拡散モデルで生成した擬似実データに基づく半教師あり LiDAR セグメンテーション

○宮脇 智也（九州大学），中嶋 一斗（九州大学），岩下 友美（JPL），倉爪 亮（九州大学）

Semi-supervised LiDAR Segmentation Based on Pseudo-Real Data Generated by Diffusion Models

○ Tomoya MIYAWAKI (Kyushu Univ.), Kazuto NAKASHIMA (Kyushu Univ.), Yumi IWASHITA (JPL), Ryo KURAZUME (Kyushu Univ.)

Abstract: In LiDAR point cloud segmentation tasks, the high cost of constructing labeled data is a significant challenge. To address this issue, methods that leverage labeled data generated by simulators have gained attention. However, domain gaps between simulated and real data often degrade generalization performance. To mitigate this, we have developed a method that uses diffusion models to transform simulation data into more realistic pseudo-real data. In this paper, we explore a semi-supervised scenario where this pseudo-real data is mixed with real data for training. By evaluating various mixing ratios and loss designs, we aim to demonstrate the effectiveness of pseudo-real data in segmentation performance.

1. 緒言

距離センサの一種である 3D LiDAR センサは，自律移動ロボットや自動運転車に広く利用されており，周囲環境を認識する上で重要な役割を果たす．特に，3D LiDAR センサで取得した点群データを個々の物体や領域に分類するセマンティックセグメンテーションタスクは，ロボット工学およびコンピュータビジョン分野の中心的タスクとして取り組まれてきた．しかし，セマンティックセグメンテーションタスクでは，学習に必要な大量のラベル付き点群を作成するために，膨大な時間とリソースを要することが大きな課題となっている．代表的な大規模ベンチマークデータセットである SemanticKITTI[1] は，ラベル付け作業に 1700 時間以上費やされたと報告されている．

この問題に対する解決策の一つとして Sim2Real が注目されている．Sim2Real は，シミュレータ上で自動生成したラベル付きシミュレーションデータを用いて認識モデルを学習し，実環境に適用する手法である．これにより，高品質かつ大量のラベル付けが可能となる．一方で，一般に学習に利用するシミュレーションデータとテストする実データの間には，物体の形状の違いだけでなく，欠損ノイズや反射強度の有無といった物理的・計測的要因によるデータ特性のギャップが存在する．このため，実環境における認識モデルの性能が著しく低下することが多い．

この Sim2Real の性能低下を防ぐため，我々の研究では，実データ特有の特徴をシミュレーションデータに再現するドメイン変換手法を開発し，両データ間のギャップを縮小することで，Sim2Real の適用性能を向上させることを目指している．具体的には，生成モデルの一種である拡散モデルを用い，シミュレーションデータに対して実データ特有の特徴である欠損ノイズ，および反射パルスレーザの強度である反射強度を再現し，より現実的な擬似実データへ変換する手法を開発している[2]．本稿では，我々の手法で生成した擬似実データを実データと混合して学習に用いる半教師ありシナリオを検討し，混合比率や損失設計を多角的に評価することで，擬似実データの有効性を明らかにする．

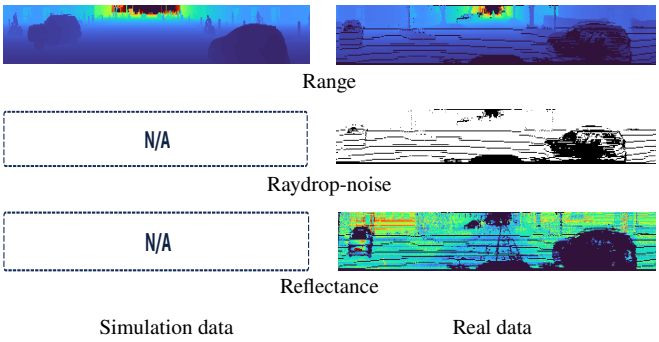

Fig. 1 Comparison between simulation data[3] and real data[4] of LiDAR range images (partial azimuth angles extracted for display)

