MapFormer: Boosting Change Detection by Using Pre-change Information Supplementary Material

Maximilian Bernhard, Niklas Strauß, Matthias Schubert
LMU Munich, MCML
{bernhard,strauss,schubert}@dbs.ifi.lmu.de

A. Hyperparameters and Training Details

We used the same training parameters as [5], except for the number of training steps which we had to increase for the used datasets. All models (except SETR) were trained for 32k iterations with a batch size of 16 (i.e. 32 images per batch), a learning rate of 6×10^{-5}, and the AdamW optimizer [4]. For SETR, we halved the batch size and the learning rate to account for its high demand of GPU memory. DynamicEarthNet and HRSCD images were split into tiles of size 512 and 500, respectively. We use a default value of $K = 10$ for most experiments and provide results for $K = 5$ and $K = 15$ to examine parameter sensitivity to K. In the DynamicMLP mechanism, we reduced the number of channels by a factor of three to fit the model into our GPU memory. An extensive listing of the used hyperparameters can be found in the configuration YAML files in the code[1]. Overall, the training schedules and data loading pipelines were chosen identically for all methods in order to ensure maximum comparability.

For MapFormer, FHD [5], ChangeFormer [1], and the baselines, we use MiT-b2 [8] pre-trained on ImageNet as backbone, whereas a ResNeSt50 [9] backbone was employed for ChangerEx [3]. UPerNet and SETR were equipped with a Swin-S and ViT-B backbone, respectively. We compare model sizes in Table 1. To counter the effect of the extreme class imbalance when training FHD, ChangeFormer, and ChangerEx on HRSCD, we had to introduce a class weight of 4 on the change class for binary change detection, which we chose empirically based on the best performance from the range of 1, 2, 4, and 8. Note that this was not necessary for MapFormer. All experiments were run on a single 40 GB Nvidia A100 or a comparable device.

Model	Params
SegFormer (MiT-b2)	24.7m
UPerNet (Swin-S)	81.2m
SETR-PUP (ViT-B)	97.1m
FHD	29.1m
ChangerEx	26.8m
ChangeFormer	25.0m
MapFormer$_{K=5}$	30.2m
MapFormer$_{K=10}$	34.5m
MapFormer$_{K=15}$	38.8m

Table 1: Number of parameters for different models considered in our experiments.

B. Additional Experiments

In Table 2, we present further experiments to investigate certain design choices in our architecture. First, we examine a modified version of our fusion module where we have shared attention weights for groups of features (8 attention groups). That is, we reduce the shape of the attention weights to $K \times (D_f/8)$ for every spatial location and reuse each of the attention weights for 8 of the D_f channels. This is inspired by multi-head attention [7] with 8 heads and reduces the computational complexity when generating the attention weights. However, we find that this version is inferior to our proposed method (-3.5% IoU).

Next, we assess MapFormer's performance if we omit $g^{(1)}$ for $(h_k)_{k=1\dots K}$ (no $g^{(1)}$ for $(h_k)_{k=1\dots K}$). That is, we use $g^{(1)}$ only for the attention weights a. Apparently, this limits the representation capability of our model, leading to an IoU drop of 2.5%.

Furthermore, we experimented with additional components for our contrastive loss module. The first extension is an additional contrastive loss applied directly on the image features (contrastive loss between $f^{(1)}$ and $f^{(2)}$), i.e., we do not only align/contrast $f^{(1)}$ and $f^{(2)}$ with $g^{(1)}$, but also $f^{(1)}$ with $f^{(2)}$ directly. However, we were not able to improve the performance with this approach (-4.1% IoU). Another version of the contrastive loss included the use of additional negative samples to counter the imbalance of changed and unchanged areas (contrastive loss with add. neg. samples). More precisely, we balanced the number of positive and

[1] e.g., https://github.com/mxbh/mapformer/blob/master/configs/conditional_bcd/dynamicearthnet/mapformer.yaml

Method	F1	IoU
MapFormer$_{K=10}$ (full)	**38.0**	**23.5**
8 attention groups	33.3	20.0
no $g^{(1)}$ for $(h_k)_{k=1...K}$	34.7	21.0
contrastive loss between $f^{(1)}$ and $f^{(2)}$	32.5	19.4
contrastive loss with add. neg. samples	37.0	22.7

Table 2: Additional experiments for MapFormer for Conditional BCD on DynamicEarthNet.

negative pixel pairs in the contrastive loss by randomly sampling pixels with a different semantic class from other spatial locations. This resulted in a strong performance with an IoU of 22.7%, but did not suffice to outperform our proposed method.

