The Change You Want To Detect: Semantic Change Detection In Earth Observation With Hybrid Data Generation

Supplementary Material

In this appendix, we provide implementation details in Sec. A, details about the deep architecture used in Sec. C and detailed quantitative results in Sec. D. Finally, we provide qualitative samples in Sec. E.

A. Implementation Details

A.1. Training of Stable Diffusion

For the VAE training, the color loss was applied only after a sufficient amount of 50,000 iterations. Fine-tuning the VAE took 160 hours A100 GPU for a total of 500,000 iterations with a batch size of 4. The diffusion UNet training required 300 hours A100 GPUs, for a total of 30,000 iterations with a batch size of 32. The training was run on 4 GPU in parallel. Finally, we trained the ControlNet for 240 hours on a A100 GPU, performing 45,000 iterations with a batch size of 16. We also observed the "sudden convergence phenomenon" mentioned in [65].

A.2. The `HySCDG` pipeline details

In the `HySCDG` pipeline, for a given image with n instances, we randomly select N_{change} instances (from 0 to $min(3, n)$) to inpaint with a semantic class change and $N_{nochange}$ (from 0 to $min(3, n)$) random shapes for inpainting without semantic class change. We draw the number of instances and shapes from the following heuristical laws (Eqs. (A) and (B)):

$$N_{change} \sim \left\lfloor \sqrt{\mathcal{U}(0, 10)} \right\rfloor, \tag{A}$$

$$N_{nochange} \sim \left\lfloor \sqrt{\mathcal{U}(0, 10) \times (1 - \frac{N_{change}}{4})} \right\rfloor. \tag{B}$$

The frequency of the inpainted instances and random shapes can be seen in Fig. B function to n the number of instances within the image.

A.3. FLAIR Dataset [18]

The FLAIR dataset [18] is composed of 77,762 VHR aerial patches 512 × 512 at 0.20m GSD. The VHR images include five channels: Red, Green, Blue, near-infrared, and a normalized Digital Surface Model derived by dense image matching (*normalization* means the altitude of the terrain is removed). For each patch, ground truth semantic segmentation is provided. The semantic map represents the land cover based on a 19 classes nomenclature (building, coniferous, deciduous, etc. details provided in Fig. 2). We only use the 16 main classes due to the scarcity of certain classes or potential ambiguity. This dataset covers approximately 817km^2 extracted from various areas in France.

A.4. Training computing times

Models were pretrained to convergence on the synthetic datasets, requiring 40 to 60 hours on a A100 GPU, depending on the configuration. For instance, Dual UNet converges faster than SCanNet, multi-task learning takes longer than binary-only, and the larger size of FSC-180k compared to SyntheWorld requires additional iterations.

For the sequential scenario, fine-tuning was done in 10 to 20 hours (V100 GPU) for HiUCD-mini, Levir-CD, S2Looking and SECOND (in increasing order) and 20 hours (A100 GPU) for HiUCD-XL. For mixed training, all trainings took from 16 to 20 hours (A100 GPU). Note that, in low data regime, each experiment was repeated 10 times, except for HiUCD-XL for which it was only 3 times for 10% and 30% experiments.

B. Failure cases of the `HySCDG` synthesis

ControlNet is the most error-prone module as it may fail to respect the semantic map. Other elements are more reliable or less sensitive: instance footprints of high quality, low influence of text prompt, high-quality inpainting from SD alone (Figure C). In this same figure, we show results with only the ControlNet module fine-tuned for generating aerial images from semantic maps, revealing a cartoonish style.

C. Pretraining and transfer experiments

C.1. Transfer Learning Scenarios

The four transfer learning scenarios considered in this work are illustrated on Fig. A. They are quantitatively detailed in Section D.

C.2. Model architectures

Our Dual UNet architecture consists of two parallel UNet-style auto-encoders: one dedicated to semantic segmentation for each image separately, and the other one for detecting changes by processing concatenated image pairs. Both UNets share the same core configuration, differing only in their input and output layers: the change detection UNet takes two concatenated images as input and outputs two classes (no-change, change), while the semantic segmentation UNet processes single images and outputs the number of semantic classes (of the dataset). Both networks use a ResNet-50 encoder pretrained on the ImageNet dataset. To enhance

