TOWARDS EXPLAINABLE LAND COVER MAPPING: A COUNTERFACTUAL-BASED STRATEGY

Cassio F. Dantas[†*], **Diego Marcos**[‡*], **Dino Ienco**[†*]
[†]UMR TETIS, INRAE, [‡]Inria, [*]University of Montpellier, France
{cassio.fraga-dantas,dino.ienco}@inrae.fr, diego.marcos@inria.fr

ABSTRACT

Counterfactual explanations are an emerging tool to enhance interpretability of deep learning models. Given a sample, these methods seek to find and display to the user similar samples across the decision boundary. In this paper, we propose a generative adversarial counterfactual approach for satellite image time series in a multi-class setting for the land cover classification task. One of the distinctive features of the proposed approach is the lack of prior assumption on the targeted class for a given counterfactual explanation. This inherent flexibility allows for the discovery of interesting information on the relationship between land cover classes. The other feature consists of encouraging the counterfactual to differ from the original sample only in a small and compact temporal segment. These time-contiguous perturbations allow for a much sparser and, thus, interpretable solution. Furthermore, plausibility/realism of the generated counterfactual explanations is enforced via the proposed adversarial learning strategy.

1 INTRODUCTION

Deep learning techniques have gained widespread popularity in the remote sensing field due to impressive results on a variety of tasks such as image super-resolution, image restoration, biophysical variables estimation and land cover classification from satellite image time series (SITS) data (Yuan et al., 2020). Of particular importance, this last task provides useful knowledge to support many downstream geospatial analyses (Inglada et al., 2017). Despite the high performances achieved by recent deep learning frameworks on this task, they remain black-box models with limited understanding on their internal behavior. Due to this limitation, there is a growing need for improving the interpretability of deep learning models in remote sensing with the objective to raise up their acceptability and usefulness, as their decision-making processes are often not transparent (Adadi & Berrada, 2018; Guidotti et al., 2019; Arrieta et al., 2020). Counterfactual explanation methods have recently received increasing attention as a means to provide some level of interpretability (Wachter et al., 2017; Verma et al., 2020; Guidotti, 2022) to these black-box models. Counterfactual explanations aim to describe the behaviour of a model by providing minimal changes to the input data that would result in realistic samples that result in the model predicting a different class.

For these perturbations to be more easily interpretable it is desirable that they are sparse and that they can be identified with some semantic element of the input data. In the case of time series, this would require to perturb a short and contiguous section of the timeline (Delaney et al., 2021). Most papers on counterfactual explanations focus on image data, while much fewer concentrate on time series (Delaney et al., 2021; Li et al., 2022; Ates et al., 2021; Guidotti et al., 2020; Lang et al., 2022; Van Looveren et al., 2021; Filali Boubrahimi & Hamdi, 2022). To the best of our knowledge, this is the first paper focusing more specifically on counterfactuals for remote sensing time series data. The proposed approach generates counterfactual explanations that are plausible (i.e. belong as much as possible to the data distribution) and close to the original data (modifying only a limited and contiguous set of time entries by a small amount). Finally, it is not necessary to pre-determine a target class for the generated counterfactual.

Paper outline In Section 2 we describe the considered study case with the associated remote sensing data. After detailing the proposed method in Section 3, we present the experimental results in Section 4. Concluding remarks and future works are outlined in Section 5.

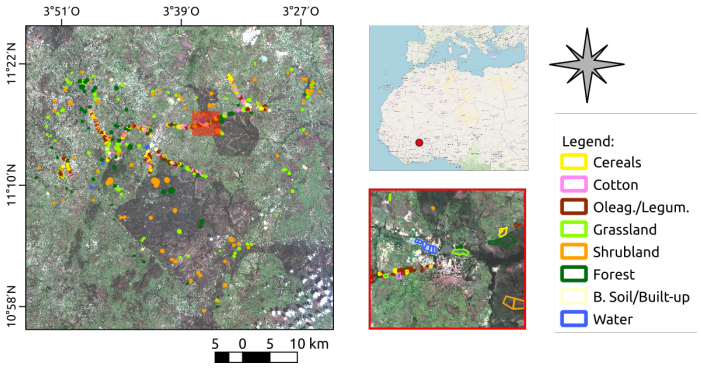

Figure 1: Study site location and corresponding ground truth.

