Received: 10 November 2025 | Revised: 23 December 2025 | Accepted: 23 March 2026

DOI: 10.4218/etrij.2025-0458

ORIGINAL ARTICLE

ETRI Journal WILEY

SMOTE3D: Geometry- and color-aware volumetric data augmentation for 3D fashion asset segmentation

Jiyoun Lim | **Jeong-Woo Son** | **Alex Lee** | **Sun Joong Kim** | **Namkyung Lee** | **Wonjoo Park**

Media Intellectualization Research Section, Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea

Correspondence
Jiyoun Lim, Media Intellectualization Research Section, Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea.
Email: kusses@etri.re.kr

Funding information
Institute of Information & Communications Technology Planning & Evaluation (IITP), Grant/Award Number: RS-2023-00215700

Abstract

This study proposes a geometry- and color-aware 3D data-augmentation framework to enhance fashion asset segmentation for digital twin applications. This study focuses on converting 2D fashion videos into 3D point-cloud data and augmenting the minority classes. The constructed dataset comprised 502 mannequin-wearing scenes and additional single-asset captures, totaling 628 instances across 16 categories and 104 items, all recorded using an iPhone 14 Pro. To address class imbalance, three augmentation strategies are introduced: SMOTE3D (normal-aware coordinate interpolation with red–green–blue jitter), scene reconfiguration, and class-consistent color shifts. Using OctFormer, OA-CNNs, and PTv3, the augmented data consistently improved the accuracy, intersection over union (IoU), and mean average precision (mAP), yielding significant gains over a Mix3D plug-in baseline, as confirmed by a paired Wilcoxon test.

KEYWORDS
3D volumetric data, data augmentation, fashion asset segmentation, imbalanced learning, synthetic data

1 | INTRODUCTION

Three-dimensional (3D) asset segmentation plays a crucial role in VR/AR services and digital twin fashion workflows in which an accurate semantic understanding of garments is essential for rendering, interaction, and content creation. However, the construction of high-quality 3D fashion datasets remains challenging owing to the cost of acquisition and annotation. Our dataset consists of (i) 502 mannequin-wearing scenes and (ii) additional single-asset captures, yielding 628 instances across 16 categories and 104 items, all recorded using an Apple iPhone 14 Pro, iOS 25.3.1(a). Despite this effort, several categories remained underrepresented, resulting in a pronounced class imbalance.

Despite recent advances in 3D semantic segmentation, progress in the fashion domain remains constrained by fundamental data limitations. In particular, 3D fashion assets with dense, part-level semantic annotations are scarce because accurate labeling requires expert knowledge and does not easily scale to deformable garments.

Moreover, existing large-scale 3D benchmarks—primarily designed for indoor scenes or autonomous driving—exhibit a considerable domain mismatch with fashion assets. Compared with rigid objects in these benchmarks, fashion items are characterized by a highly

deformable geometry, fine-grained surface details, and diverse color and material distributions that challenge direct knowledge transfer.

More importantly, long-tailed class distributions and structured occlusions induced by real garment layering are underexplored in existing datasets. These types of occlusions follow systematic and class-dependent patterns (for example, outerwear partially covering inner garments or hems overlapping footwear), making them fundamentally different from synthetic or random occlusion settings commonly adopted in previous studies. These characteristics motivate the need for domain-specific data-augmentation strategies that explicitly account for both geometric and appearance variations in 3D fashion segmentation. Consequently, existing benchmarks fail to capture adequately the long-tailed class distribution and structured occlusions inherent to real-world fashion layering.

To address these challenges, we propose three complementary augmentation strategies: (i) *SMOTE3D*: geometry- and photometry-aware oversampling method for minority classes, (ii) scene reconfiguration to increase contextual diversity, and (iii) class-consistent red–green–blue (RGB) shifts used to model appearance variations. We evaluated our approach using OctFormer, OA-CNNs, and PTv3 and compared it with Mix3D [1] as a plug-in baseline. Unlike Mix3D's global scene mixing, which primarily enhances contextual diversity, our approach targets intraclass variability at the point level by coupling geometry- and color-aware synthesis.

