Supplementary Material

Jian Song[1,2], Hongruixuan Chen[1,2], and Naoto Yokoya[1,2]
[1]The University of Tokyo, Japan
[2]RIKEN AIP, Japan
song@ms.k.u-tokyo.ac.jp, Qschrx@gmail.com, yokoya@k.u-tokyo.ac.jp

In this Supplementary Material, we will present additional examples from the SyntheWorld dataset, demonstrating its diversity. Following this, we will show our process of directing GPT-4 [15] to generate effective prompts for Stable Diffusion [17] and its variant models to create land cover textures, including examples of both prompts and generated textures.

Subsequently, we will provide the dataset divisions and relevant experimental settings for the cross-domain land cover mapping task, along with the quantitative and qualitative results of land cover mapping and building change detection tasks on a broader range of models.

Lastly, we will detail the specific information of all Blender [5] addons utilized in the creation of the SyntheWorld dataset.

S1. Expanded Examples from SyntheWorld

Fig. S1 presents the 2D UMAP [14] visualization of the features extracted using ResNet-50 [7] from the SyntheWorld dataset, along with the corresponding area image examples. During the creation of the SyntheWorld dataset, we intentionally simulated concentrated styles of regions in the real world, such as cities brimming with skyscrapers, suburbs dominated by low-rise apartments, grasslands, and farmland, as well as mountainous and desert areas characterized by earthen houses and bareland. These images of varied styles cluster together in feature space, effectively showcasing the diversity of the SyntheWorld dataset.

S2. Land Cover Texture Synthesis

We utilized GPT-4 [15] along with a series of Stable Diffusion models [20, 21] to generate textures. Its performance far surpasses the GAN [8] series models in various image generation tasks. The most remarkable aspect of Stable Diffusion is its text-to-image generation model, which, given precise and detailed prompts, can produce images highly consistent with the prompt descriptions. As it was trained on the LAION-5B dataset [19], a massive dataset comprising 50 billion image-text pairs, it has ample capability to function as a texture generator, a feature already used by the dream-texture [11] addon in the Blender community. However, if the prompts are not detailed and accurate enough, the images generated by Stable Diffusion can be highly unpredictable.

In the early stage of SyntheWorld creation, we attempted to generate textures using simple manually written prompts, but the textures produced often lacked diversity and quality. Therefore, inspired by a YouTube video [1], we employed GPT-4 as a prompt generator for the Stable Diffusion model.

As illustrated in Figure S2, we initially facilitated GPT-4's understanding of the Stable Diffusion operation process by leveraging guidance from the Stable Diffusion Manual and successful prompt examples sourced from Lexia [2]. Subsequently, we would provide the subject for the texture we wish to generate, and GPT-4 would return high-quality prompts to us. This process allowed for unlimited interaction with GPT-4 to correct and refine the prompts.

Specifically, we used the Stable Diffusion v2.1 [21] and DeepFloyd [20] models to generate relatively low-resolution textures, which were then upscaled to a final resolution of 2048×2048 using the Stable Diffusion x4 upscaler model [21]. Throughout the creation of SyntheWorld, we generated a total of 140,000 textures for seven types of geometries: roads, tree leaves, developed space, rangeland, agricultural land, bareland, and rooftops. All textures had a guidance scale of 7.5 and inference steps of 100, with each type of texture using at least 16 different GPT-4 guided prompts.

Fig. S3 showcases some examples of textures generated using GPT-4 guided prompts, with all negative prompts using those provided in Fig. S2. We found that the generated textures exhibited high quality in detail and rich diversity.

[1]https://www.youtube.com/watch?v=Lu2CrEpXe0M
[2]https://lexica.art/

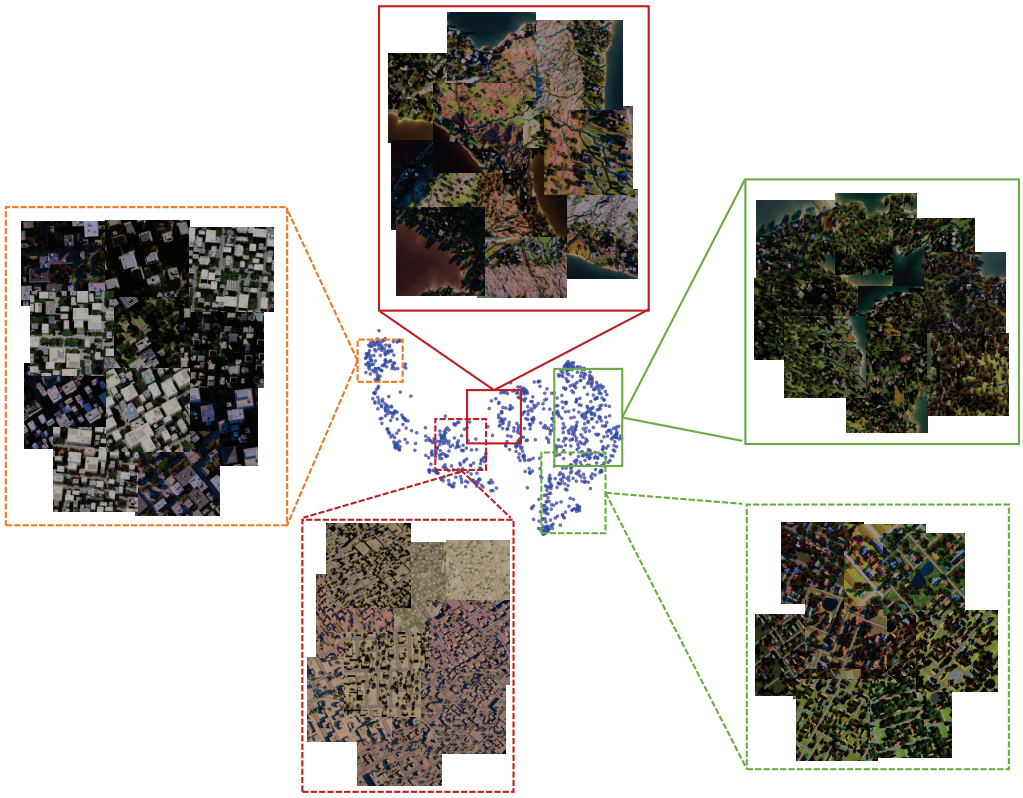


Figure S1. 2D UMAP of images from SyntheWorld encoded by ResNet-50. Solid lines represent images sourced from the terrain-based layout system with a GSD of 0.6-1m; dashed lines represent images sourced from the grid-based layout system with a GSD of 0.3-0.6m. represents mountainous or desert styles; represents represents urban styles; represents suburban or rural styles.