Class	Label	Pixels
1	*Cereals*	9 731
2	*Cotton*	6 971
3	*Oleaginous*	7 950
4	*Grassland*	12 998
5	*Shrubland*	22 546
6	*Forest*	17 435
7	*Bare soil*	1 125
8	*Water*	1 205
Total		79 961

Table 1: Ground Truth statistics.

2 Study Area

The study site covers an area around the town of *Koumbia*, in the Province of Tuy, *Hauts-Bassins* region, in the south-west of Burkina Faso. This area has a surface of about 2338 km^2, and is situated in the sub-humid sudanian zone. The surface is covered mainly by natural savannah (herbaceous and shrubby) and forests, interleaved with a large portion of land (around 35%) used for rainfed agricultural production (mostly smallholder farming). The main crops are cereals (maize, sorghum and millet) and cotton, followed by oleaginous and leguminous crops. Several temporary watercourses constitute the hydrographic network around the city of Koumbia. Figure 1 presents the study site with the reference data (ground truth) superposed on a Sentinel-2 image.

The satellite data consists of a time series of Sentinel-2 images spanning the year 2020 from January to December (Jolivot et al., 2021). All images were provided by the THEIA Pole platform[1] at level-2A, which consist of atmospherically corrected surface reflectances (cf. MAJA processing chain (Hagolle et al., 2015)) and relative cloud/shadow masks. A standard pre-processing was performed over each band to replace cloudy pixel values as detected by the available cloud masks based on the method proposed in (Inglada et al., 2016). Finally, from the spectral raw bands at 10-m of spatial resolution the NDVI (Normalized Differential Vegetation Index) was derived.

The GT (ground truth) data for the study site is a collection of (i) digitized plots from a GPS field mission performed in October 2020 and mostly covering classes within cropland and (ii) additional reference plots on non-crop classes obtained by photo-interpretation by an expert. Finally, the polygons have been rasterized at the S2 spatial resolution (10-m), resulting in 79961 labeled pixels. The statistics related to the GT are reported in Table 1.

3 Proposed Method

Architecture overview The proposed GAN (generative adversarial network)-inspired architecture is shown in Fig. 2. A counterfactual x_{CF} is obtained for each input sample x by adding a perturbation δ:

$$x_{\mathrm{CF}} = x + \delta \tag{1}$$

The perturbation δ is generated by a *Noiser* module learned with the goal to swap the prediction of the *Classifier*. A *Discriminator* module is leveraged to ensure the generation of realistic counterfactual examples.

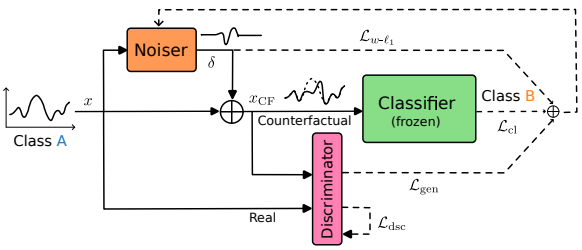

Figure 2: Schematic view of the proposed approach.