2 | RELATED WORK

2.1 | Data augmentation for 3D data

This study assumes 3D data representation is the optimal phenotype for volumetric assets. Real-world 3D segmentation applications are highly diverse—spanning from sparse autonomous driving scenes to dense fashion items—each presenting unique noise and domain-specific challenges. Despite these variations, class imbalance remains a prominent and ubiquitous challenge, consistently highlighted in recent studies as a primary factor degrading 3D object recognition accuracy.

One early approach for addressing class imbalance was the synthetic minority oversampling technique (SMOTE) [2]. SMOTE oversamples underrepresented classes to balance the dataset. However, this simple approach is not always suitable for 3D data. As an alternative, Griffiths and Boehm [3] proposed a weighted point-cloud augmentation technique. This method applies more substantial augmentation to underrepresented classes, resulting in 19% and 25% improvements in F_1 scores on the ScanNet and Semantic3D datasets, respectively.

Data generation using generative AI has also been proposed based on the development of generative AI such as the diffusion model. Yue and others [4] proposed a diffusion-based, scene-level synthesis method for building interior spaces.

Additionally, in real-world data, such as 3D LiDAR from autonomous vehicles, a class imbalance problem arises; large objects, including roads and buildings, dominate the data, whereas smaller objects, such as pedestrians and bicycles, are underrepresented. To address this issue, methods such as class-balanced grouping and dynamic weight averaging have been introduced to reduce effectively class imbalance and improve the recognition of rare objects [5–7]. A method for accessing feature-level generation for class imbalances was also proposed [8, 9].

Finally, the subspace prototype guidance (SPG) technique uses an auxiliary network to extract more accurately the features of minority classes by distinguishing different classes. This method considerably improved the segmentation performance of minority classes in indoor scenes, providing a practical approach for addressing class imbalance [10].

Class imbalance is a major challenge in fashion-item augmentation. For example, specific objects, such as shoes, may appear frequently in a dataset, whereas accessories or certain clothing items are less common. Although data augmentation can be achieved using expensive manual simulations or generative AI, these approaches require considerable time and resources. To address this challenge, approaches such as SMOTE—which improves the representation of minority classes—and penalized models—which impose higher penalties for insufficient learning in these classes—can also be employed [11]. These methods offer practical and cost-effective alternatives, particularly in multilabel contexts, such as fashion-item augmentation.

Typically, class imbalance remains a persistent challenge in various 3D data segmentation domains. Although several studies have attempted to mitigate this issue, this study proposes a novel data-augmentation technique specifically designed for wearable items, such as clothing and accessories, thereby setting it apart from existing approaches.

2.2 | Fashion data augmentation

Data augmentation has played a central role in improving the generalization and robustness of deep-learning

models, particularly in visual recognition. Several augmentation techniques have been proposed for 2D and 3D data modalities.

For 2D fashion image datasets, traditional augmentation schemes, such as flipping, cropping, rotation, and noise injection, have been extensively adopted to expand training distributions and reduce overfitting. For example, traditional augmentations have served as a benchmark for evaluating these types of augmentation pipelines in classification tasks, where even simple affine transformations have proven effective in improving accuracy [12]. Other studies extended these ideas by incorporating color jittering or background randomization, particularly in fine-grained fashion recognition or detection tasks [13].

Research on 3D fashion modeling is still in its early stages of development. The DeepFashion3D dataset [14] introduced fashion item-level 3D reconstructions but did not emphasize augmentation strategies. General-purpose 3D augmentation methods such as Mix3D [1] and VirtFusion [15] offer data fusion and synthetic scene generation. However, these approaches are typically designed for rigid indoor environments and do not consider deformable appearance-sensitive fashion items.

In contrast, our approach introduces domain-specific augmentation strategies that simulate realistic deformations and appearance changes, including color manipulation and structural variation. These operations are particularly relevant for fashion-centric 3D segmentation, in which the geometry and visual features of an object are highly entangled. To the best of our knowledge, no previously published study has jointly designed photometric and geometric perturbations specifically tailored to deformable 3D fashion assets for dense semantic segmentation.

