Don't miss the Mismatch: Investigating the Objective Function Mismatch for Unsupervised Representation Learning (Supplementary Material)

A Proofs

We measure our metrics on the mean losses during cross-validation instead of calculating the metrics for each round and taking the average. We proof that both variants are equivalent for M3 while measuring SM3 on the mean losses leads to a lower bound, given that all models converge at step s_n.

Proposition 1 *The* MM3 *of the average metric value tuples* $\frac{1}{h}\sum_{0<c\leq h} M_c^P$ *and* $\frac{1}{h}\sum_{0<c\leq h} M_c^T$ *with* $0 < c \leq h$ *is equivalent to the average* MM3 *of the individual tuples* $\frac{1}{h}\sum_{0<c\leq h}\text{MM3}(M_c^T, M_c^P)$*, given that the tuples are measured for the same steps S and converge at the same step* s_n*.*

Proof.

$$\begin{aligned}
\frac{1}{h}\sum_{0<c\leq h}\text{MM3}(M_c^T, M_c^P) &= \frac{1}{h}\sum_{0<c\leq h}\frac{1}{n}\sum_{0<i\leq n}(m_{ic}^T - m_{ic}^P) \\
&= \frac{1}{n}\sum_{0<i\leq n}\frac{1}{h}\sum_{0<c\leq h}(m_{ic}^T - m_{ic}^P) \\
&= \frac{1}{n}\sum_{0<i\leq n}\left(\frac{1}{h}\sum_{0<c\leq h} m_{ic}^T - \frac{1}{h}\sum_{0<c\leq h} m_{ic}^P\right) \\
&= \text{MM3}\left(\frac{1}{h}\sum_{0<c\leq h} M_c^T, \frac{1}{h}\sum_{0<c\leq h} M_c^P\right)
\end{aligned}$$

Corollary 1 $\frac{1}{h}\sum_{0<c\leq h}\text{M3}(m_c^T, m_c^P) = \text{M3}(\frac{1}{h}\sum_{0<c\leq h} m_c^T, \frac{1}{h}\sum_{0<c\leq h} m_c^P)$

Proposition 2 *The* MSM3 *of the average metric value tuples* $\frac{1}{h}\sum_{0<c\leq h} M_c^P$ *and* $\frac{1}{h}\sum_{0<c\leq h} M_c^T$ *with* $0 < c \leq h$ *is a lower bound of the average* MSM3 *of the individual tuples* $\frac{1}{h}\sum_{0<c\leq h}\text{MSM3}(M_c^T)$*, given that the tuples are measured for the same steps S and converge at the same step* s_n*.*

Proof.

$$\begin{aligned}
\frac{1}{h}\sum_{0<c\leq h}\text{MSM3}(M_c^T) &= \frac{1}{h}\sum_{0<c\leq h}\frac{1}{n}\sum_{0<i\leq n}\left(m_{ic}^T - \min_{0<j\leq i}(m_{jc}^T)\right) \\
&= \frac{1}{n}\sum_{0<i\leq n}\frac{1}{h}\sum_{0<c\leq h}\left(m_{ic}^T - \min_{0<j\leq i}(m_{jc}^T)\right) \\
&= \frac{1}{n}\sum_{0<i\leq n}\left(\frac{1}{h}\sum_{0<c\leq h} m_{ic}^T - \frac{1}{h}\sum_{0<c\leq h}\min_{0<j\leq i}(m_{jc}^T)\right) \\
&\geq \frac{1}{n}\sum_{0<i\leq n}\left(\frac{1}{h}\sum_{0<c\leq h} m_{ic}^T - \min_{0<j\leq i}(\frac{1}{h}\sum_{0<c\leq h} m_{jc}^T)\right) && \text{since } \frac{1}{h}\sum_{0<c\leq h}\min_{0<j\leq i}(m_{jc}^T) \leq \min_{0<j\leq i}(\frac{1}{h}\sum_{0<c\leq h} m_{jc}^T) \\
&\geq \text{MSM3}\left(\frac{1}{h}\sum_{0<c\leq h} M_c^T\right)
\end{aligned}$$

Corollary 2 $\frac{1}{h}\sum_{0<c\leq h}\text{SM3}(m_c^T, m_c^P) \geq \text{SM3}(\frac{1}{h}\sum_{0<c\leq h} m_c^T, \frac{1}{h}\sum_{0<c\leq h} m_c^P)$

B Additional Model and Training Details

CNN Encoders: We consider a family of convolutional encoders with four Conv-BatchNorm-ReLU layers. Filter widths are $[32, 64, 128, f]$ and paddings are "*valid*". For input sizes of 32×32 (Cifar10, Cifar100), kernel sizes are $[3, 3, 3, 2]$ and strides are $[2, 2, 2, 1]$; for input sizes of 64×64 (3dshapes, PCam), kernel sizes are $[4, 4, 4, 3]$ and strides are $[2, 2, 2, 2]$. Weights are initialized with the standard TensorFlow initialization (kernel_initializer="glorot_uniform", bias_initializer="zeros"). We vary f in $[4, 32, 128, 256, 512, 1024]$ for our experiments on representation sizes. For all other experiments $f = 256$.

CNN Image Decoders: We consider a family of decoders with transposed convolutions with three TranConv-BatchNorm-ReLU layers followed by a TranConv-BatchNorm-Sigmoid layer. Filter widths are $[128, 64, 32, 3]$ and paddings are "*valid*". For input sizes of 32×32 (Cifar10, Cifar100), kernel sizes are $[4, 4, 4, 3]$ and strides are $[2, 2, 2, 1]$; for input sizes of 64×64 (3dshapes, PCam), kernel sizes are $[4, 4, 5, 4]$ and strides are $[2, 2, 2, 2]$. Weights are initialized with the standard TensorFlow initialization (kernel_initializer="glorot_uniform", bias_initializer="zeros").

CNN/ResNet Head for Rotation: For our CNN encoder we use a fully-connected layer with 4 neurons and softmax activation as head to predict the four different rotations. We use the standard TensorFlow initialization (kernel_initializer="glorot_uniform", bias_initializer="zeros"). For our ResNet decoder we initialize with (kernel_initializer=RandomNormal(stddev=.01), bias_initializer="zeros").

CNN/ResNet Heads for Contrastive Learning: For our CNN encoder we use a two layer MLP with a FC-BatchNorm-ReLU layer followed by a FC-BatchNorm-Softmax layer as projection head for contrastive learning. Number of neurons are $[f, 128]$. We use the standard TensorFlow initialization (kernel_initializer="glorot_uniform", bias_initializer="zeros"). We vary f in $[4, 32, 128, 256, 512, 1024]$ for our experiments on representation sizes. For all other experiments $f = 256$. For our ResNet head the number of neurons are $[512, 128]$ and we initialize as in [9].

Target Models: For our linear target model we use a fully-connected layer with *num_classes* neurons and a softmax activation. For our two- and three-layer nonlinear models we add layers consisting of [256] and [512, 256] hidden units with batch normalization followed by ReLu activations respectively. Weights are initialized with the standard TensorFlow initialization (kernel_initializer=="glorot_uniform", bias_initializer="zeros").

Hardware: We carry out our experiments on two servers which contain four Nvidia GeForce RTX 2080 Ti GPUs respectively.

Mismatch Evaluation: In Table 4 we show additional details about training, evaluation and measurements.