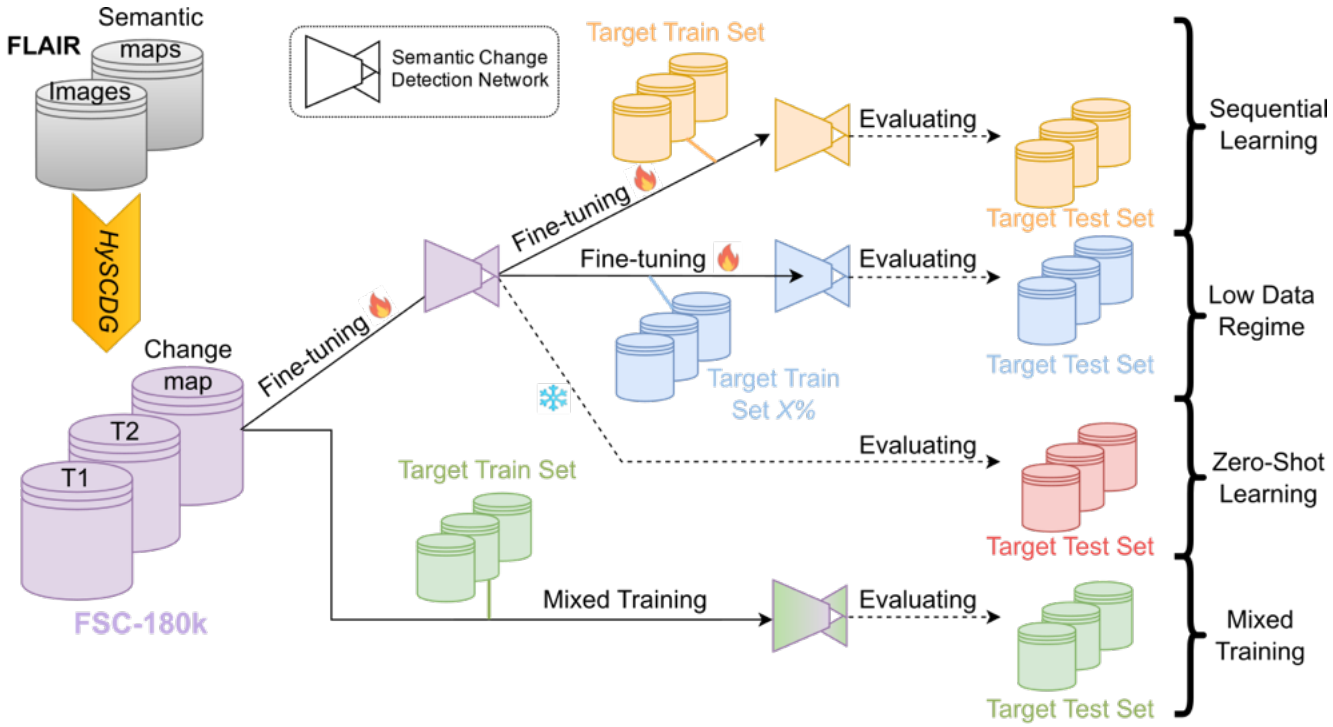

Figure A. **Transfer Learning scenarios.** We illustrate here the various configurations tested and described in Sec. 4.1.

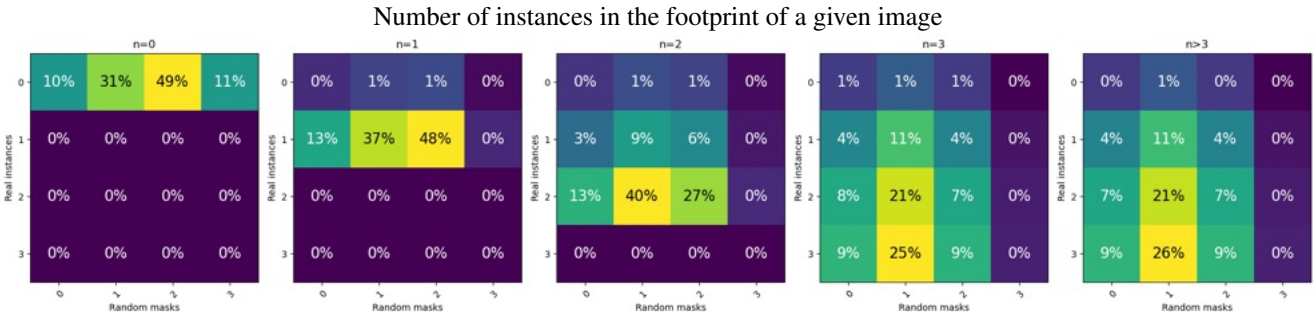

Figure B. **Heuristic rule to define instance and random mask numbers.** We illustrated the frequency of the number of instances inpainted (rows) and the number of random shapes inpainted (columns) with respect to the number n of instances contained in the footprint of the VHR images.