S3. Cross-domain Land Cover Mapping

S3.1. Dataset Division and Experimental Settings

Tab. S1 presents the division of the dataset during our cross-domain land cover mapping experiments on the OEM [26] and LoveDA [25] datasets. Our experiments tested two different semantic segmentation frameworks, U-Net [18] and DeepLabv3+ [4], along with two different CNN-based encoders, ResNet-50 [7] and EfficientNet-B4 [22], as well as a transformer-based encoder [23], MiT-B5 [27]. These models were sourced from the code repositories by Yakubovskiy [9] and Wang [24]. All experiments were carried out with a random crop of 512×512, a SGD [16] optimizer and a learning rate of 1e-3. All experiments used a batchsize of 8, in which the mixed training employed a 7:1 ratio of real to synthetic images. Each experiment was run for 100 epochs on a single Tesla A100.

S3.2. More Quantitative and Qualitative Results

Fig. S4 respectively depict the mIoU and the IoU for each category, during the continent-wise experiment with the U-Net model using the EfficientNet-B4 encoder, without employing SyntheWorld; with SyntheWorld; and the changes in mIoU and IoU for each category, respectively. We use AS to represent Asia, AF for Africa, CA for Central America, EU for Europe, NA for North America, SA for South America, and OC for Oceania.

Fig. S5 separately present the mIoU and the IoU for each category during the continent-wise experiment using the U-Net model with the EfficientNet-B4 encoder. Specifically, Fig. S5 (a) illustrates the mIoU and IoU per category when SyntheWorld is not utilized; Fig. S5 (b) shows the mIoU and IoU per category when SyntheWorld is employed; and Fig. S5 (c) shows the changes in both mIoU and IoU for each category.

Fig. S6 showcase the mIoU and IoU for each category when the U-Net model employing the MiT-B5 encoder is utilized in the continent-wise experiment. They depict the scenarios where SyntheWorld is not in use, where it is incorporated, and the corresponding changes in mIoU and IoU for each category, respectively.

We can observe that for most pairs of datasets, the SyntheWorld dataset brings substantial improvement to differ-

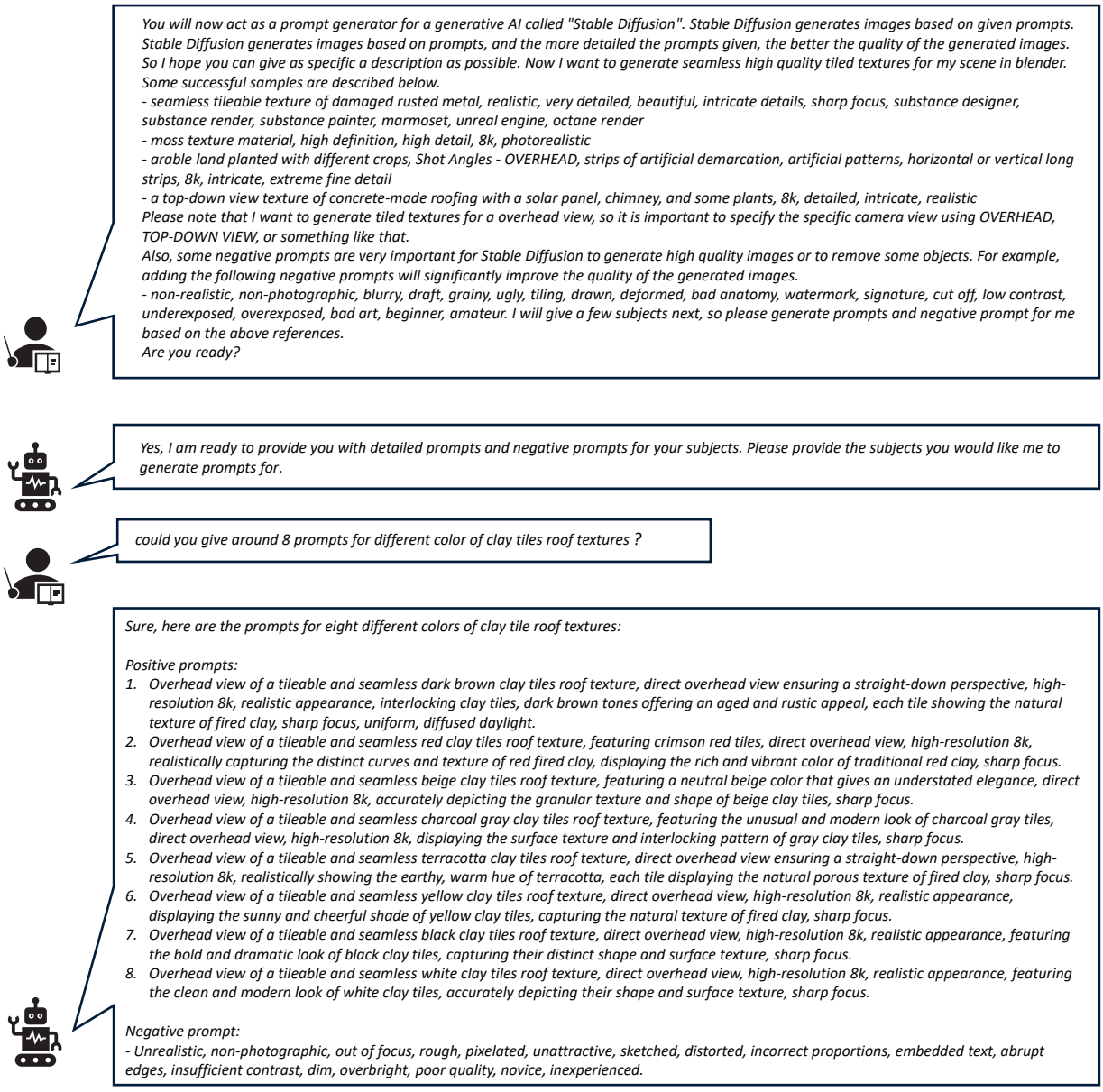
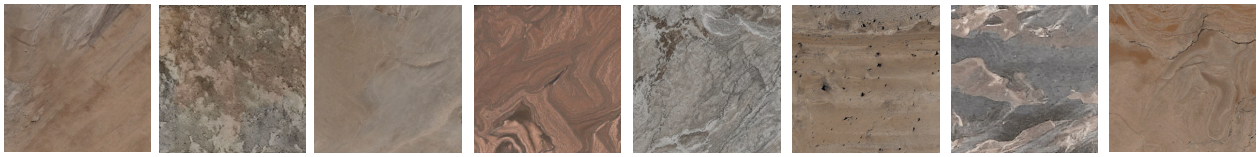