2.3 | Point-cloud semantic segmentation

Recent advances in 3D semantic segmentation led to the development of OctFormer [16] that addresses the inefficiencies of traditional sparse-voxel-based CNNs and point-cloud transformers characterized by high-computational complexity and unbalanced window partitioning. By introducing an octree-based attention mechanism, OctFormer efficiently processed point clouds and achieved superior performance on large-scale datasets, including ScanNet and ScanNet200. In particular, the octree-based attention mechanism mitigates uneven window partitioning, resulting in significantly increased processing speeds by 24.

Omniadaptive sparse CNNs (OA-CNNs) [17] have emerged as efficient alternatives to transformer-based models by incorporating a lightweight module that enables spatial adaptivity without relying on attention mechanisms. OA-CNNs reduce the performance gap with point transformers, while maintaining minimal computational overhead. They introduced adaptive receptive fields and adaptive relation convolution (ARConv), enabling the network to adapt dynamically to the scene features. This yields competitive accuracy for ScanNet v2 while maintaining minimal overhead.

Wu and others [18] proposed Point Transformer V3 (PTv3), a state-of-the-art method for point-cloud processing

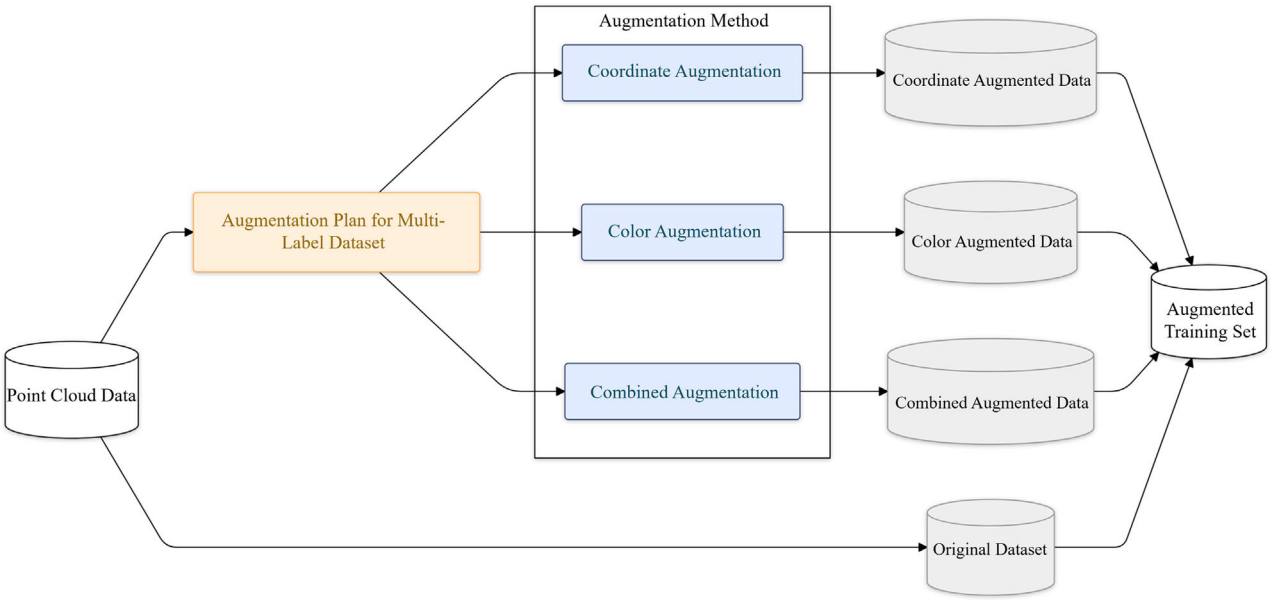

FIGURE 1 Overview of the proposed 3D augmentation pipeline (class-balanced scheduling, SMOTE3D, scene reconfiguration, and class-consistent color shifts).

that emphasizes scalability and efficiency. Unlike previous models that have focused on intricate attention mechanisms, PTv3 simplifies the analysis process to achieve a balance between accuracy and computational performance. By introducing point-cloud serialization, which converts unstructured point clouds into structured sequences via space-filling curves, we can improve data processing efficiency. In addition, PTv3 introduces serialized attention, which groups patches of serialized points to facilitate efficient attention computation, thereby improving both speed and memory consumption. PTv3 demonstrates superior performance across a range of downstream tasks, including 3D segmentation and object detection, while expanding the receptive field from 16 to 1024 points and reducing the inference latency.