Figure C. Illustrations of failure cases of synthesis with our pipeline HySCDG, and synthesis with only the ControlNet trained.

performance, features extracted at each encoding step of the semantic UNet are injected into the decoding pathway of the change detection UNet, in addition to its own skip connection features. The entire model comprises a total of 83 million parameters.

We compared our architecture with state-of-the-art change detection models that rely on different mechanisms: MambaCD [7], SCanNet [13], and ChangeFormer [2]. The latter, ChangeFormer was only evaluated through pretraining on FSC-180k, which proved unsuccessful due to the model's excessive size and the significant computational resources required for training. MambaCD and SCanNet were evaluated on SECOND and HiUCD-mini. The performances of the different models can be found in Tab. A. SCanNet was better in all cases and was therefore kept for comparison in sequential experiments, with FSC-180k pretraining.

Table A. **Comparison of model architectures** The different models were evaluated on SECOND and HiUCD-mini after being trained from scratch for the same duration. **Bold** values correspond to the best values for each metric and each dataset.

Target	Model	IoU	SeK
SECOND	MambaCD	0.52	0.14
	SCanNet	**0.54**	**0.19**
	Simple UNet	0.51	0.11
	Dual UNet	0.53	0.16
HiUCD mini	MambaCD	0.53	0.08
	SCanNet	**0.62**	0.11
	Simple UNet	0.50	0.07
	Dual UNet	0.61	**0.17**

C.3. Normalization

We compared the effects of applying or not normalization by mean and variance of the train set pixel's values on the input data. It proved to be really effective. For example, this improves SCanNet's performance on the SECOND dataset from 0.17 to 0.19 in Separated Kappa. We used it in all our experiments, using the parameters of the handled dataset. Namely, in sequential mode, normalization uses the synthetic dataset's parameters during pretraining, and the target dataset parameters during fine-tuning.

Table B. **Effect of the normalization.** The impact of input normalization was evaluated on the SECOND dataset. For both SCanNet and Dual UNet, normalization was highly effective.

Model	Normalization	IoU	SeK
SCanNet		0.52	0.17
	X	**0.54**	**0.19**
Dual UNet		0.51	0.14
	X	**0.53**	**0.16**

D. Quantitative results

In this section, we present in details the quantitative results for both binary and semantic cases, for the four transfer learning scenarios. We follow the same order as in the paper.

- **Sequential Training**: binary change detection (BCD) results are provided in Tab. C, while semantic change detection (SCD) outputs are given in Tab. G.
- **Low-Data Regime**: results for BCD can be found in Tab. D for Levir-CD et S2Looking, and in Tab. H for SECOND and HiUCD-mini for the SCD configuration.
- **Mixed Training**: Tab. E and Tab. I provide numbers, respectively, for BCD for the four datasets and SCD for SECOND and HiUCD-mini datasets. Tab. J focuses on HiUCD-XL dataset.
- **Zero-Shot Learning**: results for BCD and SCD are given in Tab. F and Tab. K respectively.

D.1. Binary change detection

Table C. **Sequential Training Evaluation (Binary Change Detection).** We report the F1 and IoU scores on four datasets. Models were initially pretrained (SyntheWorld or FSC-180k) or not (Baseline) and then fine-tuned on the target dataset. Pretraining effects remain limited on simpler datasets such as Levir-CD, but provide significant benefits for more complex datasets like HiUCD, which is also closer to our FSC-180k in terms of landscapes and characteristics. **Bold** values correspond to the best values for each metric and each dataset.

	Target → Pretraining ↓	*Levir-CD*	*S2Looking*	*SECOND*	*HiUCD mini*
F1	Baseline	**0.91**	0.61	0.70	0.75
	SyntheWorld	**0.91**	**0.63**	**0.71**	0.73
	FSC-180k	**0.91**	**0.63**	**0.71**	**0.77**
IoU	Baseline	0.83	0.44	0.53	0.61
	SyntheWorld	**0.84**	**0.46**	0.54	0.58
	FSC-180k	**0.84**	**0.46**	**0.55**	**0.63**

Table D. **Low Data Regime (Binary Change Detection).** We report the IoU scores on two binary datasets. Models were initially pretrained (SyntheWorld or FSC-180k) or not (Baseline) and then fine-tuned on a portion of the target dataset (X%). Pretraining offers significant advantages in scenarios with limited data, making it particularly valuable in many real-world applications. **Bold** values correspond to the best values for each metric and each dataset.