Table 4 Information about measurements and training

	Measurement Epochs	Convergence Criterium	Pretext Model Training Epochs	Target Model Training Epochs	Validation
Rep. Size, TMC, Augs					
CAE(Cifar10)	(0,5,20,50,100,...,400)	Patience:3	400	500	5-fold cross-validation
DCAE(Cifar10)	(0,5,20,50,100,...,400)	Patience:3	400	500	5-fold cross-validation
CCAE(Cifar100)	(0,5,20,50,100,...,400)	Patience:6	400	500	5-fold cross-validation
CCAE(PCam)	(0,10,50,100,150,200,300,...,800)	Patience:10	800	500	5-fold cross-validation
RCAE(PCam)	(0,200,400,...,2000)	Patience:30	2000	500	5-fold cross-validation
SCLCAE(3dshapes)	(0,10,50,100,150,200,300,...,600)	Patience:15	60 0	100	5-fold cross-validation
Target Task Type					
CAE(3dshapes)	(0,10,30,50,100,...,400)	Patience:3	400	100	5-fold cross-validation
DCAE(3dshapes)	(0,10,30,50,100,...,400)	Patience:3	400	100	5-fold cross-validation
CCAE(3dshapes)	(0,10,30,50,100,...,400)	Patience:3	400	100	5-fold cross-validation
RCAE(3dshapes)	(0,10,30,50,100,...,400)	Patience:3	400	100	5-fold cross-validation
SCLCAE(3dshapes)	(0,10,30,50,100,...,600)	Patience:3	600	100	5-fold cross-validation
Stability					
CAE100E(Cifar10)	(0,1,...,100)	Epoch 100	100	500	5-fold cross-validation
CAE(Cifar10)	(0,5,20,50,100,...,400)	Epoch 400	400	500	5-fold cross-validation
CAENoCrossVal(Cifar10)	(0,5,20,50,100,...,400)	Epoch 400	400	500	$5 \times$ same split
ResNets					
RResNet18(Cifar10)	(0,50,100,200,400,...,4000)	Patience:30	4000	600	5-fold cross-validation
SCLResNet18(Cifar10)	(0,50,100,200,400,...,4000)	Epoch 4000	4000	600	3-fold cross-validation
SCLResNet18(Cifar100)	(0,50,100,200,400,...,4000)	Epoch 4000	4000	600	3-fold cross-validation
SCLResNet18(PCam)	(0,400,800,...,5000)	Patience:60	5000	500	5-fold cross-validation
ResNets Rep. Size					
RResNet18(Cifar10)	(0,100,200,...,1000,1200,...,3000)	Patience:30	3000	700	5-fold cross-validation

Under (Rep. Size, TMC, Augs) we refer to all models trained with different representations, target model complexities and augmentations.

C Additional Evidence

Table 5 Detailed version of CAE (Cifar10) and DCAE (Cifar10) from Table 1

	CAE (Cifar10)			**DCAE** (Cifar10)		
	ACC	cSM3	MOFM	ACC	cSM3	MOFM
Rep. Size						
2x2x4	$27.94^{+1.47}_{-0.76}$	$\mathbf{0.00}^{+0.00}_{-0.00}$	$\mathbf{0.00}^{+29.77}_{-0.00}$	$28.06^{+0.70}_{-1.65}$	$0.05^{+0.12}_{-0.05}$	$\mathbf{0.00}^{+14.27}_{-0.00}$
2x2x32	$36.57^{+0.92}_{-0.92}$	$0.07^{+0.09}_{-0.07}$	$1.99^{+3.54}_{-1.44}$	$36.20^{+0.52}_{-0.50}$	$\mathbf{0.00}^{+0.00}_{-0.00}$	$3.20^{+10.24}_{-0.59}$
2x2x128	$41.94^{+0.45}_{-0.47}$	$0.20^{+0.42}_{-0.20}$	$10.10^{+6.24}_{-3.51}$	$41.79^{+0.46}_{-0.49}$	$0.06^{+0.11}_{-0.06}$	$5.51^{+6.79}_{-5.32}$
2x2x256	$44.69^{+0.69}_{-0.41}$	$0.75^{+0.38}_{-0.22}$	$11.14^{+5.32}_{-2.72}$	$45.42^{+0.55}_{-0.62}$	$0.69^{+0.43}_{-0.69}$	$5.17^{+2.98}_{-2.16}$
2x2x512	$48.13^{+0.58}_{-0.88}$	$0.43^{+0.38}_{-0.43}$	$5.28^{+1.93}_{-2.98}$	$49.04^{+0.29}_{-0.29}$	$0.36^{+0.75}_{-0.36}$	$1.25^{+1.81}_{-1.19}$
2x2x1024	$\mathbf{51.42}^{+0.43}_{-0.38}$	$\mathbf{0.24}^{+0.32}_{-0.24}$	$\mathbf{0.25}^{+1.26}_{-0.17}$	$\mathbf{53.82}^{+0.57}_{-0.82}$	$\mathbf{0.03}^{+0.06}_{-0.03}$	$\mathbf{0.00}^{+0.04}_{-0.00}$
Target Model						
FC	$44.69^{+0.69}_{-0.41}$	$0.75^{+0.38}_{-0.22}$	$11.14^{+5.32}_{-2.72}$	$45.42^{+0.55}_{-0.62}$	$0.69^{+0.43}_{-0.69}$	$5.17^{+2.98}_{-2.16}$
2FC	$56.72^{+0.50}_{-0.42}$	$\mathbf{0.03}^{+0.10}_{-0.03}$	$5.68^{+2.87}_{-3.04}$	$57.26^{+0.68}_{-0.52}$	$\mathbf{0.00}^{+0.00}_{-0.00}$	$5.14^{+2.68}_{-1.98}$
3FC	$\mathbf{63.05}^{+0.74}_{-0.72}$	$\mathbf{0.03}^{+0.11}_{-0.03}$	$\mathbf{3.94}^{+0.93}_{-1.61}$	$\mathbf{63.33}^{+0.23}_{-0.39}$	$\mathbf{0.00}^{+0.00}_{-0.00}$	$\mathbf{3.17}^{+1.51}_{-0.89}$
Augmentations						
All	$44.69^{+0.69}_{-0.41}$	$\mathbf{0.75}^{+0.38}_{-0.22}$	$11.14^{+5.32}_{-2.72}$	$45.42^{+0.55}_{-0.62}$	$0.69^{+0.43}_{-0.69}$	$5.17^{+2.98}_{-2.16}$
NoJitter	$46.30^{+0.50}_{-0.46}$	$0.99^{+0.59}_{-0.47}$	$\mathbf{10.99}^{+1.24}_{-3.25}$	$47.27^{+0.46}_{-0.50}$	$\mathbf{0.50}^{+0.45}_{-0.50}$	$2.33^{+1.74}_{-1.57}$
NoJitterNoFlip	$\mathbf{46.48}^{+0.51}_{-0.75}$	$1.00^{+0.48}_{-0.60}$	$12.51^{+1.46}_{-1.74}$	$\mathbf{47.38}^{+0.36}_{-0.44}$	$0.55^{+0.51}_{-0.55}$	$\mathbf{1.73}^{+4.43}_{-1.36}$