[1] http://theia.cnes.fr

Implementation and training Regarding the different components of the proposed architecture, we get inspiration from state-of-the-art satellite image time series land cover mapping literature. For the *Classifier* network we leverage the Temporal Convolutional Neural Network (TempCNN) model (Pelletier et al., 2019). This architecture has an encoder based on several one-dimensional convolutional layers to explicitly cope with the temporal dimension of the time series data followed by two fully connected layers and a final output layer to provide the multi-class decision. The same architecture is used for the *Discriminator* module, replacing the output layer with a single neuron with sigmoid activation. The *Noiser* module is implemented as a multi-layer perceptron network with two hidden layers (each with 128 neurons) and an output layer with the same dimensionality of the time series data. For each of the hidden layers, batch normalization, tangent activation function and a drop-out regularization are employed in this order while for the output layer only the tangent activation function is used. (to restrict the output domain between ± 1, like the NDVI index). The *Classifier* model is pre-trained on the training set and, then, frozen during the adversarial learning stage —devoted to learn the *Noiser* and *Discriminator* modules (see eq. (4)).

The *Noiser* module is updated w.r.t. a composite loss made of three parts detailed further below.

$$\mathcal{L}_{\text{noiser}} = \mathcal{L}_{\text{cl}} + \lambda_{\text{gen}}\mathcal{L}_{\text{gen}} + \lambda_{w\text{-}\ell_1}\mathcal{L}_{w\text{-}\ell_1} \tag{2}$$

Class-swapping loss To generate counterfactuals that effectively change the predicted class we use the following loss which enforces the reduction of the classifier's softmax output $p(y^{(i)})$ for the original label $y^{(i)}$, eventually leading to a change on the predicted class:

$$\mathcal{L}_{\text{cl}} = -\frac{1}{n}\sum_{i=1}^{n}\log(1 - p(y^{(i)})) \tag{3}$$

Note that, conversely to standard literature (Filali Boubrahimi & Hamdi, 2022; Lang et al., 2022) in which a target class for the counterfactual example is chosen a priori, here we purposely do not enforce the prediction of a predefined target class. Instead, we let the *Noiser* free to generate a perturbation δ that will change the classifier output to any other class different from y_i.

GAN-based regularization for plausibility Counterfactual plausibility is enforced via a GAN-inspired architecture, where the *Discriminator* is trained to identify unrealistic counterfactuals while, simultaneously, the *Noiser* module acts as a generator with the goal to fool the discriminator in a two player game. The following non-saturating GAN losses are used for the adversarial training:

$$\mathcal{L}_{\text{dsc}} = -\frac{1}{n}\sum_{i=1}^{n}\left[\log D(x^{(i)}) + \log\left(1 - D(x_{\text{CF}}^{(i)})\right)\right], \quad \mathcal{L}_{\text{gen}} = -\frac{1}{n}\sum_{i=1}^{n}\log\left(D(x_{\text{CF}}^{(i)})\right) \tag{4}$$

where $D(x^{(i)})$ denotes the discriminator's output for a real input $x^{(i)}$ (with expected output 1) and $D(x_{\text{CF}}^{(i)})$ its output for a fake input $x_{\text{CF}}^{(i)}$ (with expected output 0). The loss $\mathcal{L}_{\text{dsc}}$ is minimized by the *Discriminator* while $\mathcal{L}_{\text{gen}}$ is minimized the generator (i.e., the *Noiser*).

Unimodal regularization for time-contiguity To generate perturbations concentrated around a contiguous time frame we employ a weighted L1-norm penalization, with weights growing quadratically around a central time $\tilde{t}^{(i)}$ chosen independently for each sample $i \in \{1, \ldots, n\}$:

$$\mathcal{L}_{w\text{-}\ell_1} = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T} d(t, \tilde{t}^{(i)})^2 |\delta_t^{(i)}| \tag{5}$$

where, for the i-th sample, $\tilde{t}^{(i)}$ is chosen as the time step with the highest absolute value perturbation $\tilde{t}^{(i)} = \text{argmax}_t\, |\delta_t^{(i)}|$. To avoid biasing $\tilde{t}$ towards the center, we use the modulo distance $d(t, \tilde{t}) = \min\left((t - \tilde{t})\%T, (\tilde{t} - t)\%T\right)$ which treats the time samples as a circular list. Besides enforcing sparsity, penalizing the entries of δ also enforces the proximity (similarity) between x_{CF} and x.

