Supplementary Materials for paper: Weak-to-Strong 3D Detection with X-Ray Distillation

In these supplementary materials we provide detailed implementation insights: examples of both positive and negative Temporal Object Fusion, visualizations of NuScenes, Waymo and ONCE Object-Complete Frames, and additional models comparisons to provide a comprehensive overview of the work.

The supplementary materials are organized in the following way:

1. Section 1 demonstrates the performance improvements by the Semi-Supervised X-Ray Teacher method over the baseline model on the ONCE dataset, using the default training parameters provided by [6]. This improvement is notable not only with the refined training parameters [2] utilized in main work but also when applying the default parameters that were commonly used in previous research [7, 8].
2. Additional experiments and visualizations of the Object Temporal Fusion block can be found in Section 2.
3. Section 3 shows detailed information on the Object-Complete frames preprocessing and training details of the Supervised X-Ray Teacher.
4. Section 4 presents a selection of randomly chosen Object-Complete and original frames from the NuScenes , Waymo, and ONCE datasets to provide a comparative view of how these frames typically differ.

1. Additional Semi-Supervised Evaluation

In the main text, we adopted the training parameters from our previous work [2] for training our model on the ONCE dataset. This particular paper revealed that the training parameters previously used to achieve state-of-the-art (SOTA) results were suboptimal during the pretraining stage, leading to an unfair comparison of semi-supervised methods. Nevertheless, we also present results using the default parameters proposed in [3] to illustrate the robustness of our methods under various conditions and to enable a direct comparison of metrics with those reported in previous SOTA works. Other than this, we do not modify our method in any way.

2. Object Temporal Fusion

In this section, we provide a bit more detailed analysis and comparison of various Point Cloud Registration (PCR) methods utilized in the Object Temporal Fusion block. Initially, we replicated an experiment to compare different registration methods using default training parameters: the naive geometric approach, Greedy Grid [1], and GeDi [4], as outlined in Table 2. The last method's superior performance is largely due to its ability to minimize noise effects in the rotation and coordinates of boxes. For instance, the SECOND model often produces noisy boxes on unlabeled samples, leading to misaligned objects with basic alignment methods. The comparative visualizations are shown in Figure 1. From these visualizations, it is evident that both the geometric and GeDi methods encounter challenges in certain scenarios, such as with car C, while cars A and B exhibit notably better alignment using the GeDi method. This enhanced alignment is a result of advanced PCR process, which effectively reduces noise from the 3D detector. However, since registration applies corrections sequentially, an error in one iteration could lead to amplified errors in subsequent ones. This is why GeDi fails with car C but succeeds in transforming cars A and B into well-aligned, nearly complete objects, maintaining their original rotation. The Greedy Grid method is omitted in this analysis as its performance improvement is not on par with that of GeDi, and it generally aligns with the geometric approach in terms of overall quality.

3. Supervised Setting Implementation

This section discusses the specifics of Object-Completion Sampling, a technique used to optimize the processing of shape-complete objects, and provides insight into the training methodology for the Supervised X-Ray Teacher.

3.1. Object-Completion

Object-Completion is designed to utilize all available frames within a scene to create the most comprehensive representation of an object from every possible viewing angle. However, in scenarios where scenes are extended and contain numerous objects that remain within view over many