Table 6 Detailed version of CCAE (Cifar100) and RCAE (PCam) from Table 1

	CCAE (Cifar100)			**RCAE** (PCam)			
	ACC	cSM3	MOFM	ACC	cSM3	MOFM	MM3
Rep. Size							
2x2x4	$9.66^{+0.30}_{-0.45}$	$0.28^{+0.19}_{-0.13}$	$1.54^{+2.09}_{-0.98}$	$67.62^{+1.18}_{-2.05}$	$5.38^{+2.15}_{-1.66}$	$4.15^{+37.30}_{-4.15}$	$-22.26^{+2.96}_{-3.02}$
2x2x32	$17.63^{+0.36}_{-0.51}$	$0.65^{+0.28}_{-0.64}$	$3.64^{+4.14}_{-2.04}$	$72.83^{+1.17}_{-1.72}$	$3.34^{+0.74}_{-0.52}$	$7.92^{+18.08}_{-1.46}$	$-21.09^{+1.90}_{-2.48}$
2x2x128	$24.36^{+0.56}_{-0.47}$	$0.51^{+0.24}_{-0.24}$	$0.81^{+1.60}_{-0.45}$	$78.17^{+0.63}_{-0.80}$	$1.03^{+0.88}_{-0.96}$	$4.04^{+4.35}_{-0.70}$	$-23.47^{+0.55}_{-0.65}$
2x2x256	$28.36^{+0.78}_{-0.68}$	$0.17^{+0.37}_{-0.17}$	$\mathbf{0.00}^{+0.00}_{-0.00}$	$79.55^{+1.18}_{-1.07}$	$0.44^{+0.79}_{-0.44}$	$\mathbf{0.00}^{+3.95}_{-0.85}$	$-27.60^{+1.22}_{-1.00}$
2x2x512	$32.02^{+0.51}_{-0.74}$	$\mathbf{0.00}^{+0.02}_{-0.00}$	$\mathbf{0.00}^{+0.00}_{-0.00}$	$80.80^{+0.66}_{-0.63}$	$0.18^{+0.15}_{-0.18}$	$\mathbf{0.00}^{+1.13}_{-0.00}$	$\mathbf{-28.03}^{+1.28}_{-1.35}$
2x2x1024	$\mathbf{34.89}^{+0.41}_{-0.83}$	$\mathbf{0.00}^{+0.00}_{-0.00}$	$\mathbf{0.00}^{+0.00}_{-0.00}$	$\mathbf{82.20}^{+0.76}_{-0.68}$	$\mathbf{0.09}^{+0.12}_{-0.09}$	$\mathbf{0.00}^{+0.34}_{-0.00}$	$-26.56^{+1.14}_{-1.18}$
Target Model							
FC	$32.02^{+0.51}_{-0.74}$	$\mathbf{0.00}^{+0.02}_{-0.00}$	$\mathbf{0.00}^{+0.00}_{-0.00}$	$78.17^{+0.63}_{-0.80}$	$1.03^{+0.88}_{-0.96}$	$4.04^{+4.35}_{-0.70}$	$-23.47^{+0.55}_{-0.65}$
2FC	$36.01^{+0.31}_{-0.27}$	$0.08^{+0.32}_{-0.08}$	$\mathbf{0.00}^{+0.39}_{-0.00}$	$82.99^{+0.65}_{-0.62}$	$\mathbf{0.31}^{+0.31}_{-0.31}$	$0.76^{+1.76}_{-0.04}$	$-28.56^{+0.81}_{-0.72}$
3FC	$\mathbf{38.40}^{+0.34}_{-0.50}$	$0.12^{+0.28}_{-0.12}$	$0.02^{+0.44}_{-0.09}$	$\mathbf{84.18}^{+0.62}_{-0.45}$	$0.37^{+0.21}_{-0.37}$	$\mathbf{0.61}^{+3.68}_{-0.51}$	$\mathbf{-29.61}^{+0.71}_{-0.61}$
Augmentations							
All	$28.36^{+0.78}_{-0.68}$	$\mathbf{0.17}^{+0.37}_{-0.17}$	$\mathbf{0.00}^{+0.00}_{-0.00}$	$78.17^{+0.63}_{-0.80}$	$1.03^{+0.88}_{-0.96}$	$4.04^{+4.35}_{-0.70}$	$-23.47^{+0.55}_{-0.65}$
NoJitter	-	-	-	$\mathbf{80.88}^{+0.88}_{-0.94}$	$\mathbf{0.58}^{+0.34}_{-0.58}$	$\mathbf{0.10}^{+5.52}_{-0.04}$	$\mathbf{-28.60}^{+2.23}_{-1.18}$
NoJitterNoFlip	-	-	-	$80.55^{+0.73}_{-0.55}$	$1.51^{+0.57}_{-1.09}$	$7.33^{+9.76}_{-1.78}$	$-10.60^{+1.21}_{-1.88}$
NoFlip	$\mathbf{29.33}^{+0.65}_{-0.61}$	$0.20^{+0.43}_{-0.20}$	$\mathbf{0.00}^{+0.22}_{-0.00}$	-	-	-	-

Table 7 Detailed version of CCAE (PCam) and SCLCAE (3dshapes for object hue classification) from Table 1