Target Percent	1 %	10 %	30 %
Levir-CD			
Baseline	0.36	0.69	0.75
SyntheWorld	0.41 (+14%)	0.71 (+3%)	0.77 (+3%)
FSC-180k	**0.55** (+53%)	**0.73** (+6%)	**0.79** (+5%)
S2Looking			
Baseline	0.10	0.27	0.38
SyntheWorld	0.08 (-20%)	0.34 (+26%)	0.41 (+8%)
FSC-180k	**0.15** (+50%)	**0.36** (+33%)	**0.43** (+13%)

Table E. **Mixed Training Evaluation (Binary Change Detection).** Model are trained on a blend of target and synthetic/hybrid (SyntheWorld or FSC-180k) train sets, containing a ratio of x% samples from the target (including repetitions). Testing is performed on the target test set. The last column corresponds to train exclusively on target dataset (without pretraining). We report the F1 and IoU scores on four datasets. Mixed training turns out to be at least as effective as sequential one. **Bold** values correspond to the best values for each metric and each dataset.

Target		Pretraining	20%	50%	90%	100%
Levir-CD	F1	SyntheWorld	0.90	0.90	**0.91**	0.90
		FSC-180k	0.90	**0.91**	**0.91**	
	IoU	SyntheWorld	0.81	0.83	**0.84**	0.83
		FSC-180k	0.82	0.83	**0.84**	
S2Looking	F1	SyntheWorld	0.60	0.63	0.62	0.61
		FSC-180k	0.61	**0.64**	0.63	
	IoU	SyntheWorld	0.43	0.48	0.48	0.44
		FSC-180k	0.44	**0.49**	**0.49**	
SECOND	F1	SyntheWorld	0.71	0.72	0.71	0.70
		FSC-180k	0.72	**0.73**	0.72	
	IoU	SyntheWorld	0.54	0.56	0.55	0.53
		FSC-180k	**0.57**	**0.57**	0.55	
HiUCD mini	F1	SyntheWorld	0.71	0.72	0.72	0.75
		FSC-180k	0.76	0.76	**0.77**	
	IoU	SyntheWorld	0.58	0.58	0.59	0.61
		FSC-180k	0.61	0.62	**0.63**	

Table F. **Zero-Shot Evaluation (Binary Change Detection).** We report the F1 and IoU scores for the zero-shot case. Metrics for SECOND and HiUCD are available on Tab. K. Models were initially pretrained (SyntheWorld or FSC-180k). Decent performance was obtained on Levir-CD, despite the domain gap between FSC-180kand Levir-CD. However, pretraining remained completely ineffective for S2Looking. **Bold** values correspond to the best values for each metric and each dataset.

Target	Pretraining	F1	IoU
Levir-CD	SyntheWorld	0.25	0.13
	FSC-180k	**0.49**	**0.33**
S2Looking	SyntheWorld	0.0	0.0
	FSC-180k	0.04	0.02

D.2. Semantic change detection

Table G. **Sequential Training Evaluation (Semantic Change Detection).** We report the binary and semantic scores on three datasets. Models (based on the Dual UNet) were initially pretrained (SyntheWorld or FSC-180k) or not (Baseline), and, then, fine-tuned on the target dataset. Pretraining on FSC-180k proved highly effective and outperformed SyntheWorld on every semantic dataset, with greater benefits observed on HiUCD due to the higher similarity between the datasets. **Bold** values correspond to the best values for each metric and each dataset.

Target	Pretraining	IoU	Ovr. IoU	SeK	SCS
SECOND	Baseline	0.53	0.64	0.16	0.83
	SyntheWorld	0.54	0.63	0.17	0.84
	FSC-180k	**0.55**	**0.65**	**0.18**	**0.89**
HiUCD mini	Baseline	0.61	0.77	0.17	0.73
	SyntheWorld	0.58	0.78	0.15	0.73
	FSC-180k	**0.63**	**0.79**	**0.19**	**0.78**
		Sem. mIoU	Chg. mIoU	Bin. mIoU	Bin. C. mIoU
HiUCD XL	Baseline	0.58	0.17	0.34	**0.48**
	SyntheWorld	0.58	0.17	0.34	0.48
	FSC-180k	**0.60**	**0.19**	0.34	**0.48**

Table H. **Low Data Regime (Semantic Change Detection).** Models were initially pretrained (SyntheWorld or FSC-180k) or not (Baseline) and then fine-tuned on a portion of the target dataset (x%). The benefit of pretraining increases as the amount of available data decreases. These experiments highlight the relevance of pretraining on our FSC-180k in real-world cases where annotated data is often scarce and/or costly to obtain. **Bold** values correspond to the best values for each metric and each dataset.