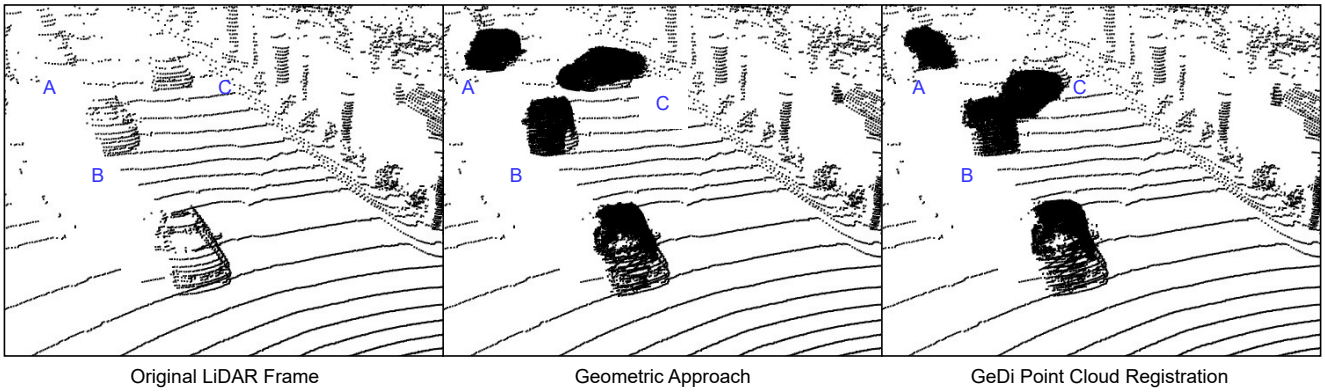

Figure 1. Visual comparison between the geometric approach (box alignment) and the GeDi [4] Point Cloud Registration technique within the Objects Temporal Fusion module. We evaluate their effectiveness in merging different views of an object into a unified point cloud. Although both methods encounter difficulties with certain objects such as car C, the GeDi method generally provides superior alignment for other vehicles, like cars A and B. This indicates that while the GeDi technique is more effective at aligning object points, it also highlights the necessity for domain adaptation to prevent the errors, as observed with car C. The illustration underscores the potential of advanced registration methods in enhancing object detection while also pointing to the need for further refinement to ensure consistent accuracy across all objects.

frames, the size of the Object-Complete Frames can become exceedingly large. Specifically, in the Waymo dataset [5], a single frame might exceed 150MB, which can significantly slow down the training process. Moreover, having an excessive number of points for one scene can be unnecessary. To address this, we employ a straightforward sampling strategy for the Waymo dataset: we split the object-complete cloud into two distinct parts — the original frame and all newly added points. We then sample a volume 1.5 times the size of the original cloud from the new points and concatenate these samples with the original points. As for the NuScenes dataset, which typically generates much smaller Object-Complete frames, we do not implement any sampling strategy.

3.2. Training

Teacher. We train teacher networks from scratch: it never saw any original point cloud, only Object-Complete ones. We set default hyperparameters except for batch size - we match total number of samples processed simultaneously, so we make it equal 8. **Knowledge Distillation.** We train students with the same configuration as if we were training teacher or original model. There's definitely a better configuration for our case and further researchers might also improve our results by just adjusting some hyperparameters.

4. Visualization

In this section, we present a series of visual comparisons that highlight the effectiveness of our Object-Complete Frame Generation process. Figures 2, 3, and 4 illustrate the pronounced differences between original frames and

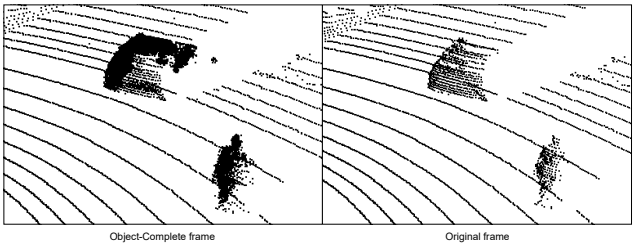

Figure 2. This figure provides a comparison of randomly selected objects from the ONCE validation set, showing the difference between Object-Complete frames (left) and original frames (right). The Object-Complete frame demonstrate the enhanced detail achieved through our frame generation process, which collects comprehensive point cloud data to construct a more complete representation of each object. This enhanced representation helps reduce the ambiguity typically associated with sparse LiDAR data, resulting in more accurate object detection.

those enhanced by Object-Complete Frame Generation, using examples from various real-world autonomous driving datasets. These comparisons visually demonstrate the advantages of our method in terms of data richness and object detection clarity. By significantly reducing sparsity and occlusions, the generated Object-Complete frames offer a more accurate and unambiguous representation of the scene, as can be seen in the enhanced details of the objects. These visualizations serve to illustrate the practical benefits of our approach, reinforcing the validity of our contributions to the field of LiDAR-based 3D object detection.