	CCAE (PCam)			**SCLCAE** (3dshapes)			
	ACC	cSM3	MOFM	ACC	cSM3	MOFM	MM3
Rep. Size							
2x2x4	$63.39^{+3.80}_{-1.58}$	$4.98^{+5.53}_{-3.53}$	$9.28^{+\infty}_{-2.09}$	$38.07^{+11.18}_{-10.06}$	$26.34^{+11.38}_{-9.69}$	∞	$-7.99^{+1.83}_{-1.28}$
2x2x32	$72.73^{+2.98}_{-2.37}$	$5.17^{+2.95}_{-2.80}$	$34.30^{+16.31}_{-10.21}$	$85.25^{+2.77}_{-2.36}$	$12.98^{+10.56}_{-12.24}$	$36.39^{+26.87}_{-29.73}$	$-57.67^{+6.98}_{-6.46}$
2x2x128	$78.66^{+0.46}_{-0.22}$	$0.32^{+0.47}_{-0.32}$	$0.10^{+3.51}_{-0.48}$	$96.54^{+0.72}_{-0.83}$	$8.14^{+1.29}_{-2.53}$	$\mathbf{22.65}^{+13.03}_{-3.92}$	$\mathbf{-66.19}^{+1.24}_{-1.71}$
2x2x256	$79.97^{+0.68}_{-0.51}$	$0.43^{+0.51}_{-0.43}$	$0.87^{+4.56}_{-0.41}$	$98.65^{+0.55}_{-0.80}$	$6.52^{+0.62}_{-0.84}$	$27.65^{+5.66}_{-4.87}$	$-65.77^{+2.19}_{-1.12}$
2x2x512	$82.34^{+0.66}_{-0.87}$	$0.20^{+0.59}_{-0.20}$	$0.07^{+0.24}_{-0.07}$	$99.41^{+0.20}_{-0.25}$	$5.96^{+1.13}_{-0.40}$	$27.78^{+8.80}_{-7.10}$	$-63.61^{+1.75}_{-1.50}$
2x2x1024	$\mathbf{83.67}^{+0.45}_{-0.61}$	$\mathbf{0.00}^{+0.01}_{-0.00}$	$\mathbf{0.00}^{+0.20}_{-0.00}$	$\mathbf{99.75}^{+0.04}_{-0.04}$	$\mathbf{4.76}^{+1.84}_{-1.36}$	$32.70^{+14.59}_{-7.15}$	$-57.92^{+2.29}_{-2.25}$
Target Model							
FC	$83.67^{+0.45}_{-0.61}$	$\mathbf{0.00}^{+0.01}_{-0.00}$	$\mathbf{0.00}^{+0.20}_{-0.00}$	$38.07^{+11.18}_{-10.06}$	$6.52^{+0.62}_{-0.84}$	$\mathbf{27.65}^{+5.66}_{-4.87}$	$-65.77^{+2.19}_{-1.12}$
2FC	$89.17^{+0.50}_{-0.38}$	$\mathbf{0.00}^{+0.00}_{-0.00}$	$\mathbf{0.00}^{+0.12}_{-0.00}$	$85.25^{+2.77}_{-2.36}$	$1.84^{+0.33}_{-0.66}$	$103.30^{+56.89}_{-21.22}$	$-70.49^{+1.41}_{-0.75}$
3FC	$\mathbf{90.59}^{+0.48}_{-0.60}$	$0.08^{+0.19}_{-0.08}$	$\mathbf{0.00}^{+0.44}_{-0.00}$	$\mathbf{96.54}^{+0.72}_{-0.83}$	$\mathbf{0.91}^{+0.18}_{-0.30}$	$258.18^{+308.48}_{-88.99}$	$\mathbf{-71.26}^{+1.20}_{-0.77}$
Augmentations							
All	$79.97^{+0.68}_{-0.51}$	$0.43^{+0.51}_{-0.43}$	$0.87^{+4.56}_{-0.41}$	$38.07^{+11.18}_{-10.06}$	$6.52^{+0.62}_{-0.84}$	$27.65^{+5.66}_{-4.87}$	$\mathbf{-65.77}^{+2.19}_{-1.12}$
NoJitter	-	-	-	$85.25^{+2.77}_{-2.36}$	$\mathbf{0.00}^{+0.02}_{-0.00}$	$0.01^{+0.07}_{-0.01}$	$-37.92^{+0.72}_{-0.89}$
NoJitterNoFlip	-	-	-	$\mathbf{96.54}^{+0.72}_{-0.83}$	$\mathbf{0.00}^{+0.00}_{-0.00}$	$\mathbf{0.00}^{+0.04}_{-0.01}$	$-35.95^{+0.63}_{-0.82}$
NoFlip	$\mathbf{80.70}^{+0.36}_{-0.57}$	$\mathbf{0.41}^{+0.21}_{-0.30}$	$\mathbf{0.04}^{+1.03}_{-0.20}$	-	-	-	-

Table 8 Detailed version of CAE and DCAE from Table 2

	CAE			DCAE		
	ACC	cSM3	MOFM	ACC	cSM3	MOFM
floor_hue	$99.96^{+0.02}_{-0.02}$	$0.01^{+0.02}_{-0.01}$	$0.95^{+0.71}_{-0.51}$	$99.97^{+0.02}_{-0.06}$	$\mathbf{0.00}^{+0.00}_{-0.00}$	$1.28^{+1.11}_{-0.81}$
wall_hue	$\mathbf{100.00}^{+0.00}_{-0.00}$	$0.02^{+0.04}_{-0.02}$	$32.03^{+6.38}_{-8.59}$	$\mathbf{99.99}^{+0.00}_{-0.01}$	$\mathbf{0.00}^{+0.01}_{-0.00}$	$24.43^{+7.56}_{-6.97}$
object_hue	$99.23^{+0.18}_{-0.35}$	$0.38^{+0.82}_{-0.37}$	$22.71^{+14.77}_{-6.45}$	$99.22^{+0.26}_{-0.31}$	$0.43^{+0.61}_{-0.43}$	$24.55^{+7.04}_{-11.17}$
scale	$74.17^{+6.07}_{-6.42}$	$0.41^{+1.64}_{-0.41}$	$\mathbf{0.00}^{+0.00}_{-0.00}$	$68.27^{+2.36}_{-2.04}$	$0.27^{+0.59}_{-0.27}$	$\mathbf{0.00}^{+0.00}_{-0.00}$
shape	$98.38^{+0.82}_{-1.19}$	$0.07^{+0.14}_{-0.07}$	$\mathbf{0.00}^{+0.04}_{-0.00}$	$97.45^{+0.46}_{-0.57}$	$0.08^{+0.20}_{-0.08}$	$\mathbf{0.00}^{+0.06}_{-0.00}$
orientation	$81.63^{+3.23}_{-2.89}$	$\mathbf{0.00}^{+0.00}_{-0.00}$	$\mathbf{0.00}^{+0.00}_{-0.00}$	$74.27^{+1.07}_{-2.34}$	$\mathbf{0.00}^{+0.00}_{-0.00}$	$\mathbf{0.00}^{+0.00}_{-0.00}$
average	**92.22**	0.15	9.28	89.86	**0.13**	8.21

Table 9 Detailed version of CCAE and RCAE from Table 2

	CCAE			RCAE			
	ACC	cSM3	MOFM	ACC	cSM3	MOFM	MM3
floor_hue	$99.94^{+0.03}_{-0.05}$	$\mathbf{0.02}^{+0.03}_{-0.02}$	$\mathbf{0.00}^{+0.00}_{-0.00}$	$92.08^{+1.79}_{-2.89}$	$56.68^{+0.94}_{-1.55}$	∞	$44.67^{+4.10}_{-3.78}$
wall_hue	$\mathbf{99.98}^{+0.01}_{-0.02}$	$0.10^{+0.11}_{-0.07}$	$\mathbf{0.00}^{+0.78}_{-0.00}$	$\mathbf{99.96}^{+0.04}_{-0.04}$	$25.17^{+9.67}_{-8.15}$	∞	$7.80^{+0.59}_{-0.85}$
object_hue	$98.96^{+0.38}_{-0.30}$	$1.55^{+0.73}_{-0.63}$	$0.63^{+5.58}_{-0.59}$	$98.79^{+0.35}_{-0.41}$	$59.65^{+8.12}_{-10.39}$	∞	$40.11^{+0.99}_{-1.88}$
scale	$66.72^{+3.20}_{-2.20}$	$0.10^{+0.41}_{-0.10}$	$\mathbf{0.00}^{+0.00}_{-0.00}$	$67.10^{+6.77}_{-3.96}$	$2.60^{+4.10}_{-2.60}$	$0.13^{+3.54}_{-0.13}$	$31.78^{+1.91}_{-2.44}$
shape	$98.75^{+0.39}_{-0.50}$	$0.03^{+0.10}_{-0.03}$	$\mathbf{0.00}^{+0.00}_{-0.00}$	$98.17^{+0.43}_{-0.39}$	$\mathbf{0.20}^{+0.45}_{-0.20}$	$0.06^{+0.82}_{-0.06}$	$\mathbf{-2.48}^{+0.65}_{-1.28}$
orientation	$72.55^{+3.17}_{-2.81}$	$0.23^{+0.91}_{-0.23}$	$\mathbf{0.00}^{+0.00}_{-0.00}$	$80.69^{+2.55}_{-2.68}$	$0.48^{+1.91}_{-0.48}$	$\mathbf{0.00}^{+1.23}_{-0.00}$	$22.26^{+0.74}_{-1.19}$
average	89.48	0.34	**0.11**	89.47	24.13	∞	24.02