Figure S2. Guide GPT-4 to generate prompts.

Dataset division	OEM							LoveDA	
	Africa	Asia	Central America	Europe	North America	South America	Oceana	Urban	Rural
# of training images	592	568	218	902	490	523	196	1156	1366
# of testing images	259	247	94	391	210	226	84	677	992

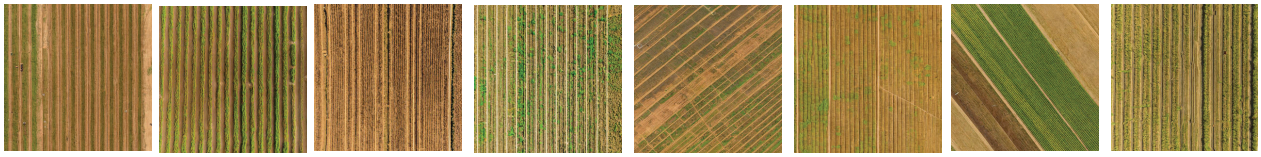
Table S1. Dataset division for cross-domain land cover mapping experiments.

Bareland



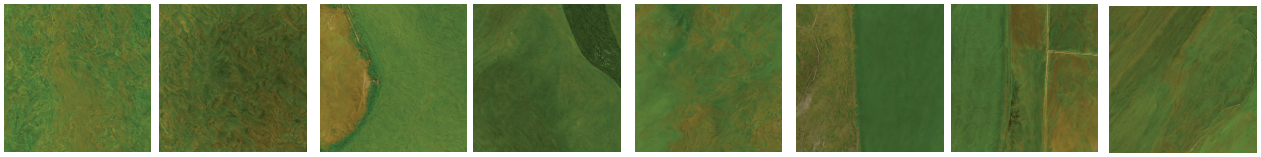
Prompt: *"Brown and tan marble textured background, shot angles - overhead, high-resolution 8k, realistic appearance, interplay of light tan and deep brown hues, organic marble patterns and veins, polished surface reflecting soft light, detailed intricacies, sharp focus, natural illumination, compatible with popular 3D engines and rendering software."*

Agricultural land



Prompt: *"Satellite view of a tileable and seamless dry cultivated land, extreme long shot, overhead view, high-resolution 8k, realistic appearance, parched and cracked soil texture, faint traces of former crop rows, sharp focus, uniform, diffused daylight."*

Rangeland



Prompt: *"satellite view meadow background of a terrain dominated by herbaceous with mud, extreme long shot, high-resolution 8k, true-to-life appearance, featuring specific species such as ryegrass, fescue, and buffalo grass, varying shades of green, natural distribution in small clusters and patches, subtle elevation changes, meticulous details, crisp focus."*

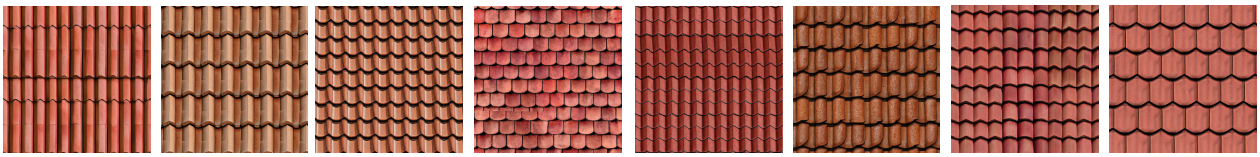
Developed space

Prompt: *"Overhead view of a parking lot, vertical image, shot angles - satellite view, high-resolution 8k, realistic appearance, monotone gray of asphalt, carefully marked parking spaces with white lines, occasional oil stains, sharp focus, natural illumination, compatible with popular 3D engines and rendering software."*

Road

Prompt: *"Satellite view of a tileable and seamless asphalt texture, featuring only one vertical road line right in the middle of the open space, extreme long shot, overhead view, high-resolution 8k, realistic appearance, dark asphalt with subtle surface imperfections, a single stark road line striking through the center, sharp focus, uniform, diffused daylight."*

Roof



Prompt: *"Satellite view of a tileable and seamless clay tiles roof texture, direct overhead view ensuring a straight-down perspective, high-resolution 8k, realistic appearance, rows of interlocking clay tiles with their distinct curved shape, showing variations of burnt orange and reddish-brown tones, each tile exhibiting the natural porous texture of fired clay, sharp focus, uniform, diffused daylight."*

Figure S3. Examples of different types of textures generated by GPT-4 guided prompts.

ent models. Meanwhile, Fig. S7 shows more visualization results of the performance improvement brought about by SyntheWorld for different pairs of datasets through different models. This further validates the robustness of the additional knowledge brought about by the SyntheWorld dataset to real datasets.