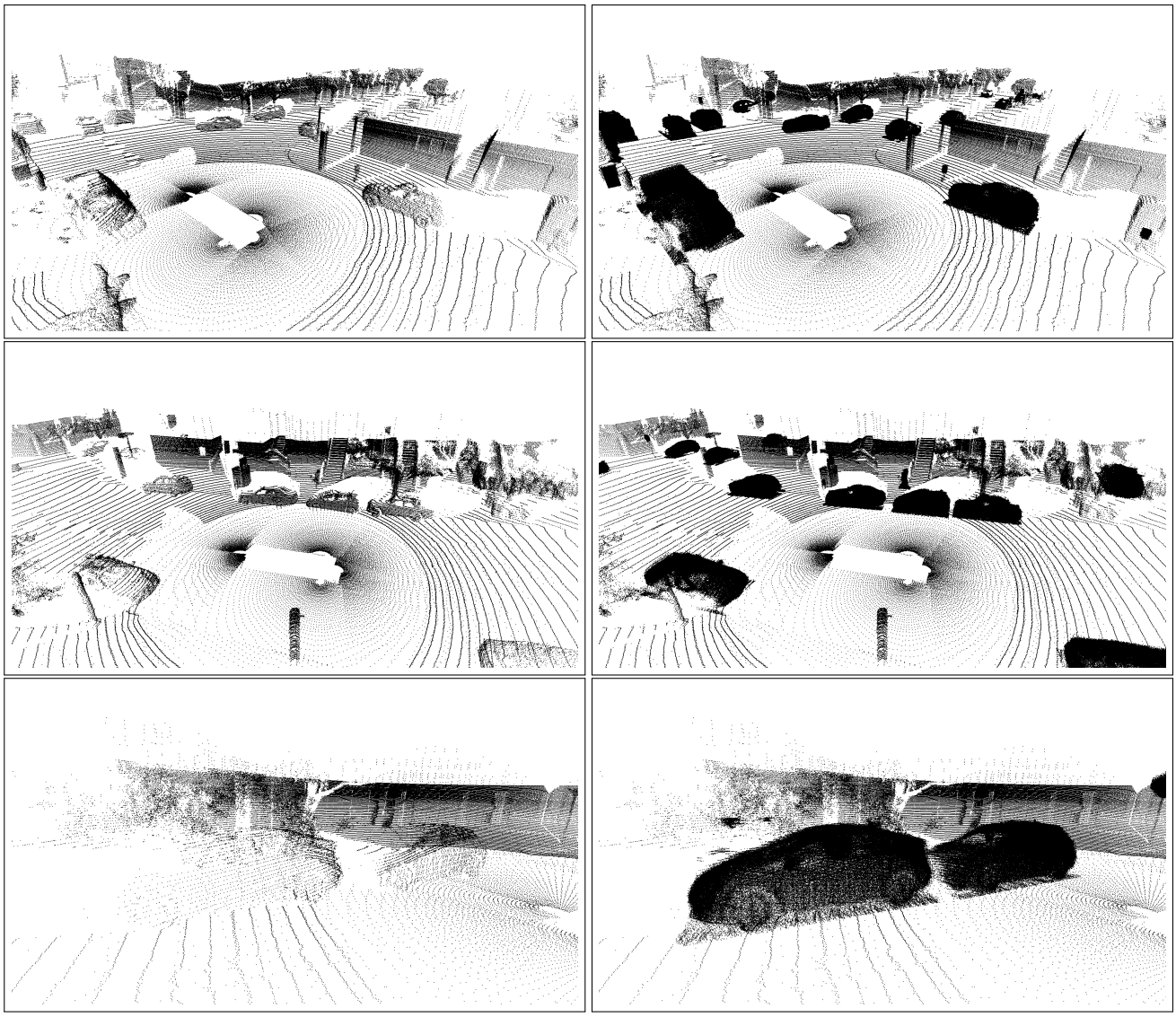

Figure 3. Visual comparison between original (left) and Object-Complete (right) frames from the Waymo dataset. This figure shows how Object-Complete Frame Generation enriches point cloud data. This enhancement significantly diminishes sparsity and occlusions, thereby reducing ambiguity and making shape-complete objects easier to detect.