Table 10 Detailed version of SCLCAE from Table 2

	SCLCAE			
	ACC	cSM3	MOFM	MM3
floor_hue	$93.53^{+2.62}_{-1.90}$	$28.18^{+11.12}_{-6.30}$	$268.27^{+161.92}_{-44.43}$	$-48.38^{+2.75}_{-2.15}$
wall_hue	$\mathbf{99.96}^{+0.03}_{-0.02}$	$\mathbf{0.29}^{+0.07}_{-0.11}$	$0.46^{+5.04}_{-0.44}$	$\mathbf{-76.40}^{+2.97}_{-3.06}$
object_hue	$98.67^{+0.82}_{-1.06}$	$2.87^{+0.69}_{-1.36}$	$8.69^{+3.20}_{-0.64}$	$-73.17^{+3.16}_{-2.77}$
scale	$83.94^{+1.15}_{-0.57}$	$2.43^{+4.02}_{-2.43}$	$\mathbf{0.00}^{+1.78}_{-0.00}$	$-44.80^{+2.06}_{-1.55}$
shape	$95.06^{+0.92}_{-0.86}$	$1.67^{+1.54}_{-1.67}$	$2.16^{+1.48}_{-0.13}$	$-67.54^{+2.38}_{-1.63}$
orientation	$45.32^{+4.14}_{-3.57}$	$2.50^{+2.55}_{-1.97}$	$6.68^{+5.21}_{-3.02}$	$-9.11^{+2.46}_{-1.71}$
average	86.08	6.32	47.71	**−53.23**

Table 11 Mismatches of other models we have tested

	Cifar10			
	ACC	cSM3	MOFM	MM3
CAE	$44.69^{+0.69}_{-0.41}$	$1.42^{+0.68}_{-0.70}$	$18.71^{+6.61}_{-4.01}$	-
CAENoCrossVal	$44.24^{+0.24}_{-0.41}$	$1.26^{+0.15}_{-0.47}$	$17.40^{+2.91}_{-2.58}$	-
CAE100E	$45.37^{+1.09}_{-1.06}$	$1.22^{+0.56}_{-0.32}$	$5.84^{+7.59}_{-1.12}$	-
RResNet18	$54.64^{+1.80}_{-2.01}$	$3.98^{+1.74}_{-3.60}$	$4.87^{+4.42}_{-3.11}$	$31.82^{+0.75}_{-0.68}$
SCLResNet18	$87.14^{+0.37}_{-0.40}$	$0.29^{+0.23}_{-0.20}$	$0.12^{+0.07}_{-0.02}$	$-31.83^{+0.12}_{-0.23}$
RResNet18R32	$39.14^{+1.53}_{-1.69}$	$1.92^{+1.75}_{-1.29}$	$4.39^{+9.53}_{-1.58}$	$48.93^{+1.17}_{-2.19}$
RResNet18R256	$47.46^{+1.44}_{-1.76}$	$3.91^{+2.94}_{-1.71}$	$6.86^{+8.51}_{-2.09}$	$39.08^{+1.78}_{-2.07}$
RResNet18R512	$50.55^{+2.91}_{-2.84}$	$7.01^{+1.83}_{-3.73}$	$9.83^{+10.55}_{-3.90}$	$36.93^{+3.15}_{-1.96}$
RResNet18R756	$51.80^{+2.04}_{-2.38}$	$7.56^{+5.01}_{-3.17}$	$8.50^{+11.90}_{-5.00}$	$35.61^{+1.45}_{-1.74}$
RResNet18R1024	$52.99^{+1.54}_{-2.58}$	$9.89^{+4.51}_{-2.41}$	$15.12^{+11.12}_{-4.18}$	$36.86^{+2.76}_{-2.20}$

Values are obtained by cross-validation, please refer to Table 4 for more details.

Table 12 Mismatches of other models we have tested

	PCam				Cifar100			
	ACC	cSM3	MOFM	MM3	ACC	cSM3	MOFM	MM3
SCLResNet18	$96.25^{+0.44}_{-0.23}$	$0.37^{+0.44}_{-0.37}$	$0.86^{+1.00}_{-0.60}$	$-53.26^{+0.52}_{-0.38}$	$59.20^{+0.19}_{-0.25}$	$0.41^{+0.10}_{-0.09}$	$0.06^{+0.05}_{-0.04}$	$0.84^{+0.06}_{-0.03}$

Values are obtained by cross-validation, please refer to Table 4 for more details.

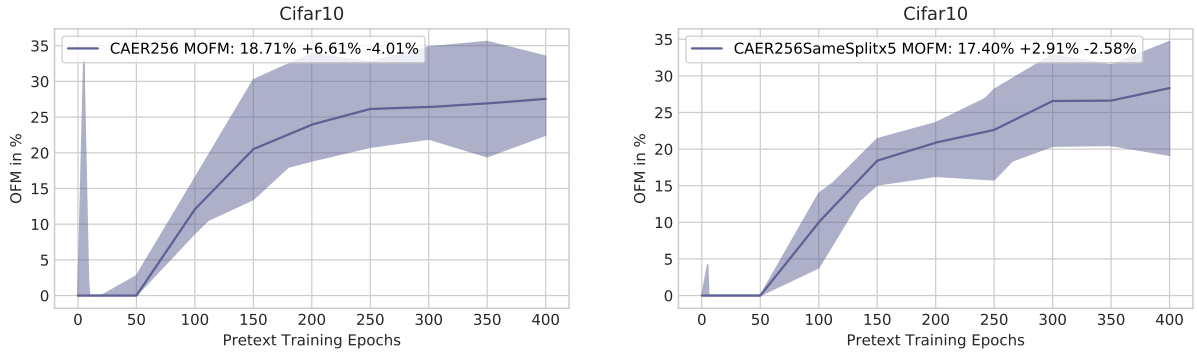

Fig. 7 (Left) OFM of a CAE trained for 400 epochs. Stability is measured by 5-fold cross-validation. (Right) OFM of a CAE trained for 400 epochs. Stability is measured by training the CAE five times on the same dataset split. Unsurprisingly, the stability of the OFM is higher when the CAE is trained on the same split instead of the different splits from the 5-fold cross-validation.

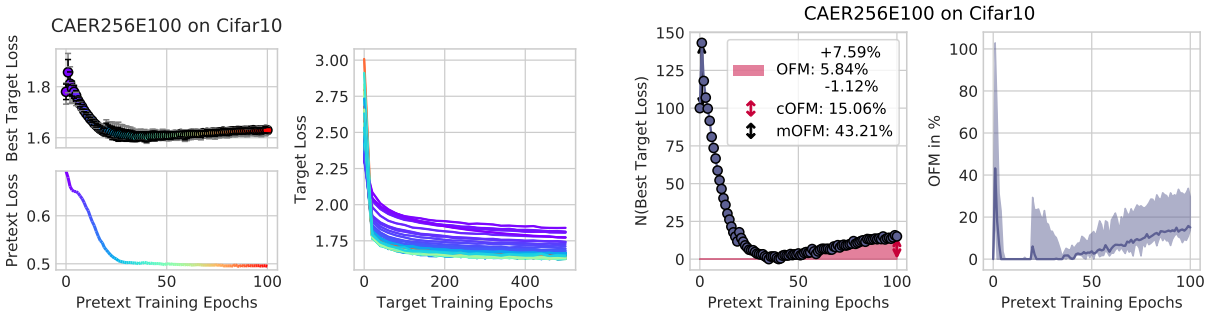

Fig. 8 (Left) Losses of a simple CAE measured for every pretext training epoch. The curve formed by the target models represents a typical target training curve in our setup. (Right) The OFM and its stability measured for every pretext training epoch of the CAE. When we compare the stability to the partially measured CAE in Figure 7, we observe a similar instability.

