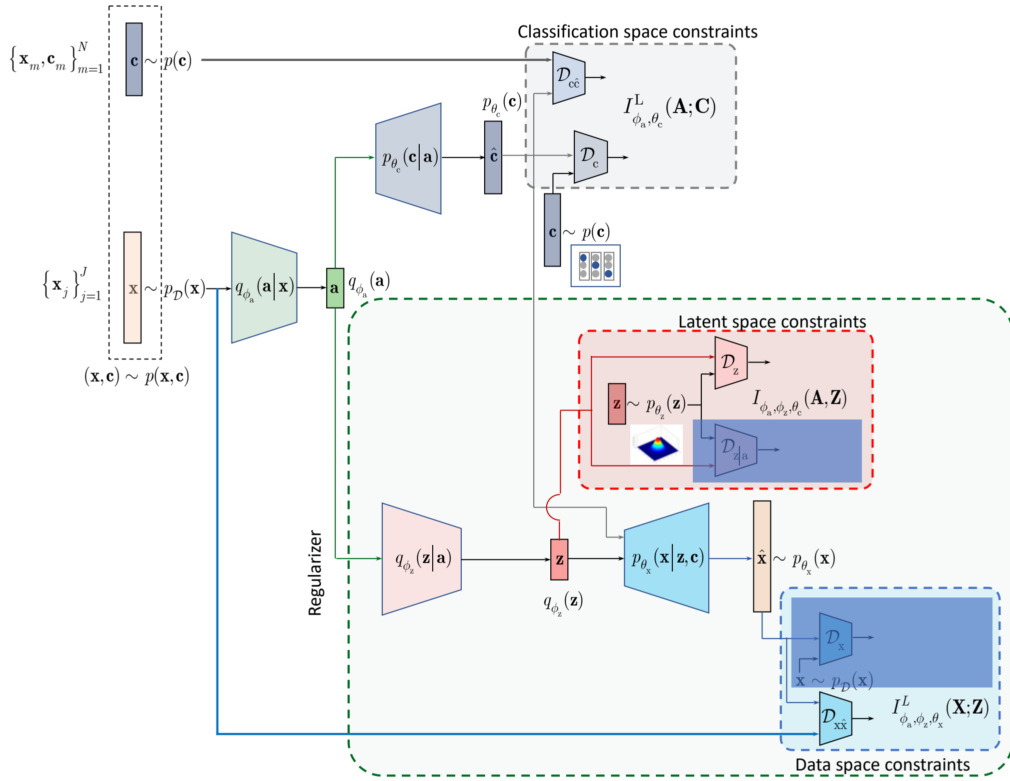


Figure 5: Semi-supervised classifier with learnable priors: the cross-entropy $\mathcal{D}_{\mathrm{c}\hat{\mathrm{c}}}$, MSE $\mathcal{D}_{\mathrm{x}\hat{\mathrm{x}}}$, class label $\mathcal{D}_{\mathrm{c}}$ and latent space regularization $\mathcal{D}_{\mathrm{a}}$. The blue shadowed parts are not used.

Encoder	
Size	Layer
$28 \times 28 \times 1$	Input
$14 \times 14 \times 32$	Conv2D, LeakyReLU
$7 \times 7 \times 64$	Conv2D, LeakyReLU
$4 \times 4 \times 128$	Conv2D, LeakyReLU
2048	Flatten
1024	FC, ReLU
500	FC, ReLU
10	FC, Softmax

Decoder	
Size	Layer
1024	Input
500	FC, ReLU
10	FC, Softmax

$\mathcal{D}_{\mathrm{c}}$	
Size	Layer
10	Input
500	FC, ReLU
500	FC, ReLU
1	FC, Sigmoid

$\mathcal{D}_{\mathrm{z}}$	
Size	Layer
10	Input
500	FC, ReLU
500	FC, ReLU
1	FC, Sigmoid

Table 10: The encoder and decoder of supervised classifier are trained based on $\mathcal{D}_{\mathrm{c}\hat{\mathrm{c}}}$, $\mathcal{D}_{\mathrm{c}}$ and $\mathcal{D}_{\mathrm{z}}$. The encoder is trained with and without batch normalization (BN) after Conv2D layers. $\mathcal{D}_{\mathrm{c}}$ and $\mathcal{D}_{\mathrm{z}}$ are trained in the adversarial way.