References

[1] David Bojanić, Kristijan Bartol, Josep Forest, Stefan Gumhold, Tomislav Petković, and Tomislav Pribanić. Challenging the universal representation of deep models for 3d point cloud registration. In *BMVC 2022 Workshop Universal Representations for Computer Vision*, 2022. 1, 5

[2] Maksim Golyadkin, Alexander Gambashidze, Ildar Nurgaliev, and Ilya Makarov. Refining the once benchmark with hyperparameter tuning. *arXiv preprint arXiv:2311.06054*, 2023. 1

[3] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Hanxue Liang, Jingheng Chen, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, et al. One million scenes for autonomous driving: Once dataset. *arXiv preprint arXiv:2106.11037*, 2021. 1

[4] Fabio Poiesi and Davide Boscaini. Learning general and distinctive 3d local deep descriptors for point cloud registration. In *IEEE Trans. on Pattern Analysis and Machine Intelligence*, (early access) 2022. 1, 2, 5

[5] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June

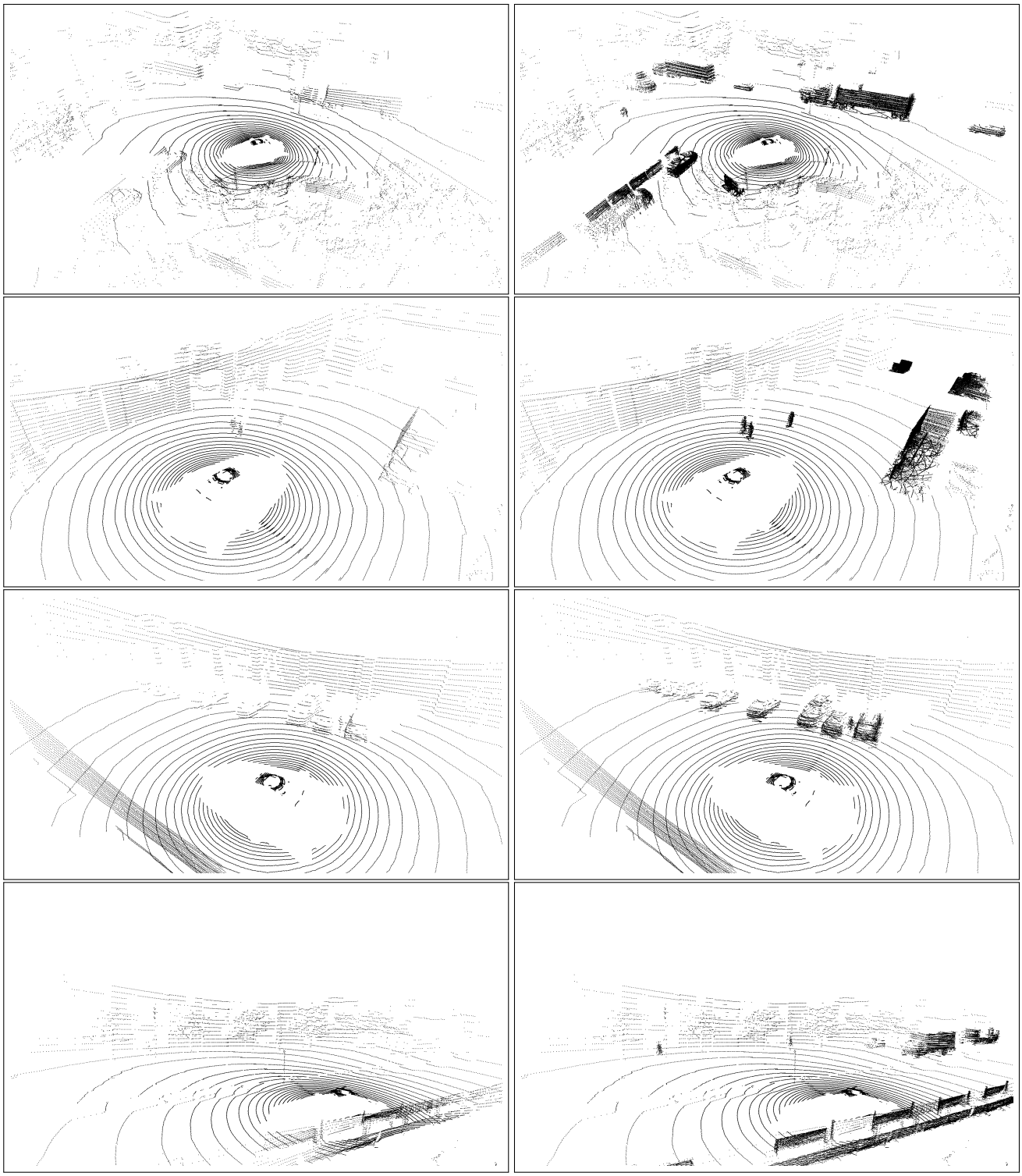

Figure 4. Visual comparison between original (left) and Object-Complete (right) frames from the NuScenes dataset. This figure shows how Object-Complete Frame Generation enriches point cloud data. This enhancement significantly diminishes sparsity and occlusions, thereby reducing ambiguity and making shape-complete objects easier to detect.