Target Percent	1 %	10 %	30 %
SECOND (scores in SCS)			
Baseline	0.31	0.49	0.69
SyntheWorld	0.37 (+18%)	0.49 (+0%)	0.64 (-6%)
FSC-180k	**0.40** (+29%)	**0.57** (+15%)	**0.74** (+8%)
HiUCD-mini (scores in SCS)			
Baseline	0.13	0.41	0.48
SyntheWorld	0.17 (+31%)	0.38 (-7%)	0.50 (+4%)
FSC-180k	**0.20** (+38%)	**0.49** (+19%)	**0.54** (+13%)
HiUCD-XL (scores in Chg. mIoU)			
Baseline	0.06	0.10	0.12
SyntheWorld	0.05 (-16%)	0.10 (+0%)	0.11 (-8%)
FSC-180k	**0.08** (+33%)	**0.13** (+30%)	**0.15** (+25%)

Table I. **Mixed Training Evaluation (Semantic Change Detection).** Model are trained on a blend of target and synthetic/hybrid (SyntheWorld or FSC-180k) train sets, containing a ratio of x% samples from the target (including repetitions). Testing is performed on the target test set. The last column corresponds to train exclusively on target dataset (without pretraining). Mixed training turns out to be more effective than sequential training, thanks to the exposure of the model to diverse and varied examples throughout the training process. **Bold** values correspond to the best values for each metric and each dataset.

Target		Pretraining	20%	50%	90%	100%
SECOND	IoU	SyntheWorld	0.54	0.56	0.55	0.53
		FSC-180k	0.57	**0.57**	0.55	
	Ovr. IoU	SyntheWorld	0.61	0.62	0.63	0.64
		FSC-180k	**0.64**	0.63	0.62	
	SeK	SyntheWorld	0.16	0.17	0.16	0.16
		FSC-180k	**0.19**	0.18	0.18	
	SCS	SyntheWorld	0.83	0.84	0.85	0.83
		FSC-180k	0.88	**0.89**	0.88	
HiUCD mini	IoU	SyntheWorld	0.58	0.58	0.59	0.61
		FSC-180k	0.61	0.62	**0.63**	
	Ovr. IoU	SyntheWorld	0.74	**0.78**	0.77	0.77
		FSC-180k	0.76	0.77	0.77	
	SeK	SyntheWorld	0.11	0.13	0.14	0.17
		FSC-180k	0.16	0.17	**0.18**	
	SCS	SyntheWorld	0.68	0.71	0.72	0.73
		FSC-180k	0.74	0.76	**0.78**	

Table J. **Mixed Training Evaluation on HiUCD-XL.** Training on a blend of target and synthetic/hybrid (SyntheWorld or FSC-180k) train sets, containing a ratio of x% samples from the target (including repetitions). Testing is performed on the target test set. 100% corresponds to fine-tuning exclusively on target dataset (without pretraining). 0% corresponds to zero-shot (after pretraining). An optimal mix ratio appears to be around 50%, which can be understood as a good compromise between exploration (pretraining data) and exploitation (target data). **Bold** values correspond to the best values for each metric.

Pretraining	Mix ratio	Sem. mIoU	Chg. mIoU	Bin. mIoU	Bin. C. mIoU
SyntheWorld	0%	0.02	0.03	0.0	0.47
	20%	0.51	0.10	0.33	0.48
	50%	0.62	0.19	**0.37**	0.51
	90%	0.60	0.17	0.32	0.47
FSC-180k	0%	0.35	0.07	0.20	0.43
	20%	0.58	0.15	0.32	0.51
	50%	**0.63**	**0.20**	0.29	**0.52**
	90%	0.62	0.18	0.30	0.51
-	100%	0.58	0.17	0.34	0.48

Table K. **Zero-Shot Evaluation (Semantic Change Detection).** Models were initially pretrained (SyntheWorld or FSC-180k). In semantic mode, only FSC-180k pretraining enables the model to detect changes and predict semantic segmentation. **Bold** values correspond to the best values for each metric and each dataset.

Target	Pretraining	F1	IoU	SCS
SECOND	SyntheWorld	0.02	0.01	0.0
	FSC-180k	**0.53**	**0.36**	**0.24**
HiUCD mini	SyntheWorld	0.02	0.01	0.0
	FSC-180k	**0.53**	**0.36**	**0.25**
		Sem. mIoU	Chg mIoU	Bin. mIoU
HiUCD XL	SyntheWorld	0.02	0.03	0.0
	FSC-180k	**0.35**	**0.07**	**0.20**

E. Qualitative results

In Fig. D, we provide some examples of images generates with HySCDG, that we extracted from our dataset FSC-180k. One can see the joint spatial and semantic accuracy of the images and maps, coupled with diverse and real-world change configurations. The random selection of instances allows to provide multiple change trajectories and not to focus on main classes (namely, *Building*, *Impervious surfaces*, and *Bare soil*). The controlled inpainting is fairly realistic, for example with the addition of an agricultural field to replace a sport ground at the top right of the first line image.