2. 欠損ノイズと反射強度のモデル化

LiDAR データにおける欠損ノイズとは，照射したレーザ光が物体表面で拡散・減衰することで，反射光の検知に必要な受光強度が十分に得られず発生するデータ欠損を指す．また，反射強度とは，LiDAR が計測する反射パルスレーザの強度を示す．これらの特徴は，照射される物体の材質や入射角によって複雑に変化するため，物理パラメータを同定しシミュレータ上で再現することは難しい．そのため，Fig. 1 に示すように，シミュレーションデータには欠損ノイズ・反射強度が存在しない場合が多い．一方で，欠損ノイズの分布の違いは LiDAR データの Sim2Real において性能低下を引き起こすことが知られている[3]．また，反射強度は今回対象とするセマンティックセグメンテーションタスクにおいて，距離値とともに入力として用いられることが多い．そのため，LiDAR Sim2Real の適応性能向上にはこれら 2 つの特徴の正確な復元が求められる．

これまでに，距離画像表現されたシミュレーションデータに対して欠損ノイズを再現する手法が提案されている．Zhao ら[5] は，教師なし画像変換手法である CycleGAN を用いることで，シミュレーションデータに

対して欠損画像を再現する手法を提案した．Nakashima ら[6] は LiDAR 距離画像の距離値と欠損確率 p の共起関係を学習した敵対的生成ネットワーク (GAN) によって欠損ノイズを再現する手法を提案している．Wu ら[3] は，実データの距離画像から反射強度画像を推定する教師あり学習によって，シミュレーションデータの反射強度再現を提案している．Xiao[7] らは，実データ・シミュレーションデータの敵対的学習に基づく点群変換と欠損画像生成，実データの教師あり学習による反射強度生成による 3 段階の Sim2Real 手法を提案している．一方，これらの手法によって欠損ノイズや反射強度などの特徴が付与されたシミュレーションデータは，実データと比べて未だギャップがあり，高精度な特徴再現手法が必要である．

3. 拡散モデルに基づく疑似実データ生成

本稿では，距離画像表現に基づく LiDAR データの拡散モデルを利用し，欠損ノイズ・反射強度を再現した疑似実データを生成する手法[2] を紹介する．

3.1 LiDAR 距離画像の拡散モデル

本研究では，LiDAR データの拡散モデルとして，R2DM[8] を採用する．R2DM は，一般的な拡散モデルと同様に，ノイズを加える拡散過程とノイズを除去する逆拡散過程の 2 つの過程から構成される．R2DM の特徴として，LiDAR のビーム角度に関する帰納バイアスを導入した連続時間 DDPM アーキテクチャを採用している点が挙げられる．逆拡散過程は，確率微分方程式 (SDE) として以下のように表される．

$$d\boldsymbol{x}_t = \left[f(\boldsymbol{x}_t, t) + \frac{1}{2} g^2(t) \nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t) \right] dt + g(t) d\bar{\boldsymbol{w}}, \quad (1)$$

ここで，状態変数 $\boldsymbol{x}_t$ は連続時間 $t \in [1, 0]$ で発展し，$f(\boldsymbol{x}_t, t)$ および $g(t)$ は拡散モデルの種類に応じて設計される係数関数，$\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t)$ はニューラルネットワークによってモデル化される学習可能なスコア関数，$\bar{\boldsymbol{w}}$ は標準ウィーナープロセスである．生成されるデータは，LiDAR 点群の距離画像および反射強度画像の 2 チャネルで表現される．本研究では，31M パラメータの公開実装を 285M パラメータに増強し，KITTI Raw データセット[4] を用いて学習を行った．

3.2 事後分布サンプリングによる条件付き生成

データの条件なし生成を学習した拡散モデルは，Eq. (1) におけるスコア関数を以下のように変形することで，その生成過程を特定の参照データや指標にしたがって制御する条件付き生成モデルとして応用可能である．

$$\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t \mid \boldsymbol{y}) = \nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t) + \nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{y} \mid \boldsymbol{x}_t), \quad (2)$$