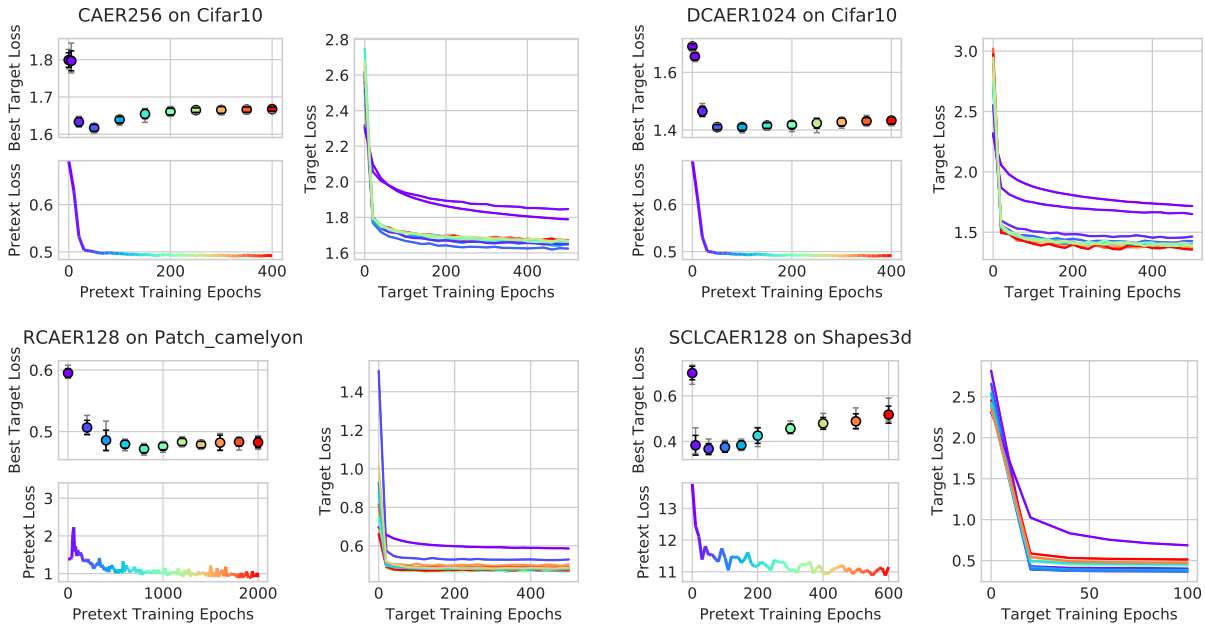

Fig. 9 Additional evidence for the Mismatch and Convergence Section 6.1: *Longer training of the pretext task tends to create easier separable representations which may mismatch with the class label*. We observe that the target loss curves converge faster for target models trained on the pretext model's representation from later pretext training epochs. Especially the purple curves show this behavior clearly.

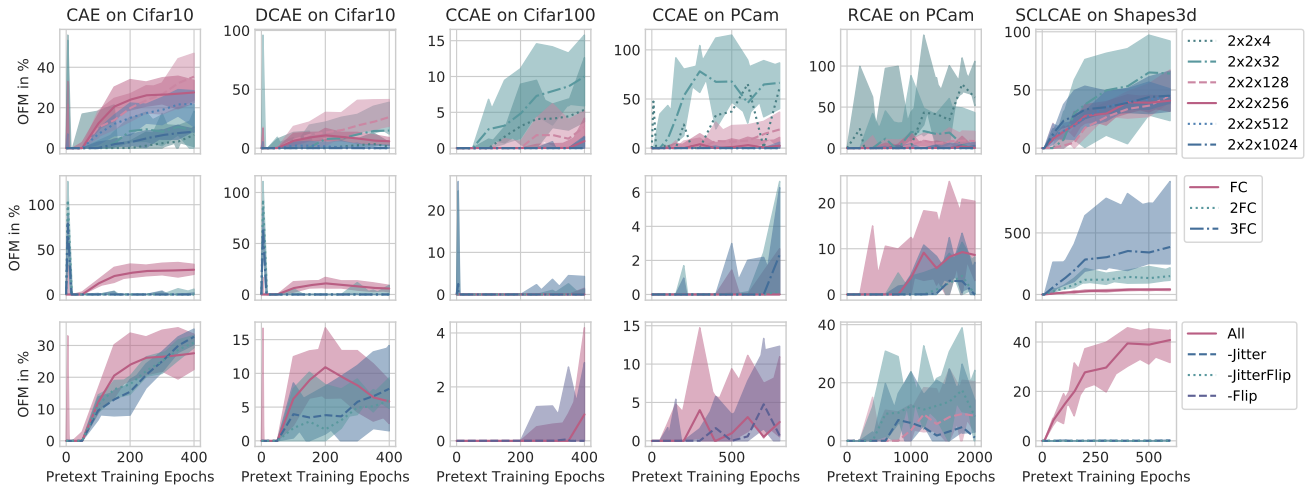

Fig. 10 Version of Figure 4 without convergence criterium. We observe similar behaviors of the mismatches as in Figure 4 in most cases. One exception is the color jitter, where the OFM starts to converge or decrease late in training. (Top) Impact of different pretext model representation sizes on the OFM for our model. (Middle) The OFM for the linear and nonlinear target models trained on our pretext model. (Bottom) The OFM for the linear target model and for the pretext models trained on fewer augmentations. First we removed the color jitter and then the vertical flip from the augmentations. For the CCAE we only removed the vertical flip. The target models of SCLCAE were trained on 3dshapes, to predict the object hue.

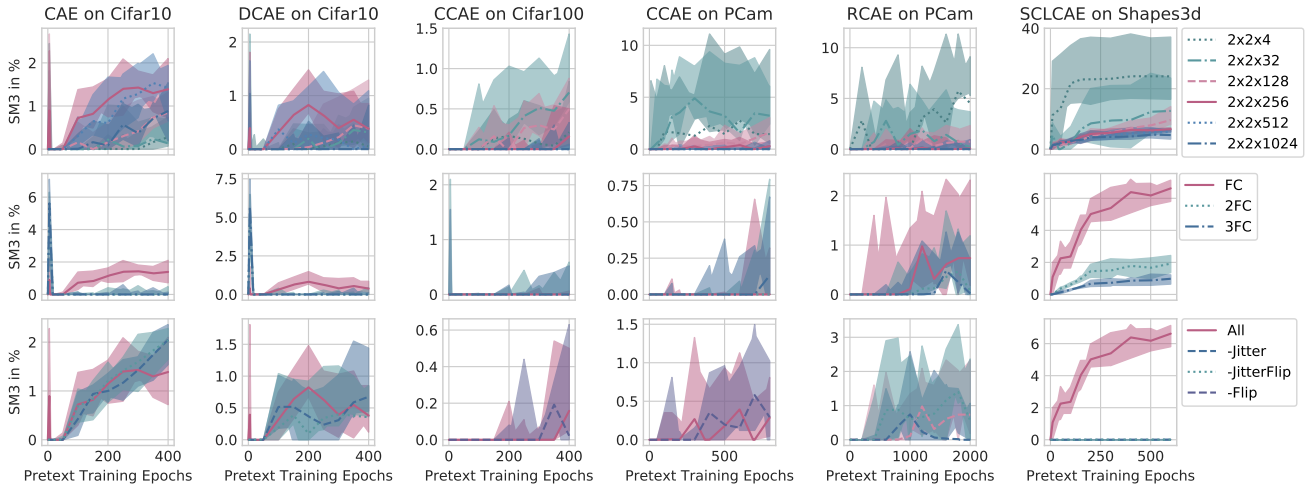

Fig. 11 Version of Figure 4 for SM3 on accuracies, without convergence criterium. Again, we observe the similar behaviors of the mismatches as in Figure 4 in most cases.