To better interpret the experimental results in Figs. S4 to S6, we further analyze the average improvements in IoU for different land cover categories and the improvement of mIoU in out-of-domain testing experiments in the continent-wise setup, which are detailed in Tab. S2.

Across different models and land cover categories, the U-Net model with the MiT-B5 backbone appears to offer the most improvement in terms of average IoU in the Bareland, Rangeland, Developed Space, Road and Tree categories. However, the U-Net model with the EfficientNet-B4 backbone has the highest IoU improvement for the 'Water' category. In terms of the category of Agriculture Land, the DeepLabv3+ with ResNet-50 backbone model shines with the highest IoU improvement. Similarly, the DeepLabv3+ with ResNet-50 model has the highest improvement in IoU in the Building category. Furthermore, considering the average improvement in mIoU, the U-Net model with MiT-B5 backbone appears to outperform the other two models.

S4. Additional Results for Building Change Detection

Tab. S3 presents the F1 scores of the FC-siam-Diff [6] model on three building change detection benchmark datasets: LEVIR-CD+ [3], SECOND [28], and WHU-CD [10]. The table compares the results of the model when run without (w / o) and with (w / o) SyntheWorld, which is not present in the main paper.

We follow a mixed training approach, maintaining a ratio of 7: 1 for real to synthetic images. We use the Adam [12] optimizer, setting the learning rate at 1e-3. Each experiment is conducted over 100 epochs and is executed on a Tesla A100 GPU.

Specifically, for the LEVIR-CD+ dataset, the model scored an F1 of 0.751 without SyntheWorld and improved to 0.766 with SyntheWorld. On the SECOND dataset, the model achieved an F1 score of 0.614 without SyntheWorld, and this score increased to 0.677 when SyntheWorld was used. Lastly, on the WHU-CD dataset, the F1 score of the FC-siam-Diff model was 0.812 without SyntheWorld and reached 0.840 with the use of SyntheWorld.

In all three cases, the use of SyntheWorld improved F1 scores, showing that it had a positive impact on the performance of the FC-siam-Diff [6] model when incorporating SyntheWorld.

We also include three tables that detail the results of the same experiment set, i.e. training data is scarce, performed using the FC-siam-Diff [6], ChangeFormer [1], and STANet [3] models.

Tab. S4 shows the performance of the FC-siam-Diff model when trained with different fractions of the real-world training set, both with and without SyntheWorld. Consistently, across all real-world datasets and at every percentage level, the incorporation of SyntheWorld boosts the model's performance.

Similar results are presented in Tab. S5 and Tab. S6, which correspond to the ChangeFormer and STANet models, respectively. Again, in each case, the addition of SyntheWorld consistently enhances the model's performance across all datasets and at each level of real-world training data usage.

These results corroborate the main finding reported in the paper, reinforcing that the SyntheWorld dataset invariably provides a significant performance boost, particularly when the amount of real training data is limited. This beneficial effect is observed not only in the DTCDSCN [13] model, but also in the FC-siam-Diff, ChangeFormer, and STANet models.

As demonstrated in Tables S7, S8, and S9, we have conducted comprehensive experiments with the DTCDSCN and FC-siam-Diff models, training them on two synthetic datasets, AICD [2] and SyntheWorld, and subsequently testing these models on three real-world datasets.

In the training phase with the synthetic datasets, we utilized the Adam optimizer and set the learning rate to 5e-4. A notable observation is that synthetic data-trained models typically achieved their best performance within fewer than 50 epochs.

In order to draw a meaningful comparison, Oracle experiments were also performed, in which the models were trained directly on real-world datasets. In these instances, we adopted a higher learning rate of 1e-3 and found that the models reached their optimal performance approximately around the 100 epoch mark.

The experiments underscored a clear trend: regardless of the real-world dataset used for testing, the models trained on the SyntheWorld dataset consistently outperformed the ones trained on the AICD dataset. Moreover, the performance of SyntheWorld-trained models, while not matching the models trained directly on the real-world datasets, came close enough to indicate a significant value of the synthetic dataset in training effective change detection models.

Further support for the effectiveness of the SyntheWorld dataset can be found in Fig. S8, which provides a visual comparison of model performance when trained with different datasets using different models.

This highlights the immense potential of the SyntheWorld dataset for building change detection. The ability of SyntheWorld to close the gap between synthetic and real-world data to an acceptable margin is an encouraging sign. It signifies that we could significantly reduce our reliance

Methods	Backbone	IoU(%)								mIoU(%)
		Bareland	Rangeland	Developed space	Road	Tree	Water	Agriculture land	Building	
U-Net	EfficientNet-B4	1.10	0.96	1.40	0.97	0.30	**5.81**	3.17	0.51	1.78
U-Net	MiT-B5	**3.46**	**1.68**	**1.68**	**5.53**	**2.87**	1.26	5.44	3.30	**3.16**
DeepLabv3+	ResNet-50	1.53	1.24	0.95	1.46	2.15	3.88	**7.32**	**3.40**	2.74

Table S2. Average improvement in IoU for different land cover categories and average mIoU improvement in out-of-domain testing experiments under the continent-wise setup for different models.

Datasets	FC-siam-Diff	
	w/o	w/
LEVIR-CD+ [3]	0.751	**0.766**
SECOND* [28]	0.614	**0.677**
WHU-CD [10]	0.812	**0.840**

Table S3. F1 score resulting from the use or non-use of SyntheWorld across three building change detection benchmark datasets, assessed with the FC-siam-Diff model. * means to use the part of building change label in SECOND.

Datasets	1%		5%		10%	
	w/o	w/	w/o	w/	w/o	w/
LEVIR-CD+ [3]	0.414	**0.558**	0.635	**0.658**	0.686	**0.759**
SECOND* [28]	0.381	**0.444**	0.545	**0.590**	0.507	**0.603**
WHU-CD [10]	0.420	**0.519**	0.680	**0.717**	0.732	**0.738**

Table S4. Comparison of F1 scores from the FC-siam-Diff model trained with and without SyntheWorld, applied on three different real-world datasets at varying ratios of real image use. * means to use the part of building change label in SECOND.