式変形にはベイズの定理 $p(\boldsymbol{x} \mid \boldsymbol{y}) \propto p(\boldsymbol{x}) p(\boldsymbol{y} \mid \boldsymbol{x})$ を用いている．第 1 項の事前分布項は Eq. (1) に基づき $\boldsymbol{x}$ のみで学習可能な条件なしスコアであり，第 2 項の尤度項は観測方程式に基づいて解析的に計算することができる．本手法では，シミュレーションデータ $\boldsymbol{y}$ と生成したい疑似実データ $\boldsymbol{x}$ の間に以下の観測方程式を仮定する．

$$\boldsymbol{y} = H\boldsymbol{x} + z, \quad (3)$$

H は反射強度チャネルを欠落させる欠損行列，z はガウスノイズである．この観測方程式に基づき，尤度項を以下のように定義する．

$$\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{y} \mid \boldsymbol{x}_t) \approx r_t^{-2} \left[\left(H^\dagger \boldsymbol{y} - H^\dagger H \hat{\boldsymbol{x}}_t \right)^\top \frac{\partial \hat{\boldsymbol{x}}_t}{\partial \boldsymbol{x}_t} \right]^\top, \quad (4)$$

ここで $\hat{\boldsymbol{x}}_t$ は Tweedie の公式[9] を用いて $\boldsymbol{x}_t$ から推定した $\boldsymbol{x}_0$ の暫定値であり、$H^\dagger$ は H のムーア・ペンローズ擬似逆行列である。

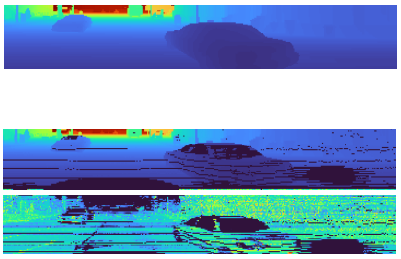

Fig. 2 Examples of pseudo-real data. we can see that our method can successfully reproduce the missing noise and reflectance intensity characteristics of real data while maintaining the scene structure of the simulation data.

3.3 適応的なマスキングによる欠損ノイズ・反射強度の再現

Eq. (3) の定式化は反射強度に関する順方向の劣化 $\boldsymbol{y} \leftarrow \boldsymbol{x}$ は扱えるが，欠損ノイズの逆方向劣化 $\boldsymbol{y} \rightarrow \boldsymbol{x}$ は考慮できない．このため，Eq. (4) の尤度項を全画素に適用すると，欠損ノイズを含まないシミュレーションデータ $\boldsymbol{y}$ との整合性を取るために，拡散モデルが生成しようとする自然な欠損パターンが抑制され，過剰に補間された不自然なデータが生成されてしまう．この問題に対し，ここでは Tweedie 推定 $\hat{\boldsymbol{x}}_t$ から生成されるマスク $\boldsymbol{m}_t$ を用いて尤度項を適応的にマスク化する．Tweedie 推定 $\hat{\boldsymbol{x}}_t$ は，拡散モデルの事前分布が学習したシーン構造や欠損パターンを反映しているため，その距離値に基づいてマスク $\boldsymbol{m}_t$ を構築する．これにより，尤度項を以下のように修正する．

$$\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t \mid \boldsymbol{y}) = \nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t) + \boldsymbol{m}_t \odot \nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{y} \mid \boldsymbol{x}_t), \quad (5)$$

この操作により，マスク $\boldsymbol{m}_t$ でマスクされた画素では拡散モデルによって生成された欠損ノイズを保持しつつ，それ以外ではシミュレーション入力との整合性を保つことができる．生成データの例を Fig. 2 に示す．元のシミュレーションデータのシーン構造を維持しつつ，実データに近い欠損ノイズと反射強度の特徴が再現されていることがわかる．

4. 半教師あり LiDAR セグメンテーション

本実験では，拡散モデルを用いた LiDAR ドメイン変換手法で生成した擬似実データを，実際の LiDAR データと混合して学習に用いる半教師ありシナリオでのセグメンテーション性能を評価する．