C. Dataset Statistics and Details

Statistics We provide general statistics for the datasets used in this paper in Table 3. While both DynamicEarthNet [6] and HRSCD [2] consider land use/land cover classes, the resolution of the images differs (3m vs. 0.5m). Also, HRSCD surpasses DynamicEarthNet in the number of annotated tiles (5,280 vs. 232,800). However, the percentage of change pixels in the image pairs is much higher for DynamicEarthNet (4.5% vs. 0.8%), i.e., the class imbalance between "change" and "no-change" is more severe for HRSCD. A major difference between the two datasets is that DynamicEarthNet provides manual annotations [6], whereas the annotations in HRSCD come from the Urban Atlas project of the European Environment Agency (EEA)[2]. Furthermore, DynamicEarthNet images contain four bands (RGB + near-infrared). However, we only use the RGB information to be able to use pre-trained image encoder backbones and for better comparability with other methods. When splitting DynamicEarthNet into training, validation, and test set, we made sure that each semantic class is represented in each split.

Creating Pairs on DynamicEarthNet Since DynamicEarthNet consists of multi-temporal time series (24 months), there are two natural options to generate bi-temporal image pairs. The first is to only use images from consecutive months, i.e., $[(t_1, t_2), (t_2, t_3), ..., (t_{23}, t_{24})]$. The second option is to consider all possible pairs, i.e., $[(t_1, t_2), (t_1, t_3), ..., (t_1, t_{24}), (t_2, t_3), (t_2, t_4)...]$. The second option leads to significantly more image pairs ($24 \cdot 23 = 552$ vs. 23), but introduces high correlations between the samples. Consequently, we use only consecutive pairs for validation and testing (as suggested in [6]), but train with all possible pairs of the time series to augment the data.

Differences between DynamicEarthNet and HRSCD From our experiments, we gained the high-level insight that semantic segmentation is more closely connected to change detection on DynamicEarthNet than it is on HRSCD. For example, the bi-temporal SegFormer baseline is competitive among the bi-temporal SOTA methods on DynamicEarthNet, while it is clearly outperformed on HRSCD (see Table 2 in the paper). Also, replacing $m^{(1)}$ with a semantic segmentation prediction $\hat{m}^{(1)}$ yields strong results with MapFormer, whereas the performance on HRSCD cannot be improved compared to other bi-temporal methods (see Table 2 in the paper). We argue that the reason for this behavior is twofold: first, the binary change ground-truth on DynamicEarthNet is generated directly from the semantic segmentation ground truths. In contrast, HRSCD has additional binary change labels that do not perfectly agree with the semantic segmentation labels. Thus, a perfect semantic segmentation model would achieve a perfect binary change performance on DynamicEarthNet, but not on HRSCD. The second reason is that the annotations on DynamicEarthNet were manually created on the images themselves, which is not the case for HRSCD. Therefore, the DynamicEarthNet annotations can be considered to be of higher quality (see E Further Qualitative Results).

The overall performance gap between the datasets can be explained by the lower resolution of DynamicEarthNet (3m vs. 0.5m GSD) to a certain extent. The lower resolution makes it more difficult to classify surface patterns in DynamicEarthNet images, and also results in many small detail changes that are particularly hard to detect. All in all, it is much more difficult to detect changes with the naked eye in DynamicEarthNet than in HRSCD (see E Further Qualitative Results).

