Supplementary Material

A. Details of the experimental protocol

Experiments on the KITTI dataset. We have discussed in Section 5.1 of the main paper three different protocols ($P1$, $P2$, $P3$) to evaluate lifelong learning. Each protocol corresponds to a sequence of conditions (e.g. Clean→Foggy→Cloudy for P1) and uses a different urban environment sequence for each condition, which we refer to as A, B, and C in the paper. For each protocol, we train models on 11 different permutations of A, B and C, which we list below for reproducibility (following KITTI's notation [3]), and report mean and std results.

1. Scene-02 → Scene-01 → Scene-06
2. Scene-06 → Scene-01 → Scene-18
3. Scene-20 → Scene-01 → Scene-18
4. Scene-02 → Scene-18 → Scene-20
5. Scene-06 → Scene-01 → Scene-02
6. Scene-20 → Scene-18 → Scene-01
7. Scene-02 → Scene-06 → Scene-01
8. Scene-18 → Scene-20 → Scene-02
9. Scene-20 → Scene-06 → Scene-01
10. Scene-18 → Scene-06 → Scene-02
11. Scene-06 → Scene-20 → Scene-18

B. Transformation sets used for the auxiliary meta-domains

We report in the Table 12 of this supplementary how the transformation sets used for our experiments in Section 5 of the main paper are built. We indicate as Ψ_1, Ψ_2, and Ψ_3 the sets used for the digits/PACS experiments (as in Section 5.1), and as Ψ_4 the set used for the semantic segmentation experiments on KITTI. For the description of a single transformation, we refer to the documentation of the PIL library [45] which is the one we used (see in particular [43, 44])—with the exception of *Invert*, *Gaussian noise* and *RGB-rand*. For these three last transformations, we give their details below. Given an RGB image $\mathbf{x}$ with pixels in range $[0, 255]$:

- **Invert** applies the transformation $\hat{\mathbf{x}} = |\mathbf{x} - 255|$.
- **Gaussian noise** perturbs pixels with values that are sampled from a Gaussian distribution with standard deviation σ defined by the chosen level.
- **RGB-rand** perturbs the pixels of each channel by adding factors r, g, b, each sampled from a uniform distribution defined in [-level, +level].

C. Domain randomization improves domain generalization performance

In Section 1 we presented domain randomization as a means to increase robustness of the model at hand in out-of-domain contexts—and, in turn, lighten the adaptation process and mitigating the catastrophic forgetting. We report in Table 4 and 5 of this supplementary respectively out-of-domain performance of models trained on MNIST [32] and on the Sketch domain (from PACS [33], see Figure 5), with and without domain randomization (relying on transformation set Ψ_2 when using domain randomization). Similar results for digits were also shown in previous work [60]. We would like to stress that this protocol is different from the ones used to carry out the experiments in the main manuscript; we are not assessing continual learning performance in this Appendix, but out-of-domain performance of models trained on a single domain (MNIST [32] and Sketches [33]). This experiment only serves as a support to our motivation for using domain randomization, expressed in Section 1.

Domain generalization MNIST models

	MNIST-M	SYN	SVHN
w/o DR	41.2 ± 1.3	35.1 ± 0.6	23.5 ± 1.6
w/ DR	65.6 ± 5.1	53.7 ± 2.4	40.4 ± 1.3

Table 4. Performance of models trained on MNIST [32] when tested on MNIST-M [19], SYN [19] and SVHN [40]. First and second row report results of models trained without and with domain randomization, respectively. These results are related to models trained on a single domain, hence they are not comparable with the ones from the main manuscript.

Domain generalization Sketches models

	Cartoons	Paintings	Photos
w/o DR	31.3 ± 3.0	24.4 ± 4.3	31.1 ± 4.3
w/ DR	48.3 ± 4.6	28.5 ± 7.7	36.8 ± 5.7

Table 5. Performance of models trained on the Sketches domain when tested on Cartoons, Paintings and Photos domains (from PACS [33]). First and second row report results of models trained without and with domain randomization, respectively. These results are related to models trained on a single domain, hence they not comparable with the ones from the main manuscript.

D. Additional experiments

We report in Tables 6, 7, and 8 additional results associated with protocol $P1$ of the digits experiments. We report in Table 9 additional results associated with protocol $P3$ of the semantic segmentation experiment on KITTI. All results in Tables 6– 9 are referred to the *Meta-DR* method.