4.1 実験設定

本実験にて使用するデータセットを Table 1 に示す．シミュレーションデータセットとして，SynLiDAR[7] を使用する．SynLiDAR は，Unreal Engine 4 ベースの仮想空間で生成された 32 クラスのラベル付き点群データセットであり，198,396 フレームの点群データを含む．本シミュレーションデータセットにあらかじめ Sec. 3

Table 1 Dataset used in our experiments

Dataset	Domain	#Classes	#Samples	Resolution
SynLiDAR[7]	Simulation	32†	198,396	64 × 1024
SemanticKITTI[10]	Real	25†	43,552	64 × 1024

† We use the common 19 classes.

で説明した手法を適用し，擬似実データセットを生成しておく．実データセットとして，SemanticKITTI[10]を使用する．SemanticKITTI は，実際の都市環境で収集された 25 クラスのラベル付き点群データセットであり，43,552 フレームの点群データを含む．training, validation, test の 3 つのサブセットに分割されており，本実験では training セットを学習に，validation セットを評価に用いる．本実験では，両データセットで共通する 19 クラスを使用し，19 クラスのセマンティックセグメンテーションタスクを評価する．セマンティックセグメンテーションを行うモデルには，RangeNet[11]を用いる．結果の評価には，推定された領域と真値の領域の重畳度を示す intersection-over-union (IoU) を算出する．

4.2 比較手法

本稿では，実データと擬似実データを混合して学習に用いる際の混合比率，ミニバッチ作成手法および損失設計を多角的に評価する．まず混合比率に関して，実データの training セット（19,130 フレーム）を 1%, 20%, 50%, 100%の 4 段階でサンプリングし，それぞれに対して擬似実データセット全量を加えて学習データセットを構成する．学習戦略として，単純な混合学習（Config-A）に加え，実データと擬似実データの学習をマルチタスク学習の枠組みで扱う 2 つの手法（Config-B, C）を評価する．

- **Config-A 単純混合 (Naive Mix)**: 実データと擬似実データを区別なく混合し，ランダムにミニバッチを作成して学習する．
- **Config-B 均等混合 (Balanced Mix)**: ミニバッチを実データと擬似実データから常に同数ずつサンプリングして構成する．損失は各ドメインで個別に計算し，固定の重み（0.5）で和を取る．損失は以下のように定義される．

$$\mathcal{L} = 0.5\mathcal{L}_r + 0.5\mathcal{L}_s, \tag{6}$$

  ここで，$\mathcal{L}_r$ および $\mathcal{L}_s$ はそれぞれ実データと擬似実データの損失である．
- **Config-C 不確実性重み付け (Uncertainty Weighting)**: ミニバッチの構成は均等混合と同様にし，各ドメインの損失の重みを不確実性に基づいて学習中に動的に調整する[12]．損失は以下のように定義される．

$$\mathcal{L} = \frac{1}{2\sigma_r^2}\mathcal{L}_r + \frac{1}{2\sigma_s^2}\mathcal{L}_s + \log\sigma_r + \log\sigma_s, \tag{7}$$

  ここで，$\mathcal{L}_r$ および $\mathcal{L}_s$ はそれぞれ実データと擬似実データの損失，σ_r および σ_s は各ドメインの不確実性を表すパラメータであり，学習中に最適化される．

また，ベースラインとして，実データの training セットを 1%, 20%, 50%, 100%の 4 段階でサンプリングしたデータセットのみを用いて学習する手法も評価する．