D. Evaluation Protocol for Conditional and Cross-modal Semantic Change Detection

In Conditional and Cross-modal Semantic Change Detection, the are two ways to generate semantic segmentation predictions for $I^{(2)}$. The first is to directly use the semantic segmentation head's output $\hat{m}^{(2)}$ (after argmax along the class dimension). The second option is to additionally utilize $m^{(1)}$ and $\hat{b}$, and compute the prediction via

$$\tilde{m}^{(2)} = (1 - \hat{b}) \cdot m^{(1)} + \hat{b} \cdot \hat{m}^{(2)}.$$

That is, $\tilde{m}^{(2)}$ simply repeats the pre-change map $m^{(1)}$ for regions with $\hat{b} = 0$ and only uses $\hat{m}^{(2)}$ otherwise. However, in our experiments, we consider $\hat{m}^{(2)}$ instead of $\tilde{m}^{(2)}$ as the post-change semantic segmentation prediction for the following reason: The evaluation metric SCS (see Equation (1) in the paper, proposed in [6]) measures the model's performance for binary change detection and semantic segmentation separately via BC and SC, respectively. Finally, the two terms are combined via the arithmetic mean to obtain SCS. If we use $\tilde{m}^{(2)}$ for evaluation, the binary change prediction

[2] https://land.copernicus.eu/local/urban-atlas

	DynamicEarthNet	HRSCD
Years	2018/2019	2005/2006/2012
Spatial Distribution	global	France
Source	PlanetFusion	BD ORTHO
GSD (m)	3	0.5
#Classes	7 (6 used)	5
#Images (overall)	54,750	582
Orig image size (px)	1024 x 1024	10k x 10k
Tile size (px)	512	500
#Annotated tiles (used)	5,280	232,800
#Training pairs	38,640	76,400
Changed pixels	4.5%	0.8%

Table 3: Dataset statistics for DynamicEarthNet and HRSCD.

affects the semantic segmentation, i.e., BC and SC are no longer separate. For instance, every false negative pixel in $\hat{b}$ automatically leads to a misclassified pixel in $\tilde{m}^{(2)}$ since the old, outdated semantic class is simply repeated for this pixel. Thus, to ensure an independent assessment of BC and SC, we directly use $\hat{m}^{(2)}$ as semantic segmentation prediction in SCD tasks. In practice, however, one may consider $\tilde{m}^{(2)}$ more suitable as it does not alter the given semantic map $m^{(1)}$ in supposedly unchanged areas, which may be desirable.

E. Further Qualitative Results

We provide additional qualitative results for HRSCD and DynamicEarthNet in Figures 1 and 2, respectively. In the examples from HRSCD (Figure 1), we can see that MapFormer is generally able to identify the changed areas for this dataset. However, we observe that the model struggles to correctly detect the boundaries of the changed segments. Looking at the input images, the semantic maps $m^{(1)}$ and $m^{(2)}$, and the binary change ground truth b, we find that the boundaries of the ground-truth annotations are also rather imprecise (especially in the last example). Thus, we conclude that the quality of the annotations is the main limiting factor for MapFormer's performance here. This is consistent with the observation that downsampling the semantic input $m^{(1)}$ by a factor of 32 only decreased the BC IoU performance by a relative of 9% (vs. 23.9% on DynamicEarthnet, see Table 2 in the paper).

In the DynamicEarthNet examples (Figure 2), we observe that detecting change is much more challenging on this dataset than for HRSCD. Nevertheless, the overall look of our predictions is relatively accurate. For the last example, we have chosen a sample where MapFormer fails. Here, our model is not able to detect that large parts of the vegetation in the pre-change image $I^{(1)}$ become soil in the post-change image $I^{(2)}$. Such changes are barely noticeable from the visual input, leading to the overall relatively low binary change IoU scores of all models (our best model achieved 23.5%).

References

[1] Wele Gedara Chaminda Bandara and Vishal M Patel. A transformer-based siamese network for change detection. *arXiv preprint arXiv:2201.01293*, 2022. 1

[2] Rodrigo Caye Daudt, Bertrand Le Saux, Alexandre Boulch, and Yann Gousseau. Multitask learning for large-scale semantic change detection. *Computer Vision and Image Understanding*, 187:102783, 2019. 2

[3] Sheng Fang, Kaiyu Li, and Zhe Li. Changer: Feature interaction is what you need for change detection. *arXiv preprint arXiv:2209.08290*, 2022. 1

[4] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 1

[5] Gensheng Pei and Lulu Zhang. Feature hierarchical differentiation for remote sensing image change detection. *IEEE Geoscience and Remote Sensing Letters*, 2022. 1