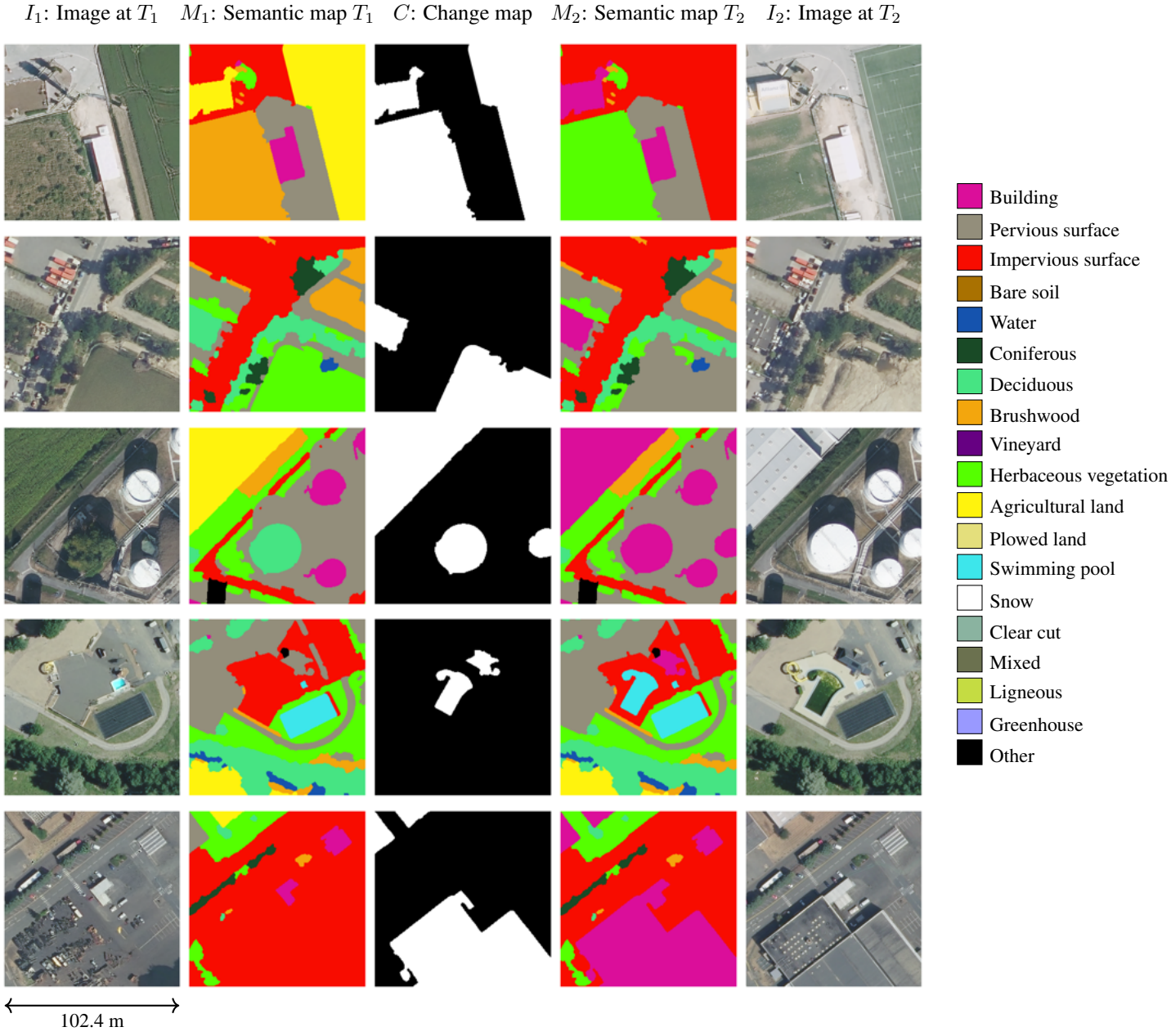

Figure D. **FSC-180k samples.** Hybrid change pair images from our pretraining dataset, generated by HySCDG. The first column T_1 corresponds to the unpainted images, the second to the corresponding semantic map M_1, the middle column provides the change map, whereas the two last columns are the FLAIR samples used for generating the three first elements from the row (VHR image and semantic map). Best viewed in color.