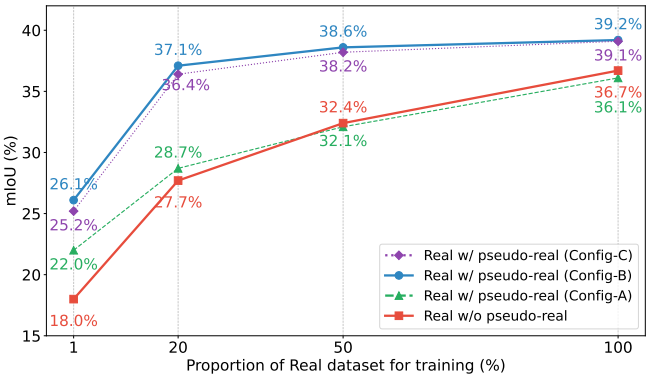

Fig. 3 Comparison of semi-supervised learning strategies. We show mIoU (%) scores on the 19-class semantic segmentation task. The plot compares a baseline (training on real data only) against three proposed configurations (Config-A, B, C) that mix real data with our pseudo-real data at various ratios.

4.3 実験結果

Fig. 3 に，各手法で半教師あり学習を行った場合のセグメンテーション性能を示す．横軸が実データの使用割合，縦軸が 19 クラスの IoU の平均値である mIoU を表す．まず，Config-A の単純混合では，実データの使用割合が増加するにつれて性能が向上するものの，100%使用した場合には実データのみで学習した場合と比較して性能が低下してしまった．一方，ミニバッチを実データと擬似実データから均等にサンプリングして構成し，マルチタスク学習の枠組みで損失を設計した Config-B および Config-C では，実データを 20%混ぜた段階で，実データのみを 100%使用した場合と同等の性能を達成した．また，実データを 100%使用した段階では，Config-B および Config-C はそれぞれ mIoU 39.2%および 39.1%を達成し，実データのみで学習した場合（mIoU 36.7%）と比較して性能が向上した．これは，単純混合（Config-A）では学習が擬似実データに偏る一方，均等サンプリング（Config-B, C）では両ドメインからバランス良く学習できるため，擬似実データを有効なデータ拡張として活用し，実データへの汎化性能を向上できたためと考えられる．また，Config-B と Config-C の性能はほぼ同等であり，今回の実験設定では不確実性重み付け（Config-C）が特に有効であるとは言えなかった．

次に Table 2 に実データを 100%使用した場合の各クラスの IoU の詳細を示す．mIoU が最も高かった Config-B の性能向上の中でも特筆すべきクラスとして，歩行者 (person) クラスがある．実データのみでの学習では IoU が約 5%であったのに対し，擬似実データを混合することで約 20%まで性能が向上した．Fig. 4 に改善結果の一例を示す．Fig. 4 は Config-B で 100%実データを混ぜて学習した場合と実データのみで学習した場合の典型的なセグメンテーション結果を示している．実データのみで学習した場合（中央）では ■ person が ■ trunk として誤分類されている一方，擬似実データを加えることで正しく分類できていることがわかる．

Table 2 Quantitative comparison of Sim2Real methods on semantic segmentation (SynLiDAR → SemanticKITTI).

	Intersection-over-Union (IoU, %) ↑																			
semi-supervised method	Car	Bicycle	Motorcycle	Truck	Bus	Person	Bicyclist	Motorcyclist	Road	Parking	Sidewalk	Other-ground	Building	Fence	Vegetation	Trunk	Terrain	Pole	Traffic-sign	Mean (mIoU)
Baseline	89.5	0.6	12.1	34.2	20.7	5.2	28.8	0.0	92.0	41.2	75.2	0.2	71.7	26.4	70.5	30.7	68.2	24.3	6.4	36.7
Config-A	82.3	3.0	24.5	34.7	16.5	14.6	44.4	0.0	90.7	36.4	70.9	0.5	63.4	22.2	61.4	27.5	65.4	20.4	7.3	36.1
Config-B	88.7	1.1	34.1	31.5	21.2	20.5	51.1	0.0	92.4	42.0	76.0	0.3	68.9	23.6	66.6	31.1	63.1	26.0	7.5	39.2
Config-C	87.4	0.7	24.9	49.9	21.8	17.4	46.1	0.0	92.3	41.0	75.0	0.2	67.6	26.5	67.5	30.3	67.3	22.2	5.4	39.1