4 RESULTS

In this section, we analyse the class transitions induced by the counterfactual generation process and discuss some examples of generated counterfactual explanations. In the appendix, we also discuss per-class average perturbations (sec. A.1), assess the plausibility of the generated counterfactual examples via an anomaly detection strategy (sec. A.2) and perform an ablation analysis (sec. A.3).

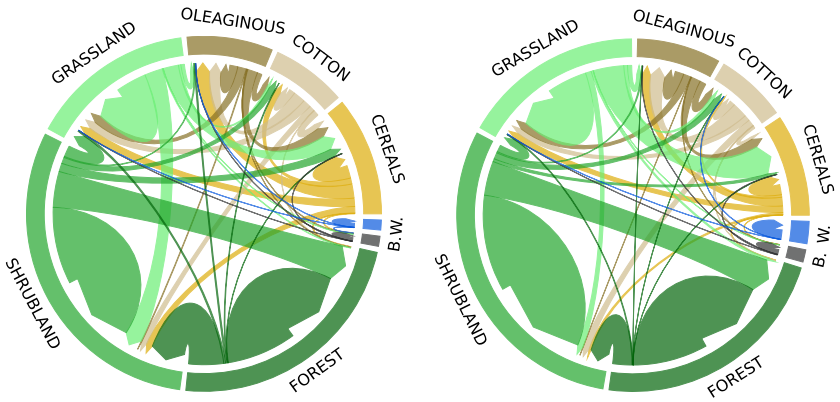

Figure 3: Summary of class transitions induced by the counterfactuals. Training data (left) and test data (right), where B. stands for *Bare Soil* and W. for *Water* classes.

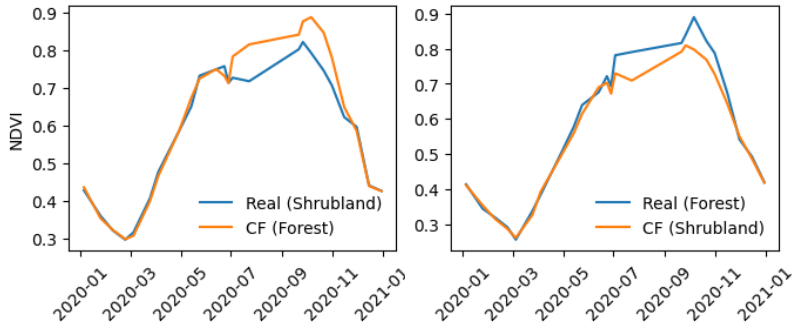

Figure 4: Examples of original time series with corresponding counterfactual from classes *Shrubland* (4) and *Forest* (5) on both ways.

Experimental setup The *Koumbia* study case described in Section 2 was split into training, validation and test sets containing respectively 50-17-33% of the 79961 samples. Each data sample corresponds to a (univariate) NDVI time series with 24 time samples. First, the *Classifier* was trained over 1000 epochs with batch size 32 and Adam optimizer with learning rate 10^{-4} and weight decay of same value. The model weights corresponding to the best obtained F1-score on the validation set were kept. Then, with the classifier weights frozen, the *Noiser* and *Discriminator* modules are simultaneously trained over 100 epochs with batch size 128 and Adam optimizer. Finally, we empirically set $\lambda_{\text{gen}} = 5{\cdot}10^{-1}$ and $\lambda_{w\text{-}\ell_1} = 5{\cdot}10^{-2}$ on the reported results.

Visualizing class relationships Class transitions induced by the counterfactual samples are summarized in Fig. 3, with arrow widths proportional to the number of counterfactuals leading to that specific transition. The left (resp. right) graph was generated by feeding the network with each of the training (resp. test) data samples. They present very similar behavior, indicating that the proposed method generalizes well to unseen data. We recall that class transitions are not pre-defined; on the contrary, our method allows input samples to freely split-up into multiple target classes. Transitions obtained in such a way thus bring up valuable insights on the relation between classes.