[6] Aysim Toker, Lukas Kondmann, Mark Weber, Marvin Eisenberger, Andrés Camero, Jingliang Hu, Ariadna Pregel Hoderlein, Çağlar Şenaras, Timothy Davis, Daniel Cremers, et al. Dynamicearthnet: Daily multi-spectral satellite dataset for semantic change segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21158–21167, 2022. 2

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[8] 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. 1

[9] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. In *Proceed-*

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2736–2746, 2022. 1

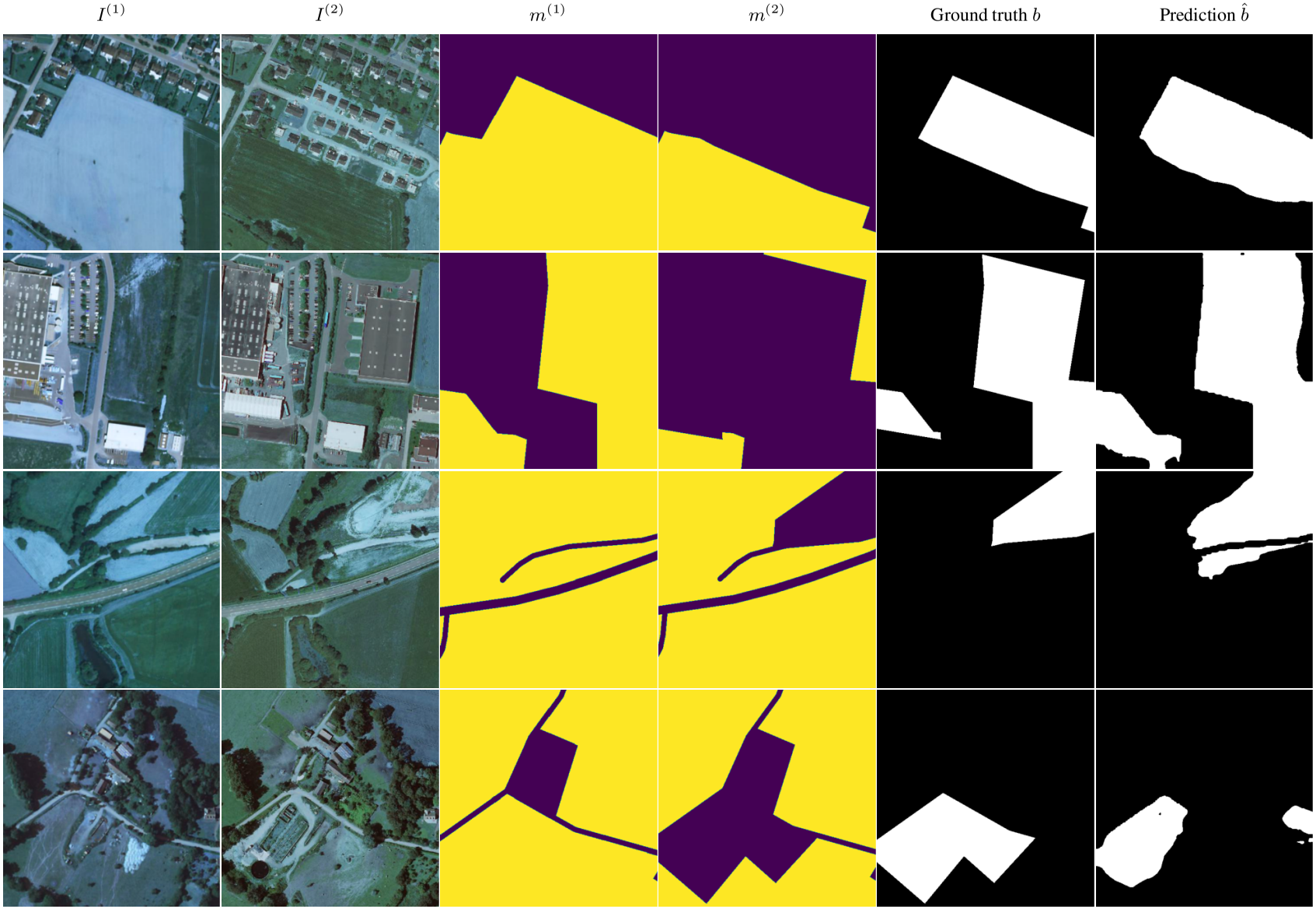

Figure 1: Additional qualitative results on HRSCD for Conditional Binary Change Detection with MapFormer.