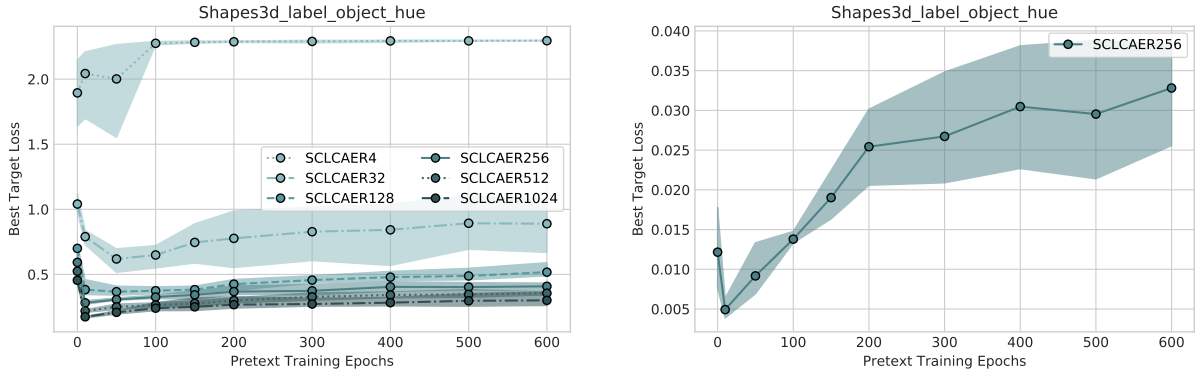

Fig. 12 (Left) We describe why the *OFM* for higher representation sizes does not decrease in the setup where we train the contrastive pretext model (SCLCAE) on 3dshaped and the target model to predict the object hue. Here, the target models trained on the untrained pretext models with larger representation sizes already achieve high performance due to a higher amount of color-selective, random features. Additionally, learning of the pretext model does not lead to a high performance gain, which leads again to a small interval for normalization. Therefore, forgetting useful features for the target task later in training leads to a high mismatch. (Right) We describe why the *OFM* does not decrease for more complex target models in the same setup. The nonlinear target model can make better sense of specific random pretext features for classification, which leads to a very low target loss at pretext model initialization. Since the pretext model does not learn many useful features for the target task later, this leads again to a small interval for normalization. Therefore, the *OFM* gets very large later in training, when the pretext model starts to forget useful features for the target task.

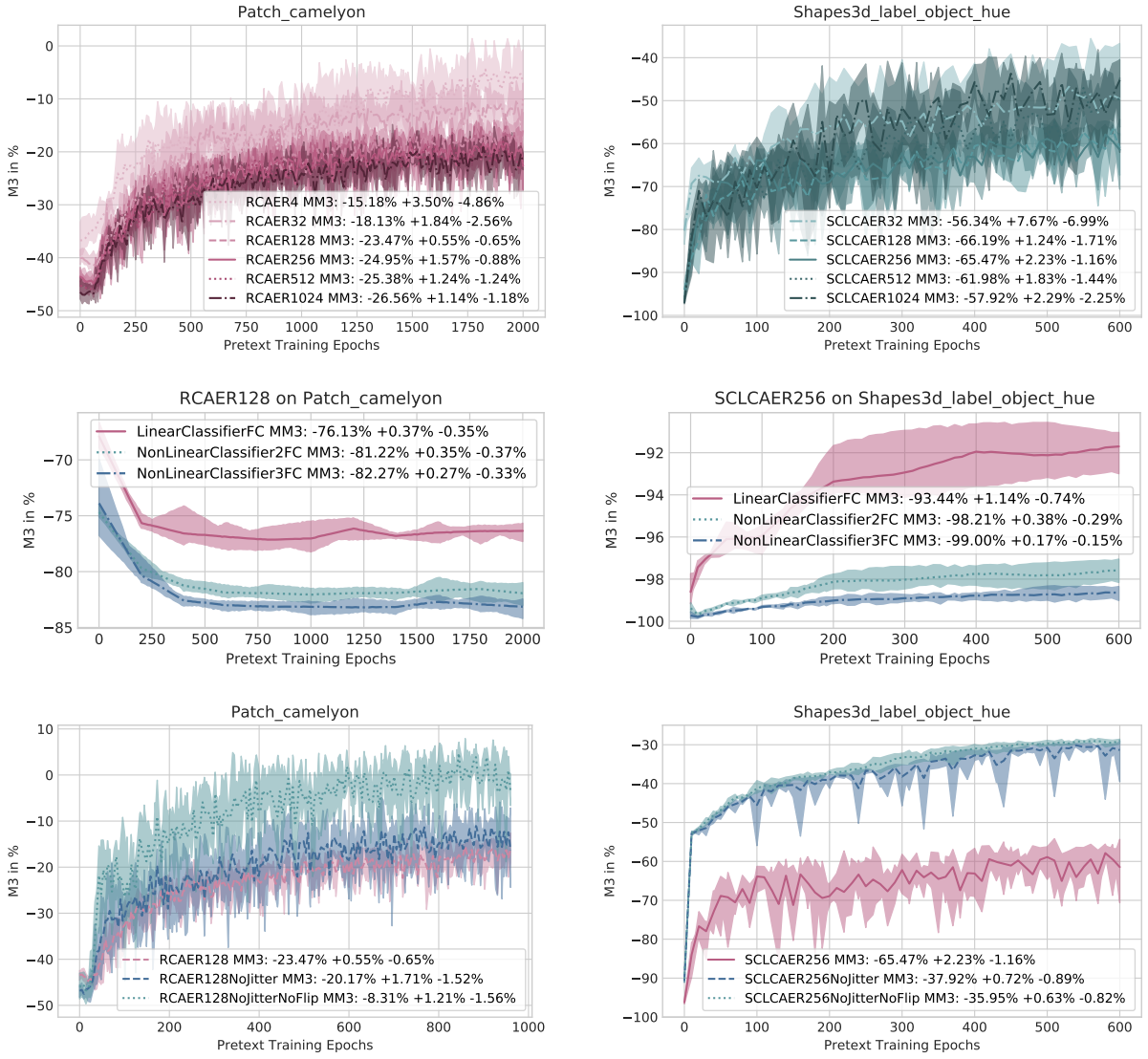

Fig. 13 Version of Figure 4 for M3 on accuracies. We observe that besides early spikes, M3 decreases when we add complexity to the target model. Additionally, we observe that M3 decreases when we add augmentations in this case. We measure the M3 for RCAE between the classification error of the target task and the classification error of predicting the rotations of rotated images from PCam (pretext task). For SCLCAE the pretext task metric measures the ability of the model to correctly detect the representation of each given image in a batch of representations of transformed images. Here we show M3 without a convergence criterium.