Fig. E and Fig. F provide several examples of semantic prediction results with our Dual UNet architecture for SECOND and HiUCD-mini datasets, respectively. Experiments were performed in sequential scenario: without synthetic nor hybrid training data (Baseline), pretraining on SyntheWorld, and pretraining on FSC-180k.
One can first see that the Dual UNet architecture alone is sufficient for reliable dense predictions on VHR images, validating our quantitative assessment related to model architectures (Tab. A). Secondly, it can be noticed that working with FSC-180k allows for more consistent results over various classes, landscapes, and datasets. It is also worth noting that the GSD varies between the pretraining dataset (0.2m) and the target datasets (0.1m for HiUCD and approximately 0.45m for SECOND).

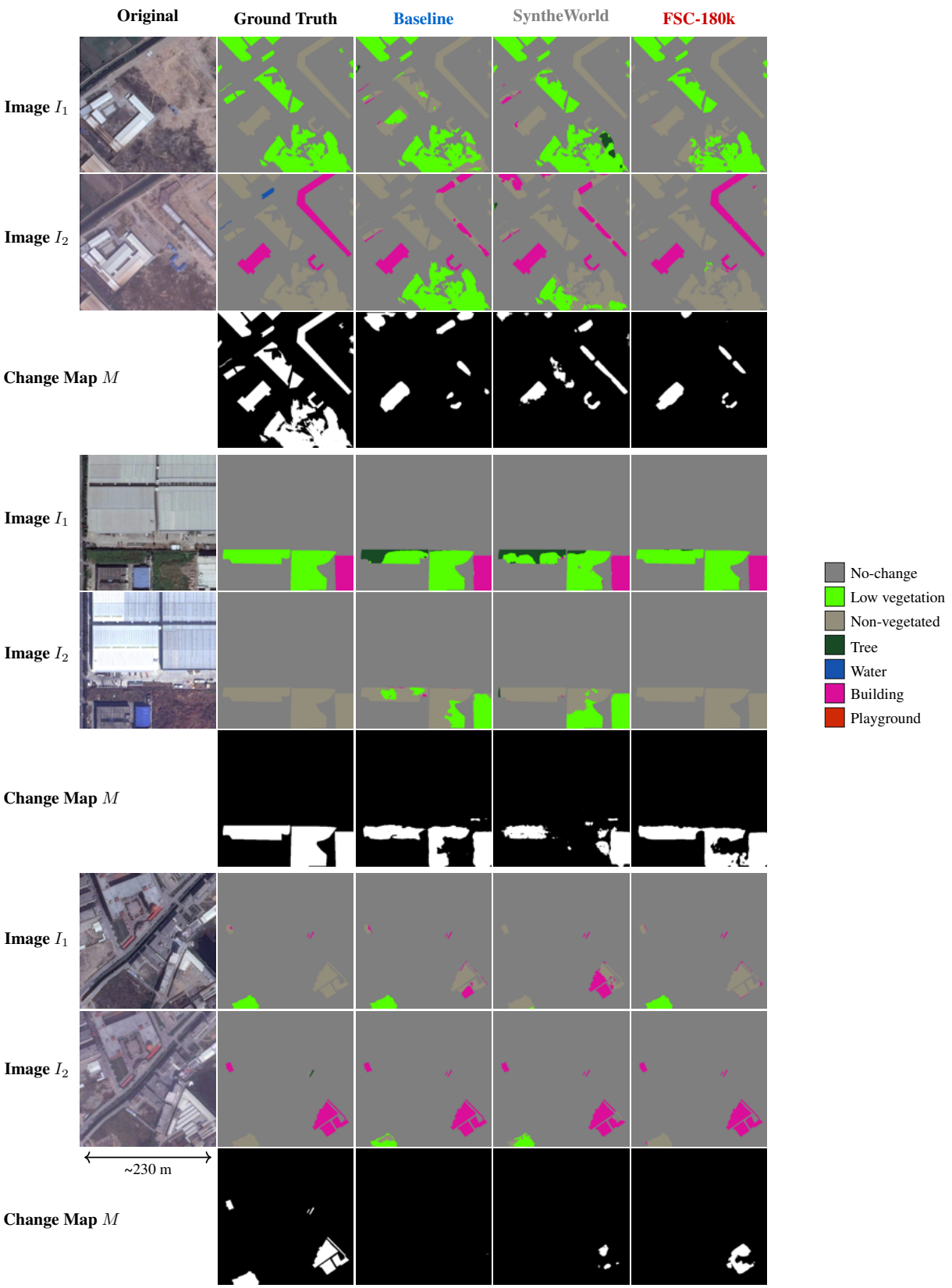

Figure E. **Qualitative change detection on the SECOND dataset.** For each pair of images, we show the land cover prediction and the binary change detection maps for models pretrained on (SyntheWorld or FSC-180k [Ours]) or not (Baseline) and then fine-tuned on the target dataset, in a sequential manner. The second column is composed of the ground truth maps. Best viewed in color.