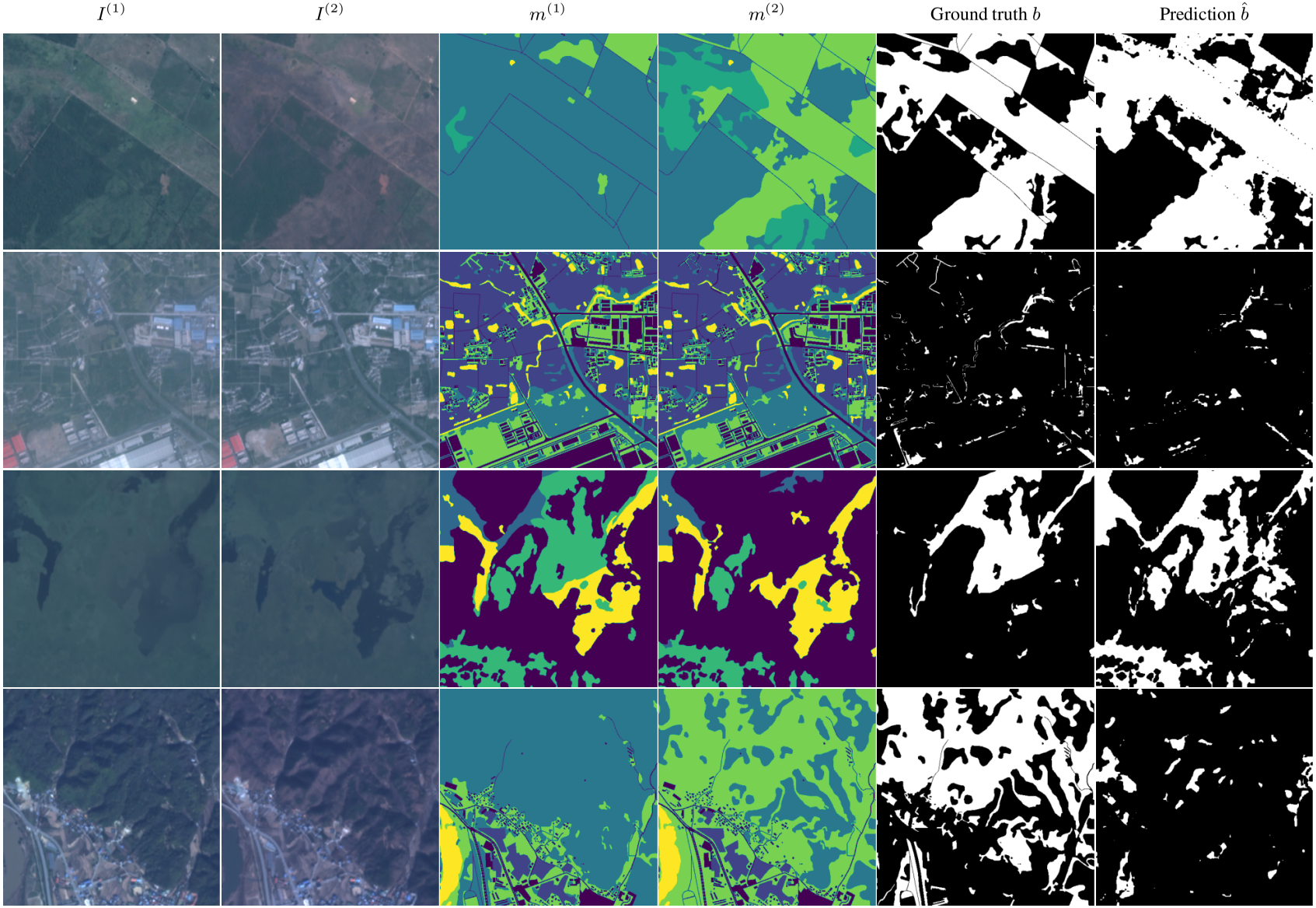

Figure 2: Additional qualitative results on DynamicEarthNet for Conditional Binary Change Detection with MapFormer.