Digits experiment: hyper-parameter β

	Training Protocol: P1			
	MNIST (1)	**MNIST-M (2)**	**SYN (3)**	**SVHN (4)**
$\beta = 0.0$	83.7 ± 6.4	68.8 ± 3.4	92.3 ± 0.4	86.9 ± 0.1
$\beta = 0.1$	90.6 ± 2.5	73.7 ± 1.6	93.6 ± 0.1	87.9 ± 0.0
$\beta = 1.0$	94.3 ± 0.7	76.5 ± 0.6	94.4 ± 0.0	89.5 ± 0.2

Table 6. Performance of models trained with *Meta-DR* with different values for β ($\gamma = 0.0$). Results averaged over 3 runs, and models trained using Ψ_3. Performance evaluated on all domains at the end of the training sequence $P1$.

Digits experiment: hyper-parameter γ

	Training Protocol: P1			
	MNIST (1)	**MNIST-M (2)**	**SYN (3)**	**SVHN (4)**
$\gamma = 0.0$	83.7 ± 6.4	68.8 ± 3.4	92.3 ± 0.4	86.9 ± 0.1
$\gamma = 0.1$	91.5 ± 1.3	76.5 ± 0.7	94.8 ± 0.3	89.7 ± 0.5
$\gamma = 1.0$	89.7 ± 0.5	74.6 ± 0.1	95.4 ± 0.1	91.9 ± 0.0

Table 7. Performance of models trained with *Meta-DR* with different values for γ ($\beta = 0.0$). Results averaged over 3 runs, and models trained using Ψ_3. Performance evaluated on all domains at the end of the training sequence $P1$.

Digits experiment: hyper-parameter α

	Training Protocol: P1			
	MNIST (1)	**MNIST-M (2)**	**SYN (3)**	**SVHN (4)**
$\alpha = 0.001$	85.5 ± 1.6	70.7 ± 0.7	94.5 ± 0.3	91.1 ± 0.0
$\alpha = 0.01$	87.1 ± 1.1	72.7 ± 0.5	95.1 ± 0.1	91.5 ± 0.0
$\alpha = 0.1$	92.0 ± 0.6	75.1 ± 0.5	95.4 ± 0.3	91.9 ± 0.2

Table 8. Performance of models trained with *Meta-DR* with different values for the meta-learning rate α ($\beta = \gamma = 1.0$). Results averaged over 3 runs, and models trained using Ψ_3. Performance evaluated on all domains at the end of the training sequence $P1$.

Sem. segm. experiment: hyper-parameter β

	Training Protocol: P3		
	Clone (1)	**Sunset (2)**	**Morning (3)**
$\beta = 0.0$	60.3 ± 11.5	63.6 ± 7.7	76.0 ± 10.0
$\beta = 0.001$	62.3 ± 9.2	67.1 ± 8.6	73.8 ± 9.2
$\beta = 0.01$	61.7 ± 9.4	65.8 ± 7.0	73.8 ± 9.9
$\beta = 0.1$	61.6 ± 11.0	67.1 ± 7.7	74.9 ± 8.2
$\beta = 1.0$	65.4 ± 5.3	68.1 ± 3.7	74.5 ± 3.7
$\beta = 10.0$	64.1 ± 7.6	66.6 ± 6.9	73.8 ± 8.4

Table 9. Performance (mIoU) of models trained with *Meta-DR* with different values for β ($\gamma = 0.0$). Results averaged over 10 permutations of urban environments. Performance evaluated on all domains at the end of the training sequence $P3$.

We extend the results reported in Table 1 in the main manuscript by testing different values for the memory size and further comparison against GEM [37]; these are reported in Table 10, for the protocol $P1$. Note that all methods were implemented with SGD optimizer here (learning rate $\eta = 0.01$), for comparability. We further report in Table 11 results obtained by averaging over the 24 possible digit permutations.

Digits experiment: memory size

Methods	**M. size**	**MNIST(1)**	**MNIST-M(2)**	**SYN(3)**	**SVHN(4)**
GEM [37]	200	93.77 ± 0.8	75.68 ± 1.1	93.51 ± 0.3	84.58 ± 1.1
	300	94.51 ± 0.7	76.37 ± 1.5	93.68 ± 0.4	84.84 ± 1.1
	400	95.19 ± 0.4	77.09 ± 0.9	93.86 ± 0.3	85.16 ± 0.5
GEM + DR	200	93.59 ± 0.5	76.34 ± 1.2	95.80 ± 0.2	89.82 ± 0.6
	300	93.81 ± 0.7	77.66 ± 0.6	95.65 ± 0.3	89.86 ± 0.6
	400	94.23 ± 0.8	77.83 ± 1.2	95.81 ± 0.2	89.96 ± 0.5
ER [6]	200	95.78 ± 0.3	79.88 ± 0.5	93.23 ± 0.2	86.29 ± 0.4
	300	96.41 ± 0.3	81.32 ± 0.5	93.50 ± 0.2	86.20 ± 0.4
	400	96.63 ± 0.3	82.07 ± 0.5	93.69 ± 0.2	86.43 ± 0.2
ER + DR	200	95.52 ± 0.5	82.54 ± 0.7	95.74 ± 0.2	89.96 ± 0.4
	300	95.63 ± 0.4	84.26 ± 0.7	95.94 ± 0.1	90.02 ± 0.3
	400	96.45 ± 0.3	85.50 ± 0.3	95.88 ± 0.2	89.94 ± 0.3
ER + *Meta-DR*	200	96.05 ± 0.4	84.19 ± 0.6	96.42 ± 0.1	91.46 ± 0.2
	300	96.64 ± 0.4	85.66 ± 0.4	96.56 ± 0.1	91.40 ± 0.2
	400	97.12 ± 0.3	86.81 ± 0.3	96.73 ± 0.2	91.75 ± 0.2