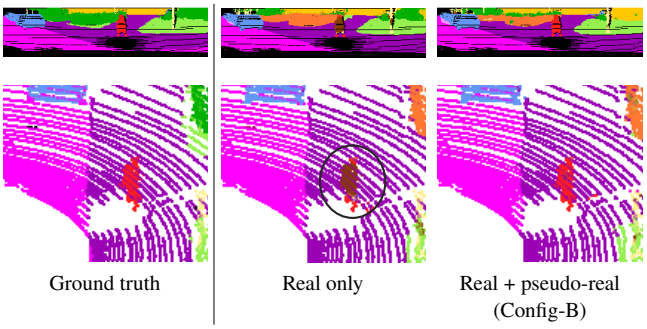

Fig. 4 Qualitative comparison of typical segmentation results on the semi-supervised setting. ■ person ■ trunk.

5. 結言

本稿では，拡散モデルを用いて生成した擬似実データを実データと混合して学習に用いる半教師ありシナリオを検討し，混合比率や損失設計を多角的に評価することで，擬似実データの有効性を検証した．実験の結果，擬似実データと実データを単純に混合するだけでは性能向上に繋がらず，学習戦略が重要であることが示唆された．両ドメインをマルチタスク学習の枠組みで扱うことで，少量の実データでも高い性能を達成し，最終的には実データのみで学習した場合を上回る性能を達成した．今後は，擬似実データの生成手法の改良や，より高度な半教師あり学習手法の検討を通じて，LiDAR セグメンテーションの性能向上を目指す．

謝辞

本研究は JSPS 科研費 JP23K16974 の助成を受けたものである．

参考文献

[1] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, J. Gall, and C. Stachniss: Towards 3D LiDAR-based semantic scene understanding of 3D point cloud sequences: The SemanticKITTI Dataset, The International Journal on Robotics Research (IJRR), 40.8-9, pp. 959–967 (2021).

[2] 宮脇 智也, 中嶋 一斗, 劉 瀟文, 岩下 友美, and 倉爪 亮: 拡散モデルの条件付き生成を用いた LiDAR データの Sim2Real ドメイン変換, 第 27 回 画像の認識・理解シンポジウム MIRU2024 (2024.8.6-9), IS-3–105.

[3] B. Wu, X. Zhou, S. Zhao, X. Yue, and K. Keutzer: SqueezeSegV2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a LiDAR point cloud, Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) (2019), pp. 4376–4382.

[4] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun: Vision meets robotics: The KITTI dataset, The International Journal of Robotics Research (IJRR), 32.11, pp. 1231–1237 (2013).

[5] S. Zhao, Y. Wang, B. Li, B. Wu, Y. Gao, P. Xu, T. Darrell, and K. Keutzer: ePointDA: An End-to-End Simulation-to-Real Domain Adaptation Framework for LiDAR Point Cloud Segmentation, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 4 (2021), pp. 3500–3509.

[6] K. Nakashima, Y. Iwashita, and R. Kurazume: Generative Range Imaging for Learning Scene Priors of 3D LiDAR Data, Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2023), pp. 1256–1266.

[7] A. Xiao, J. Huang, D. Guan, F. Zhan, and S. Lu: Transfer learning from synthetic to real LiDAR point cloud for semantic segmentation, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, 3 (2022), pp. 2795–2803.

[8] K. Nakashima and R. Kurazume: LiDAR Data Synthesis with Denoising Diffusion Probabilistic Models, Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) (2024), pp. 14724–14731.

[9] J. Ho, A. Jain, and P. Abbeel: Denoising diffusion probabilistic models, Advances in Neural Information Processing Systems (NeurIPS), vol. 33 (2020), pp. 6840–6851.

[10] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall: SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences, Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019), pp. 9297–9307.

[11] A. Milioto, I. Vizzo, J. Behley, and C. Stachniss: RangeNet++: Fast and accurate LiDAR semantic segmentation, Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2019), pp. 4213–4220.

[12] A. Kendall, Y. Gal, and R. Cipolla: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, Proceedings of the IEEE conference on computer vision and pattern recognition (2018), pp. 7482–7491.