Datasets	1%		5%		10%	
	w/o	w/	w/o	w/	w/o	w/
LEVIR-CD+ [3]	0.357	**0.469**	0.520	**0.581**	0.602	**0.671**
SECOND* [28]	0.329	**0.428**	0.483	**0.521**	0.503	**0.564**
WHU-CD [10]	0.227	**0.290**	0.517	**0.644**	0.565	**0.665**

Table S5. Comparison of F1 scores from the ChangeFormer model trained with and without SyntheWorld, applied on three different real-world datasets at varying ratios of real image use. * means to use the part of building change label in SECOND.

Datasets	1%		5%		10%	
	w/o	w/	w/o	w/	w/o	w/
LEVIR-CD+ [3]	0.541	**0.600**	0.575	**0.643**	0.688	**0.741**
SECOND* [28]	0.536	**0.546**	0.554	**0.628**	0.619	**0.645**
WHU-CD [10]	0.295	**0.344**	0.524	**0.595**	0.687	**0.719**

Table S6. Comparison of F1 scores from the STANet model trained with and without SyntheWorld, applied on three different real-world datasets at varying ratios of real image use. * means to use the part of building change label in SECOND.

Training Data	Models	
	DTCDSCN	FC-siam-Diff
AICD [2]	0.160	0.133
SyntheWorld	**0.364**	**0.425**
Oracle	0.793	0.751

Table S7. Comparison of the best F1 scores achieved on the LEVIR-CD+ dataset test set by the DTCDSCN and FC-siam-Diff models. Each model was trained on each synthetic dataset and tested on the LEVIR-CD+ dataset. The Oracle results indicate the performance of each model when trained and tested on the LEVIR-CD+ dataset.

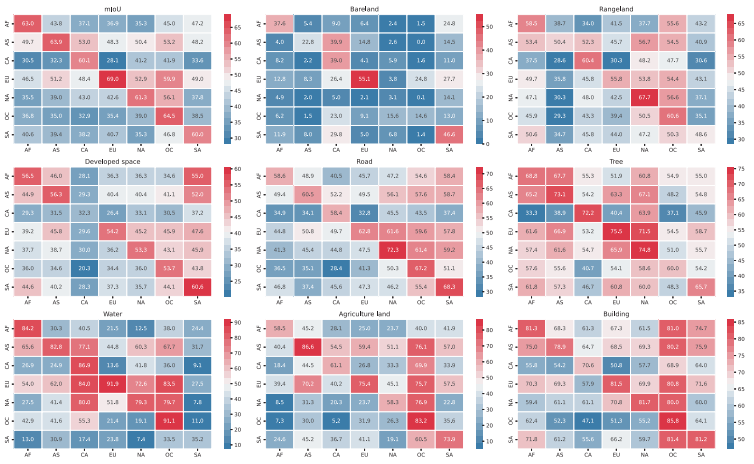
Training Data	Models	
	DTCDSCN	FC-siam-Diff
AICD [2]	0.348	0.324
SyntheWorld	**0.451**	**0.461**
Oracle	0.712	0.614

Table S8. Comparison of the best F1 scores obtained on the SECOND dataset test set by the DTCDSCN and FC-siam-Diff models. The models were independently trained on each synthetic dataset and subsequently tested on the SECOND dataset. The Oracle performance is derived from training and testing each model on the SECOND dataset.

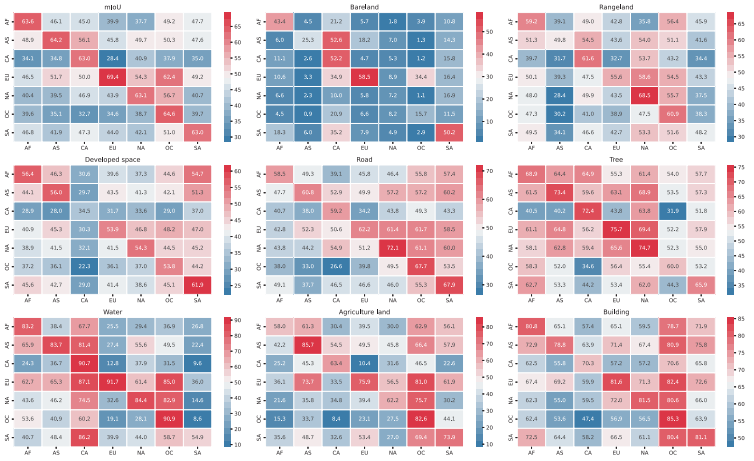
Training Data	Models	
	DTCDSCN	FC-siam-Diff
AICD [2]	0.231	0.236
SyntheWorld	**0.550**	**0.540**
Oracle	0.769	0.812

Table S9. Performance comparison of the best F1 scores on the WHU-CD test set, attained by the DTCDSCN and FC-siam-Diff models. The Oracle results represent the performance of the models when both training and testing are performed on the WHU-CD dataset.

on large volumes of real-world data for training effective change detection models.



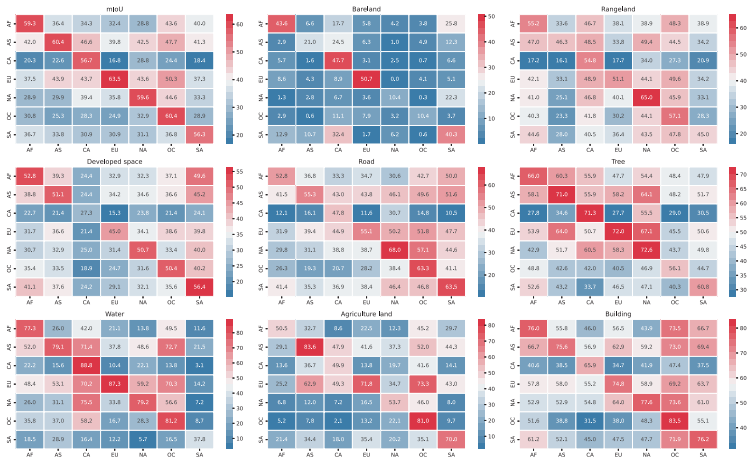
(a) Without SyntheWorld.