Table 10. Comparison between models trained via GEM [37] and ER [6] algorithms, with and wihout DR, and *Meta-DR*. Memory size is varied from 200 to 400 samples. For comparability, all models were trained using the SGD optimizer, as performed in the PACS experiments in the main manuscript. For what concerns the episodic memory, the number of samples per domain is indicated in the 2nd column.

Digits experiment: 24 permutations

Methods	**MNIST**	**MNIST-M**	**SYN**	**SVHN**
GEM [37]	96.48(2.1)	81.53(6.4)	90.09(5.5)	78.16(5.8)
GEM [37] + DR	96.09(2.7)	83.45(7.6)	90.86(6.5)	83.01(6.1)
ER [6]	97.23(1.3)	84.65(3.7)	92.49(2.5)	82.53(2.8)
ER [6] + DR	97.04(1.4)	86.31(4.2)	94.77(1.9)	87.01(2.3)
ER [6] + *Meta-DR*	97.67(1.1)	87.94(4.0)	95.61(1.6)	88.82(2.0)

Table 11. Average results for the 24 possible digit permutations that can be obtained from the set of available domains {MNIST, MNIST-M, SYN, SVHN}. For what concerns the episodic memory, the number of samples per domain is set to 100.

Image transformations (for auxiliary meta-domains or data augmentation)						
			Set Ψ			
Transformations	**Range**	**No. Levels**	Ψ_1	Ψ_2	Ψ_3	Ψ_4
Brightness	$[0.2, 1.8]$	90	✓	✓	✓	✓
Color	$[0.2, 1.8]$	90	✓	✓	✓	✓
Contrast	$[0.2, 1.8]$	90	✓	✓	✓	✓
RGB-rand	$[1, 120]$	90				✓
Solarize	$[255, 75]$	90	✓	✓	✓	
Grayscale	–	1	✓	✓	✓	
Invert	–	1	✓	✓	✓	
Rotate	$[-60, 60]$	30		✓	✓	
Gaussian noise	$[0.0, 30.0]$	30			✓	
Blur	–	1			✓	
Number of transformations N			2	2	2	2

Table 12. Details of the different transformation sets applied to images, which are either used to create the auxiliary meta-domains or for data augmentation.

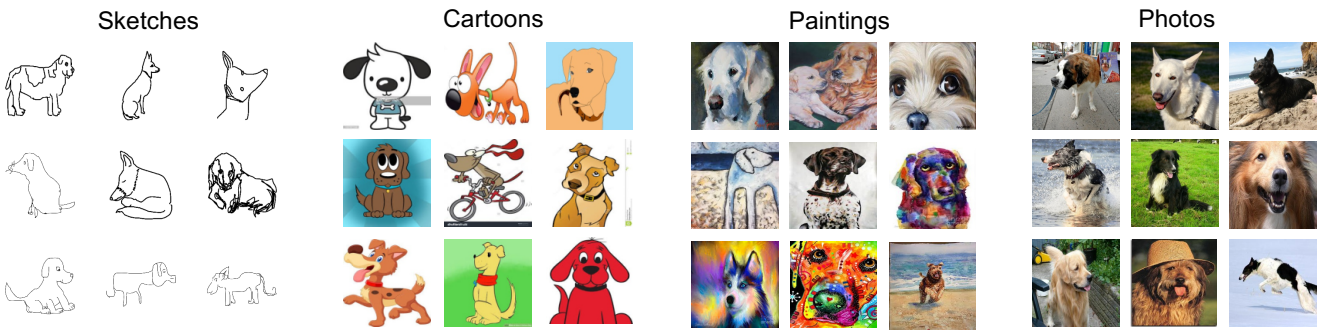

Figure 5. Samples from the 'dog' class of PACS dataset [33]