Counterfactual examples Two illustrative examples of counterfactual explanations are shown in Fig. 4. It is interesting to observe the similarity between the generated counterfactual and a real data example from the same class (on the neighboring plot). To transform a *Shrubland* sample into a *Forest* one, NDVI is added between the months of July and October. The opposite is done to obtain the reverse transition, which is reassuring. One can verify that the obtained counterfactuals do look realistic besides differing from the real signal only on a contiguous time window. These two properties have been explicitly enforced via the losses in eqs. (4) and (5).

5 Conclusion

In this paper, we presented a new framework to generate counterfactual SITS samples of vegetation indices (NDVI) for the land cover classification task. The proposed method overcomes the restriction to apriori define the source and the target classes for the counterfactual generation process while leveraging adversarial learning to ensure realistic counterfactual samples. A possible future work would be to extend the framework to the case of multivariate time series satellite data.

REFERENCES

Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, 6, 2018.

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58:82–115, 2020.

Emre Ates, Burak Aksar, Vitus J. Leung, and Ayse K. Coskun. Counterfactual explanations for multivariate time series. In *International Conference on Applied Artificial Intelligence (ICAPAI)*, pp. 1–8, 2021. doi: 10.1109/ICAPAI49758.2021.9462056.

Eoin Delaney, Derek Greene, and Mark T Keane. Instance-based counterfactual explanations for time series classification. In *International Conference on Case-Based Reasoning*, pp. 32–47. Springer, 2021.

Soukaïna Filali Boubrahimi and Shah Muhammad Hamdi. On the mining of time series data counterfactual explanations using barycenters. In *ACM CIKM*, pp. 3943–3947. ACM, 2022. ISBN 9781450392365. doi: 10.1145/3511808.3557663.

Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, pp. 1–55, 2022.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5), Sep. 2019.

Riccardo Guidotti, Anna Monreale, Francesco Spinnato, Dino Pedreschi, and Fosca Giannotti. Explaining any time series classifier. In *IEEE International Conference on Cognitive Machine Intelligence (CogMI)*, pp. 167–176, 2020. doi: 10.1109/CogMI50398.2020.00029.

O. Hagolle, M. Huc, D. Villa Pascual, and G. Dedieu. A multi-temporal and multi-spectral method to estimate aerosol optical thickness over land, for the atmospheric correction of formosat-2, landsat, venμs and sentinel-2 images. *Rem. Sens.*, 7(3):2668–2691, 2015.

J. Inglada, A. Vincent, M. Arias, and B. Tardy. iota2-a25386, July 2016. URL `https://doi.org/10.5281/zenodo.58150`.

J. Inglada, A. Vincent, M. Arias, B. Tardy, D. Morin, and I. Rodes. Operational high resolution land cover map production at the country scale using satellite image time series. *Remote. Sens.*, 9(1): 95, 2017.

A. Jolivot, V. Lebourgeois, L. Leroux, M. Ameline, V. Andriamanga, B. Bellón, M. Castets, A. Crespin-Boucaud, P. Defourny, S. Diaz, M. Dieye, S. Dupuy, R. Ferraz, R. Gaetano, M. Gely, C. Jahel, B. Kabore, C. Lelong, G. le Maire, D. Lo Seen, M. Muthoni, B. Ndao, T. Newby, C. L. M. de Oliveira Santos, E. Rasoamalala, M. Simoes, I. Thiaw, A. Timmermans, A. Tran, and A. Bégué. Harmonized in situ datasets for agricultural land use mapping and monitoring in tropical countries. *Earth System Science Data*, 13(12):5951–5967, 2021. doi: 10.5194/essd-13-5951-2021.