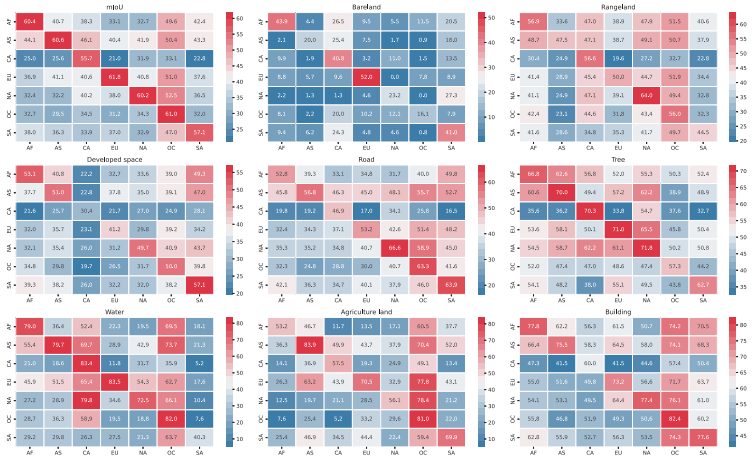
(b) With SyntheWorld.

(c) Performance changes.

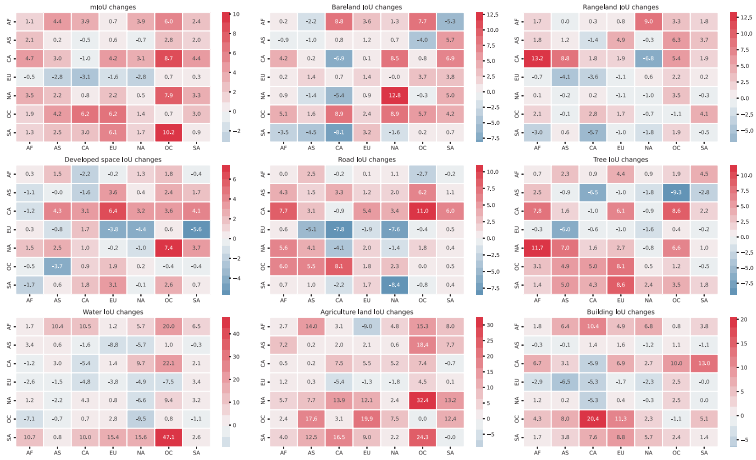
Figure S4. Results of continent-wise in-domain and out-of-domain land cover mapping experiments of OEM dataset. The x-axis represents the target domain, and the y-axis represents the source domain. U-Net with EfficientNet-B4 encoder is used for all experiments.



(a) Without SyntheWorld.

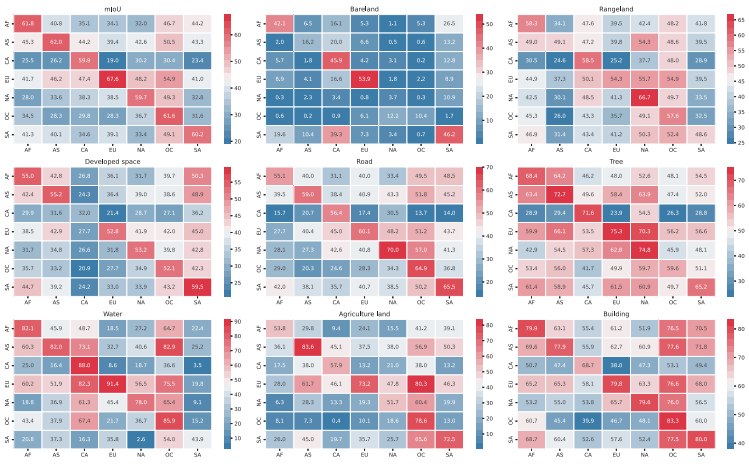


(b) With SyntheWorld.

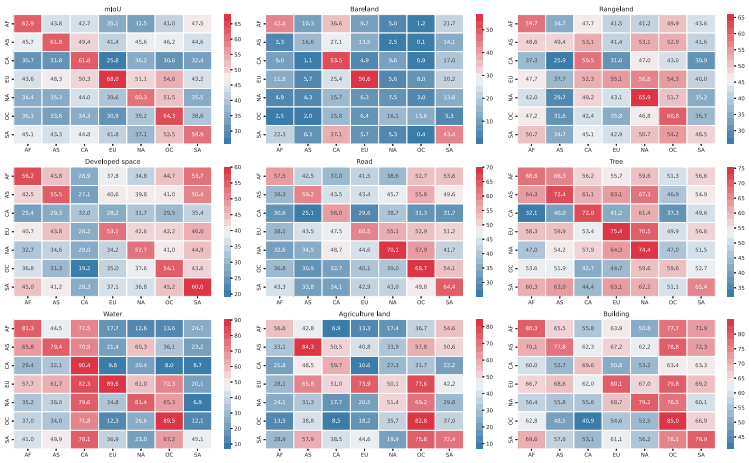


(c) Performance changes.

Figure S5. Results of continent-wise in-domain and out-of-domain land cover mapping experiments of OEM dataset. The x-axis represents the target domain, and the y-axis represents the source domain. DeepLabv3+ with ResNet-50 encoder is used for all experiments.