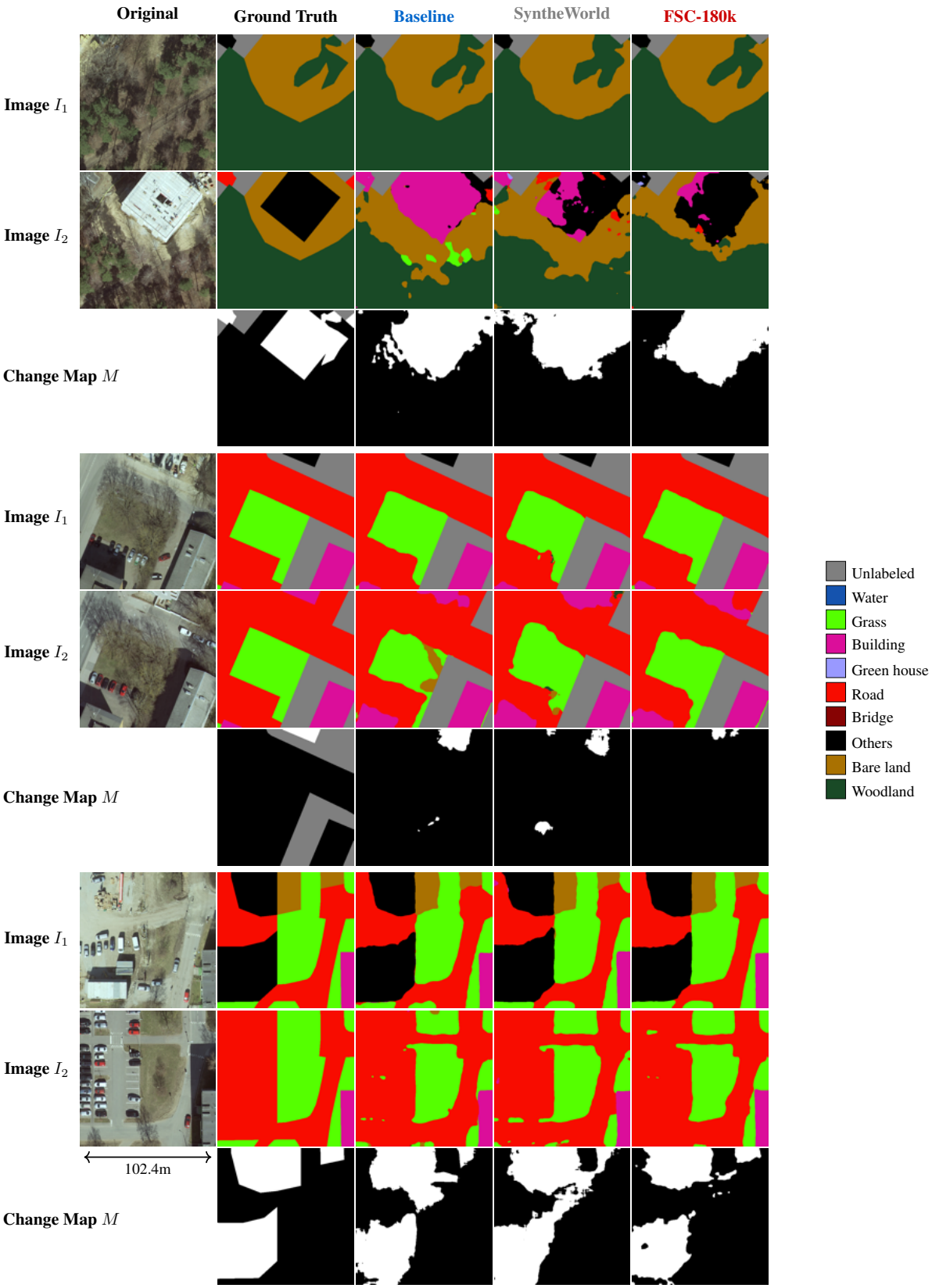

Figure F. **Qualitative change detection on the HiUCD-mini dataset.** For each pair of images, we show the land cover prediction and the binary change detection maps for models pretrained on (SyntheWorld or FSC-180k [Ours]) or not (Baseline) and then fine-tuned on the target dataset, in a sequential manner. The second column is composed of the ground truth maps. Best viewed in color.