Jana Lang, Martin Giese, Winfried Ilg, and Sebastian Otte. Generating sparse counterfactual explanations for multivariate time series. *arXiv preprint arXiv:2206.00931*, 2022.

Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. *AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018.

Peiyu Li, Soukaina Filali Boubrahimi, and Shah Muhammad Hamd. Motif-guided time series counterfactual explanations. *arXiv preprint arXiv:2211.04411*, 2022.

C. Pelletier, G. I. Webb, and F. Petitjean. Temporal convolutional neural network for the classification of satellite image time series. *Remote. Sens.*, 11(5):523, 2019.

Arnaud Van Looveren, Janis Klaise, Giovanni Vacanti, and Oliver Cobb. Conditional generative models for counterfactual explanations. *arXiv preprint arXiv:2101.10123*, 2021.

Sahil Verma, John Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*, 2020.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

Q. Yuan, H. Shen, T. Li, Z. Li, S. Li, Y. Jiang, H. Xu, W. Tan, Q. Yang, J. Wang, J. Gao, and L. Zhang. Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sensing of Environment*, 241:111716, 2020.

A Appendix

A.1 Average perturbation examples

Examples of average perturbation profiles for two different class transitions are depicted in Fig 5. It is interesting to notice how the perturbations correspond roughly to the opposite of each other, which is quite suitable since they correspond to opposite transitions between the same two classes.

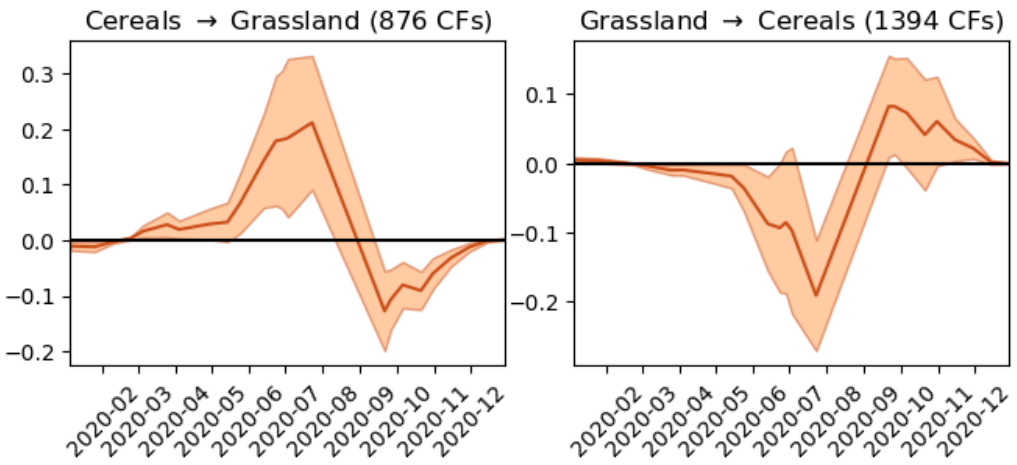

Figure 5: Examples of average counterfactual perturbations between classes *Cereals* and *Grassland* on both ways. Shaded area corresponds to the standard deviation.

A.2 Plausibility analysis

In this section, we quantify to what extent the proposed counterfactual explanations fit the original data distribution. To do so, we run an anomaly detection method, Isolation Forest (Li et al., 2018), on both the original data and corresponding counterfactuals. To attest the importance of the proposed adversarial training for the generation of realistic/plausible counterfactuals, we perform an ablation study confronting the proposed model trained with and without the generator loss in Eq. (4). Fig. 6 shows contingency matrices relating the isolation forest outputs on the original data (rows) and on the corresponding counterfactual explanations (columns). Two counterfactual generation approaches are investigated: the proposed method (left matrix) and its non-adversarial variant (right matrix). In the figures, diagonal entries correspond to matching isolation forest outputs –i.e., same prediction (inlier/outlier) for both real and counterfactual data. Later, in Table 2 we compute some metrics on such contingency matrices to further quantify and summarize the behaviour of the compared methods. The proposed counterfactual model achieves impressive results, even leading to more samples identified as inliers than the real data itself (23806 against 23755), since proposed approach converts less inliers into outliers (164) than the other way around (215).