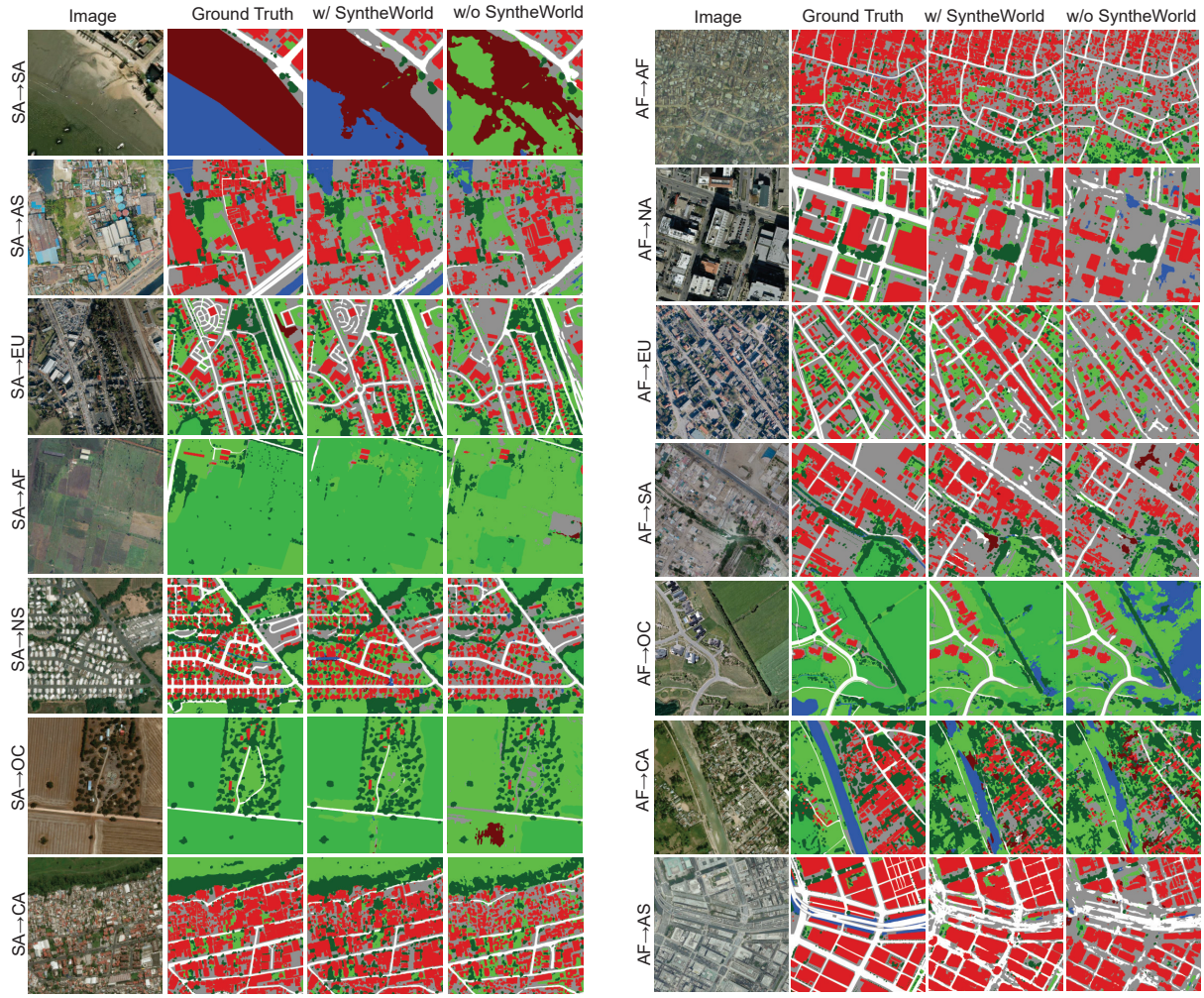
(a) Without SyntheWorld.



(b) With SyntheWorld.

(c) Performance changes.

Figure S6. Results of continent-wise in-domain and out-of-domain land cover mapping experiments of OEM dataset. The x-axis represents the target domain, and the y-axis represents the source domain. U-Net with MiT-B5 encoder is used for all experiments.



(a) UNet-EfficientNet-B4

(b) DeepLabv3+ with ResNet-50

Figure S7. Qualitative results of continent-wise in-domain and out-of-domain land cover mapping experiments using different models on the OEM dataset.

S5. Attribution of Utilized Blender Addons

The development process of SyntheWorld is based on Blender 3.4. We utilized and modified a multitude of community addons, combining them for the generation of SyntheWorld. Detailed information on all the addons used during our development process is shown in Tab. S10.

References

[1] Wele Gedara Chaminda Bandara and Vishal M Patel. A transformer-based siamese network for change detection. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 207–210. IEEE, 2022. 5