We set the parameters $\beta_{\mathrm{x}} = \beta_{\mathrm{c}} = 1$ to compare our system with the vanila AAE. However, these parameters can be also optimized in practice.

The parameters of encoder and decoder are shown in Table 10.

The performance of this classifier without and with batch normalization is shown in Table 11.

Encoder model	runs 1	2	3	mean	std
MNIST 100					
without BN	2.15	2.05	1.78	1.99	0.19
with BN	1.57	1.56	1.92	**1.68**	0.21
MNIST 1000					
without BN	1.55	1.47	1.53	1.52	0.04
with BN	1.37	1.34	1.73	**1.48**	0.22
MNIST all					
without BN	0.78	0.7	0.82	0.77	0.06
with BN	0.79	0.77	0.76	**0.77**	0.02

Table 11: The performance of classifier based on $\mathcal{D}_{c\hat{c}} + \mathcal{D}_c + \mathcal{D}_z + \mathcal{D}_{x\hat{x}}$ for the encoder with and without batch normalization.

Encoder	
Size	Layer
$28 \times 28 \times 1$	Input
$14 \times 14 \times 32$	Conv2D, LeakyReLU
$7 \times 7 \times 64$	Conv2D, LeakyReLU
$4 \times 4 \times 128$	Conv2D, LeakyReLU
2048	Flatten
1024	FC, ReLU
500	FC, ReLU
10	FC, Softmax

$\mathcal{D}_x$	
Size	Layer
$28 \times 28 \times 1$	Input
$14 \times 14 \times 64$	Conv2D, LeakyReLU
$7 \times 7 \times 64$	Conv2D, LeakyReLU
$4 \times 4 \times 128$	Conv2D, LeakyReLU
$4 \times 4 \times 256$	Conv2D, LeakyReLU
4096	Flatten
1	FC, Sigmoid

Decoder	
Size	Layer
1024	Input
500	FC, ReLU
10	FC, Softmax

$\mathcal{D}_c$	
Size	Layer
10	Input
500	FC, ReLU
500	FC, ReLU
1	FC, Sigmoid

$\mathcal{D}_z$	
Size	Layer
10	Input
500	FC, ReLU
500	FC, ReLU
1	FC, Sigmoid

Table 12: The network parameters of semi-supervised classifier are trained based on $\mathcal{D}_{c\hat{c}}$, $\mathcal{D}_c$ and $\mathcal{D}_z$. The encoder is trained with and without batch normalization (BN) after Conv2D layers. $\mathcal{D}_c$ and $\mathcal{D}_z$ are trained in the adversarial way.

2.2.2 Semi-supervised training with latent space regularization and adversarial reconstruction

This model is similar to the previously considered model but in addition to the MSE reconstruction term representing $\mathcal{D}_{x\hat{x}}$ it also contains the adversarial reconstruction term $\mathcal{D}_x$ as defined by (16) in the main part of paper. In our implementation, we consider the regularization of the latent space based on the adversarial term $\mathcal{D}_z$ as shown in Figure 6. The training is based on:

$$\mathcal{L}^{\mathrm{LP}}_{\mathrm{SS-AAE}}(\boldsymbol{\theta}_c, \boldsymbol{\theta}_x, \boldsymbol{\phi}_a, \boldsymbol{\phi}_z) = \mathcal{D}_z + \mathcal{D}_{x\hat{x}} + \mathcal{D}_{c\hat{c}} + \mathcal{D}_c + \alpha_x \mathcal{D}_x. \tag{23}$$

The parameters of encoder and decoder are shown in Table 12.

The performance of this classifier without and with batch normalization is shown in Table 13.