Table 1. Comparison of the performance of X-Ray Teacher in semi-supervised setting vs other methods using SECOND baseline model with a default configuration. Models were trained on different splits of unlabeled data (Small, Medium, Large) and evaluated on the ONCE validation split with Mean Average Precision (mAP). The integration of X-Ray Teacher with the Mean Teacher and Proficient Teacher methods is referred to as X-Ray MT and X-Ray PT, respectively. Higher metric values indicate superior model performance in 3D Object Detection. The best results are highlighted in **bold**. Values in parentheses indicate the performance difference between the original and X-Ray approaches. Our approach consistently outperforms the state-of-the-art for Semi-Supervised 3D Object Detection in terms of mAP across all splits.

Method	SECOND
Train (5k labeled samples)	
Pretraining	51.89
Small (100k unlabeled samples)	
Pseudo Label	51.22 (-0.67)
Noisy Student	52.39 (+0.50)
Mean Teacher	55.34 (+3.45)
SESS	53.39 (+1.50)
3DIoUMatch	53.81 (+1.92)
NoiseDet	58.00 (+6.11)
Proficient Teacher	57.72 (+5.83)
X-Ray Teacher (ours)	**59.65**(+7.76)
Medium (500k unlabeled samples)	
Pseudo Label	50.40 (-1.49)
Noisy Student	55.34 (+3.45)
Mean Teacher	58.27 (+6.38)
SESS	55.79 (+3.90)
3DIoUMatch	56.25 (+4.36)
NoiseDet	60.06 (+8.17)
Proficient Teacher	59.89 (+8.00)
X-Ray Teacher (ours)	**62.42** (+10.53)
Large (1M unlabeled samples)	
Pseudo Label	49.76 (-2.13)
Noisy Student	56.37 (+4.48)
Mean Teacher	59.28 (+7.39)
SESS	57.99 (+6.10)
3DIoUMatch	57.07 (+5.18)
NoiseDet	61.16 (+9.27)
Proficient Teacher	61.40 (+9.51)
X-Ray Teacher (ours)	**63.57** (+11.68)

Table 2. Comparison of different registration methods for Object Complete Frame Generation. We used the default SECOND model in the semi-supervised setting with X-Ray Teacher (our modification of Mean Teacher without Exponential Moving Average) on three unlabeled splits (Small, Medium, Large) processed with three different registration methods and evaluated it on the ONCE validation split with Mean Average Precision (mAP). Higher metric values indicate superior model performance in 3D Object Detection. The best results are highlighted in **bold**. Our analysis shows that the choice of registration method has a noticeable impact on the performance of X-Ray Teacher. The GeDi registration method consistently outperforms the other techniques across all data splits, achieving the highest mAP scores. This underlines the importance of sophisticated registration techniques in the generation of more accurate and complete point clouds.

Small	
Method	**mAP**
Box Geometry	59.04
Greedy Grid [1]	59.11
GeDi (our preprocessing)[4]	**59.65**
Medium	
Box Geometry	62.03
Greedy Grid [1]	62.31
GeDi (our preprocessing) [4]	**62.42**
Large	
Box Geometry	62.56
Greedy Grid [1]	62.81
GeDi (our preprocessing) [4]	**63.57**

2020. 2

[6] OpenPCDet Development Team. Openpcdet: An open-source toolbox for 3d object detection from point clouds. https://github.com/open-mmlab/OpenPCDet, 2020. 1

[7] He Wang, Yezhen Cong, Or Litany, Yue Gao, and Leonidas J. Guibas. 3dioumatch: Leveraging iou prediction for semi-supervised 3d object detection. In *CVPR*, 2020. 1

[8] Junbo Yin, Jin Fang, Dingfu Zhou, Liangjun Zhang, Cheng-Zhong Xu, Jianbing Shen, and Wenguan Wang. Semi-supervised 3d object detection with proficient teachers. In *European Conference on Computer Vision*, pages 727–743. Springer, 2022. 1