The non-adversarial variant, on the other hand, obtains considerably more degraded results, as it converts as many as 4338 real inlier samples into outliers (about 20 times more). Such a gap becomes evident when looking at the corresponding accuracy and normalized mutual information (NMI) computed w.r.t. the isolation forest results on the original data (cf. Table 2). Such scores measure to what degree the inlier/outlier partitioning obtained on the counterfactual samples (for each of the two compared variants) matches the one obtained on the original data. The higher they are the better the two partitions match. The obtained results clearly show that counterfactual plausibility is achieved thanks to the adversarial training process.

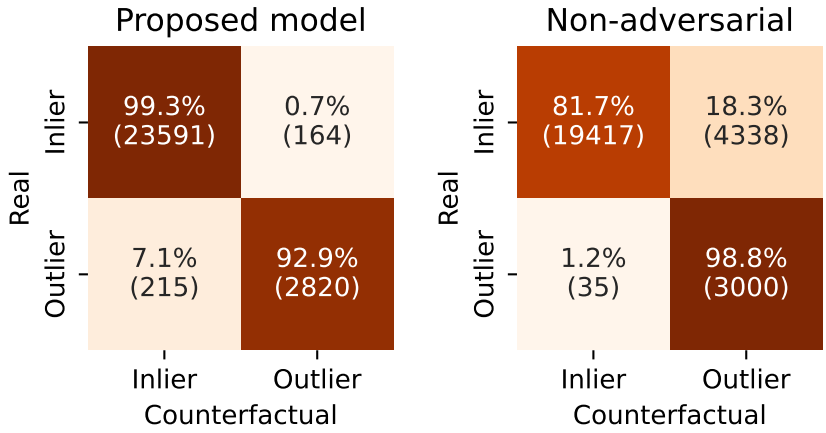

Figure 6: Isolation forest results on real (rows) and counterfactual data (columns). Proposed model with (left) and without (right) adversarial loss during training. Row-normalized percentages.

Method	Accuracy	NMI	Inliers ratio
Proposed	**98.6%**	**0.808**	**88.9%**
Non-adversarial	83.7%	0.337	72.6%

Table 2: Plausibility analysis using different performance metrics. Isolation Forest results on the real data were used as ground truth for the accuracy and NMI scores.

A.3 Other ablation studies

In Table 3 we compare the number of successful class-swapping counterfactual samples as well as the average ℓ_2 and ℓ_1 norms of the perturbations δ generated by the proposed model and two variants ignoring the generator loss ($\mathcal{L}_{\text{gen}}$) and the weighted-ℓ_1 loss ($\mathcal{L}_{w\text{-}\ell_1}$), respectively.

One can see that the removal of the auxiliary losses significantly bumps the class-swapping rate, but it happens at the expense of either: 1) counterfactual plausibility, as shown in the Section A.2 for the removal of $\mathcal{L}_{\text{gen}}$; 2) counterfactual proximity/similarity, as demonstrated by the dramatic increase on the norm of the generated perturbations (or, equivalently, the distance between x and x_{CF}) upon removal of $\mathcal{L}_{w\text{-}\ell_1}$.

Method	Class-swap CF	Average $\|\delta\|_2$	Average $\|\delta\|_1$
Proposed	43.8%	0.24 ± 0.18	0.76 ± 0.54
Without $\mathcal{L}_{\text{gen}}$	83.7%	0.97 ± 0.47	1.69 ± 0.99
Without $\mathcal{L}_{w\text{-}\ell_1}$	99.6%	4.79 ± 0.07	23.3 ± 0.53

Table 3: Ablation study on test data.