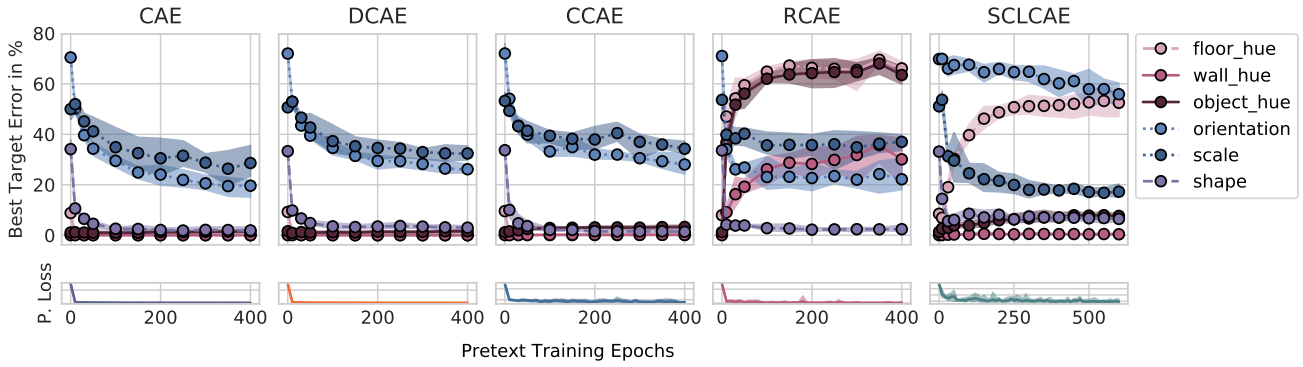

Fig. 14 Version of Figure 5 for SM3 on accuracies.

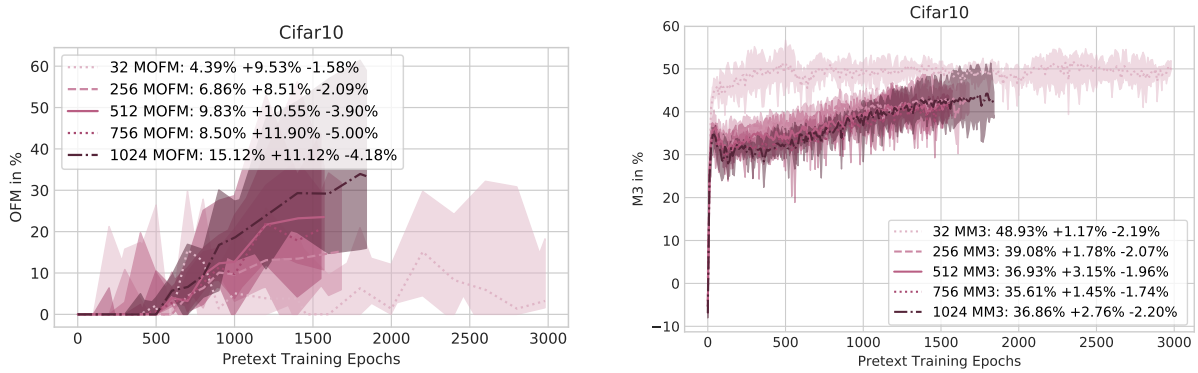

Fig. 15 (Left) The OFM of ResNets with different representation sizes trained on the pretext task of predicting rotations on Cifar10. Target models are trained for Cifar10 classification. (Right) MM3 of those ResNets. We vary representation sizes in $[32, 256, 512, 756, 1024]$ by adding a 1×1 convolution layer on top of the ResNet18. Thereby the number of filters corresponds to the representation size. In contrast to our observations on our small model the largest representation we tested leads to a high OFM. A reason for that could be that the larger representation size helps the model to solve the pretext task and since there is a mismatch with the target task, a better understanding of this task leads to a higher mismatch. We note that a representation size of 1024 is still very small for unsupervised learning. An even larger representation size could therefore still lead to a lower mismatch.

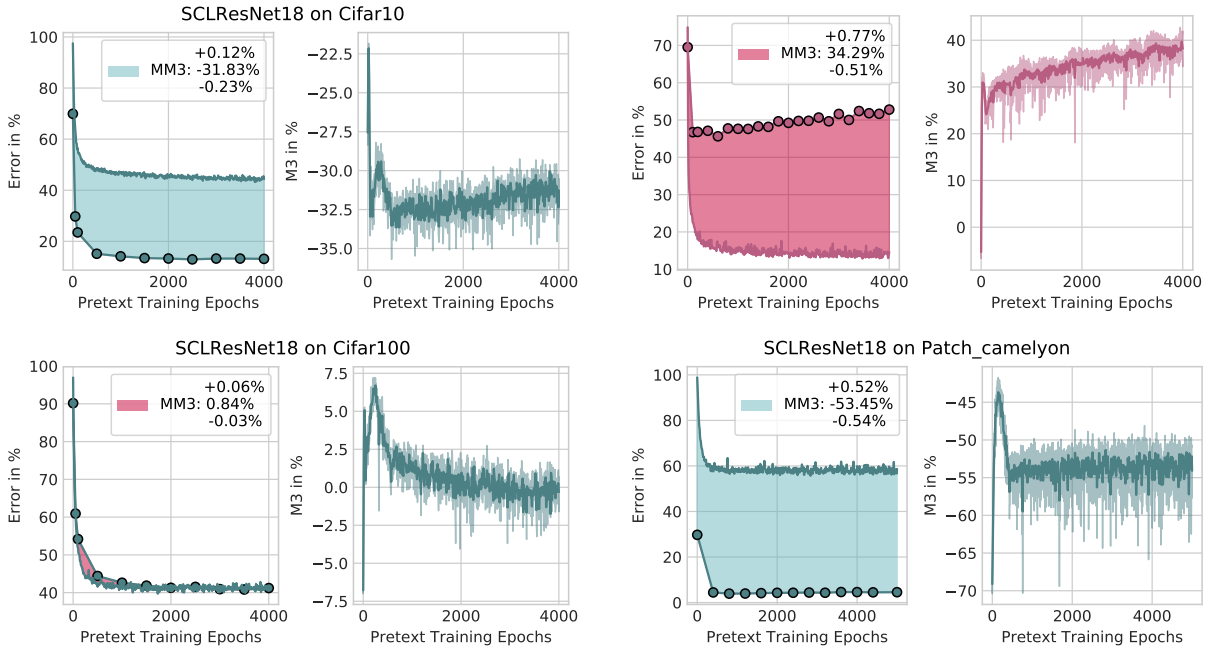

Fig. 16 MM3 for different pretext tasks trained with a ResNet18 model as backbone. The mismatches are shown for the entire training process. In contrast to the prediction of rotations (top right), SCLResNet18 has a high negative MM3 for Cifar10. This indicates that learning the contrastive pretext task is better suited for distinguishing Cifar10 classes then the prediction of rotations. Furthermore, the error of the constrastive pretext task is significant, which indicates that the model still underfits the pretext task with this setup and there is more room for improvement. For the 100 classes of Cifar100, MM3 becomes slightly positive in the contrastive learning setup. For contrastive learning on the PCam dataset and rotation prediction on Cifar10, we observe an increasing mismatch during training.