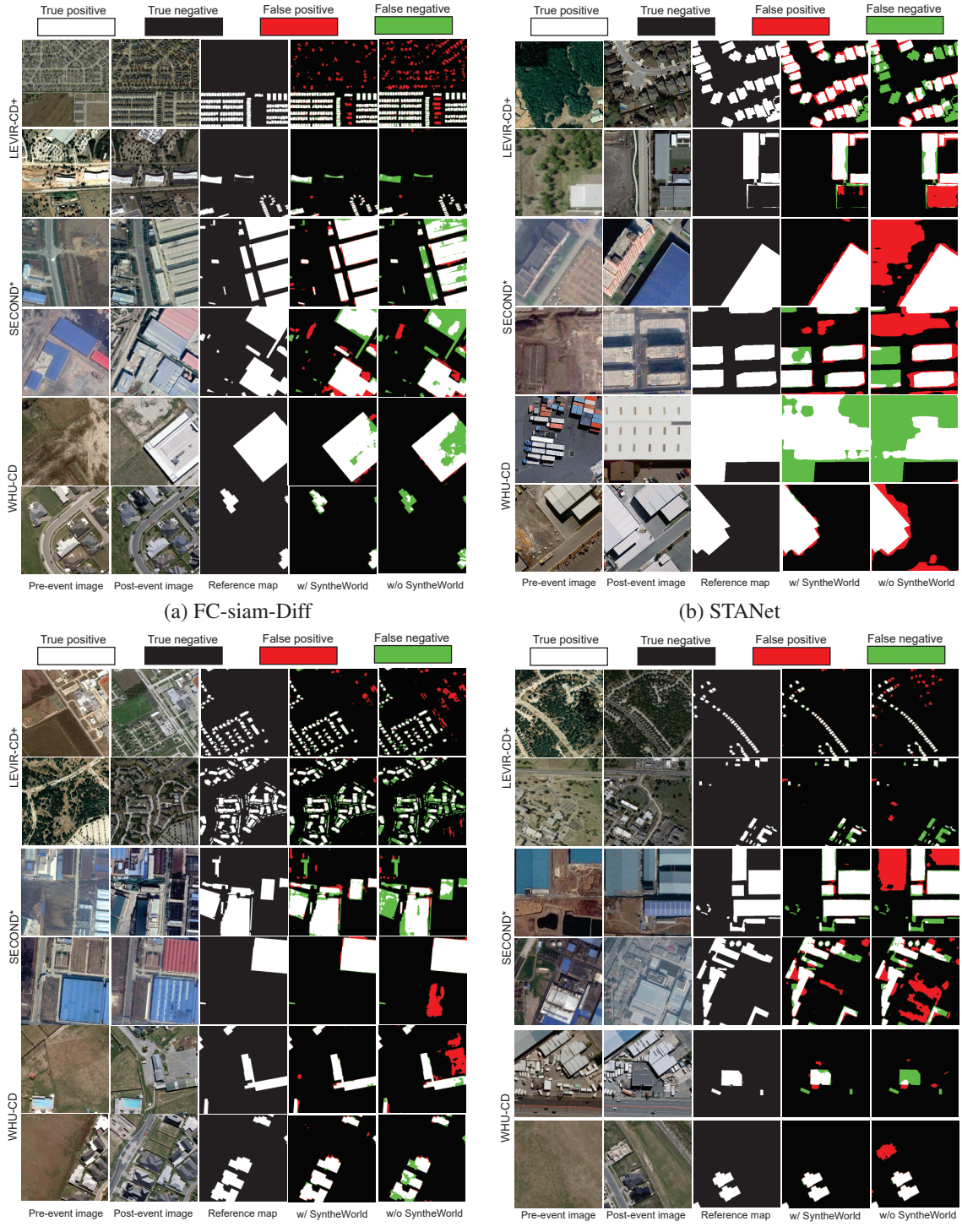
[2] Nicolas Bourdis, Denis Marraud, and Hichem Sahbi. Constrained optical flow for aerial image change detection. In *2011 IEEE international geoscience and remote sensing symposium*, pages 4176–4179. IEEE, 2011. 5, 6

[3] Hao Chen and Zhenwei Shi. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. *Remote Sensing*, 12(10):1662, 2020. 5, 6

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 2

[5] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 1

[6] Rodrigo Caye Daudt, Bertrand Le Saux, and Alexandre Boulch. Fully convolutional siamese networks for change



(a) FC-siam-Diff (b) STANet

(c) DTCDSCN (d) ChangeFormer

Figure S8. Qualitative results of the building change detection task using different models and datasets, with and without the incorporation of the SyntheWorld dataset.

Name	Author	Version	License	URL
Realtime River Generator	specoolar	1.1	RF	https://blendermarket.com/products/river-generator
Next Street	Next Realm	2.0	RF	https://blendermarket.com/products/next-street
Objects Replacer	Georeality Design	1.06	GPL	https://blendermarket.com/products/objects-replacer/docs
Albero	Greenbaburu	0.3	RF	https://blendermarket.com/products/albero---geometry-nodes-powered-tree-generator
Hira Building Generator	HiranojiStore	0.9	RF	https://blendermarket.com/products/hira-building-generator
Procedural Building Generator	Isak Waltin	1.2.1	CC-BY 4.0	https://blendermarket.com/products/building-gen
Pro Atmo	Contrastrender	1.0	GPL	https://blendermarket.com/products/pro-atmo
Modular Buildings Creator	PH Felix	1.0	RF	https://blendermarket.com/products/modular-buildings-creator
Next Trees	Next Realm	2.0	RF	https://blendermarket.com/products/next-trees
SceneCity	Arnaud	1.9.3	RF	http://www.cgchan.com/store/scenecity
Flex Road Generator	EasyNodes	1.1.0	RF	https://www.cgtrader.com/3d-models/scripts-plugins/modelling/blender-mesh-curve-to-road
Buildify	Pavel Oliva	1.0	RF	https://paveloliva.gumroad.com/l/buildify

Table S10. List of Blender Addons used in the study.

detection. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 4063–4067. IEEE, 2018. 5

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 2

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 1

[9] Pavel Iakubovskii. Segmentation models pytorch. https://github.com/qubvel/segmentation_models.pytorch, 2020. 2

[10] Shunping Ji, Shiqing Wei, and Meng Lu. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1):574–586, 2018. 5, 6

[11] Carson Katri. dream-textures. https://github.com/carson-katri/dream-textures, 2022. 1

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[13] Yi Liu, Chao Pang, Zongqian Zhan, Xiaomeng Zhang, and Xue Yang. Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model. *IEEE Geoscience and Remote Sensing Letters*, 18(5):811–815, 2020. 5

[14] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 1

[15] OpenAI. Gpt-4 technical report, 2023. 1

[16] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951. 2

[17] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2

[19] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 1

[20] StabilityAI. deep-floyd. https://github.com/deep-floyd, 2023. 1

[21] StabilityAI. Stable diffusion version 2. https://github.com/Stability-AI/stablediffusion, 2023. 1

[22] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 2

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[24] Junjue Wang. Loveda repository. https://github.com/Junjue-Wang/LoveDA, 2021. 2

[25] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021. 2

[26] Junshi Xia, Naoto Yokoya, Bruno Adriano, and Clifford Broni-Bediako. Openearthmap: A benchmark dataset for global high-resolution land cover mapping. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6254–6264, 2023. 2

[27] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 2

[28] Kunping Yang, Gui-Song Xia, Zicheng Liu, Bo Du, Wen Yang, Marcello Pelillo, and Liangpei Zhang. Semantic change detection with asymmetric siamese networks. *arXiv preprint arXiv:2010.05687*, 2020. 5, 6