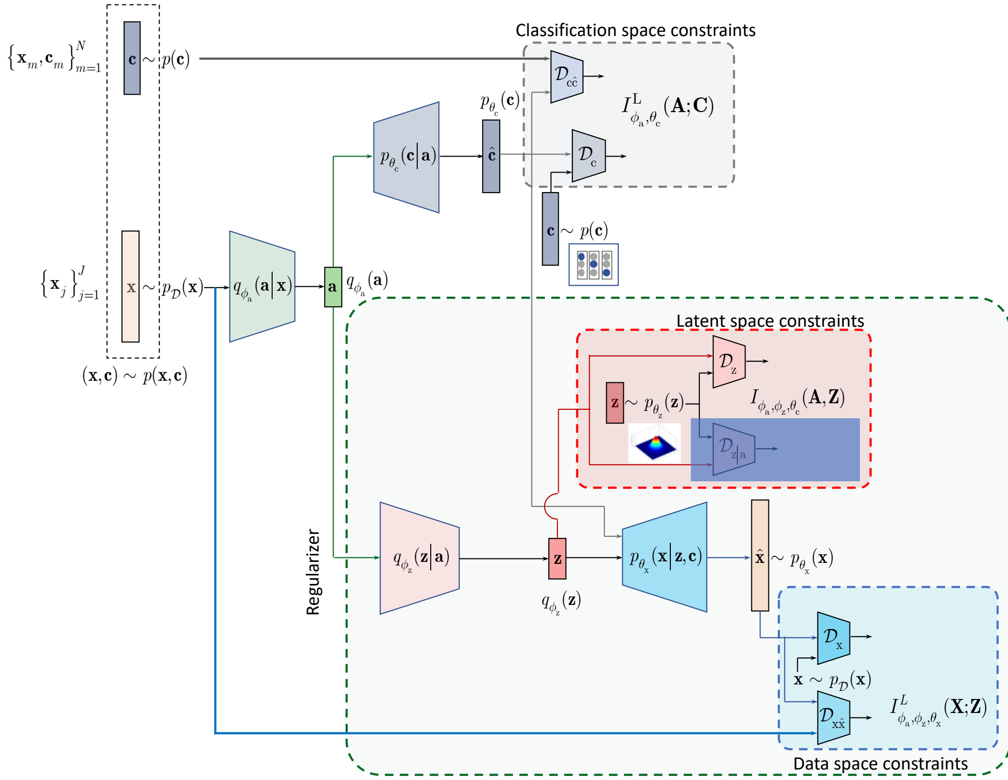


Figure 6: Semi-supervised classifier with learnable priors: the cross-entropy $\mathcal{D}_{\mathrm{c}\hat{\mathrm{c}}}$, MSE $\mathcal{D}_{\mathrm{x}\hat{\mathrm{x}}}$, adversarial reconstruction $\mathcal{D}_{\mathrm{x}}$, class label $\mathcal{D}_{\mathrm{c}}$ and latent space regularizer $\mathcal{D}_{\mathrm{z}}$. The blue shadowed parts are not used.

Encoder model	α_{c}	runs 1	runs 2	runs 3	mean	std
		MNIST 100				
without BN	0.005	2.85	3.36	2.77	2.99	0.32
	0.0005	2.58	2.49	3.08	2.72	0.32
	1	19.62	19.96	15.97	18.52	2.21
with BN	0.005	1.56	1.33	1.35	**1.41**	0.13
	0.0005	1.68	1.66	2.02	1.79	0.20
	1	20.85	13.6	21.67	18.71	4.44
		MNIST 1000				
without BN	0.005	2.29	2.35	2.11	2.25	0.12
	0.0005	1.69	1.88	2.24	1.94	0.28
	1	3.47	3.30	4.12	3.63	0.43
with BN	0.005	1.18	1.21	1.09	**1.16**	0.06
	0.0005	1.44	1.28	1.29	1.34	0.09
	1	4.14	2.94	2.48	3.19	0.86
		MNIST all				
without BN	0.005	0.97	1.01	1.04	1.01	0.04
	0.0005	0.88	0.85	0.93	0.89	0.04
	1	1.31	1.28	1.47	1.35	0.10
with BN	0.005	0.81	0.83	0.75	0.80	0.04
	0.0005	0.73	0.78	0.75	**0.75**	0.03
	1	0.88	0.86	1.27	1.00	0.23

Table 13: The performance of classifier based on $\mathcal{D}_{\mathrm{c}\hat{\mathrm{c}}} + \mathcal{D}_{\mathrm{c}} + \mathcal{D}_{\mathrm{z}} + \mathcal{D}_{\mathrm{x}\hat{\mathrm{x}}} + \alpha_{\mathrm{x}}\mathcal{D}_{\mathrm{x}}$ for the encoder with and without batch normalization.