We used these three models to demonstrate the practical effectiveness of the proposed data-augmentation technique for diverse 3D segmentation tasks. Unlike the standard SMOTE applied to tabular features, we propose SMOTE3D that applies interpolation in the spatial *xyz* and photometric RGB spaces, accounting for geometry-consistent sampling.

3 | FASHION 3D DATA AUGMENTATION

In this study, we propose a data-augmentation strategy to reduce class imbalance in 3D multiclass datasets within the fashion domain. Scene-level data collection often yields multiple objects per scene, resulting in multiclass segmentation challenges. To address this issue, we employed a data-augmentation approach to balance the distribution of individual classes across the dataset. Furthermore, we leveraged SMOTE3D to enhance the data diversity by adjusting the distribution of point clouds in the 3D dataset and proposed a consistent color augmentation technique. The proposed methodology is illustrated in Figure 1.

3.1 | Data-augmentation strategy for multiclass datasets

Data collected at the scene level typically contain multiple objects in a single scene. Consequently, scene-level segmentation involves solving multiclass segmentation problems. In this study, we focused on segmenting fashion items, including clothing and accessories. To address this challenge, we applied data augmentation to a multiclass dataset to ensure balanced class distributions and mitigate class imbalance.

Algorithm 1. Class-balanced augmentation for multiclass scenes.

Require: Dataset $\mathcal{D} = \{(x_d, Y_d)\}_{d=1}^{N}$, where $Y_d \subseteq S$ (multilabel scene), class set S, variance σ, lower/upper bounds (B_L, B_U) policy, augmentation ops $\mathcal{A}$ (SMOTE3D, scene-reconfig, color-shift), max iters K

Ensure: Balanced dataset $\mathcal{D}'$ with class level counts within $[B_L, B_U]$

Init counts: $c(s) \leftarrow |\{d : s \in Y_d\}|$ for all $s \in S$
Target level: $T \leftarrow \max_{s \in S} c(s)$ ▷ or a robust target, for example, percentile
Bounds: $B_L \leftarrow T - \sigma$, $B_U \leftarrow T + \sigma$
Deficit: $\alpha(s) \leftarrow \max(0, B_L - c(s))$ for all s
for $t = 1$ to K
  if $\forall s \in S:\ B_L \le c(s) \le B_U$ **then**
    break ▷ converged
  end if
  Pick minority class: $s^\star \leftarrow \arg\max_s \alpha(s)$ ▷ ties: pick higher loss/lower F_1 class if available
  Choose source: select scene or single-asset $(x_{\text{src}}, Y_{\text{src}})$ with $s^\star \in Y_{\text{src}}$
  Choose op: pick $A \in \mathcal{A}$ by rule:
    - if $s^\star$ is *rare* & *small*: prefer *SMOTE3D* on class mask of $s^\star$
    - if context helps (co-occurrence needed): prefer *scene-reconfig* (swap/compose compatible items)
    - if appearance diversity needed: apply *class-consistent color-shift*

  Synthesize: $(\tilde{x}, \tilde{Y}) \leftarrow A(x_{\text{src}}, Y_{\text{src}}, s^\star)$
  Admit check: ▷ avoid overshoot of other labels in a multilabel scene
  if $\exists s \in \tilde{Y}$ s.t. $c(s) \ge B_U$ **then**
  optionally **prune** $\tilde{x}$ regions not in $s^\star$ or **resample** another source/op
    continue
  end if
  Insert: $\mathcal{D}' \leftarrow \mathcal{D}' \cup \{(\tilde{x}, \tilde{Y})\}$
  Update counts: $c(s) \leftarrow c(s) + \mathbb{1}[s \in \tilde{Y}]$ for all s; $\alpha(s) \leftarrow \max(0, B_L - c(s))$
end for
Return $\mathcal{D}'$ ▷ now $c(s) \in [B_L, B_U]$ for most classes

We propose the data-augmentation strategy outlined in Algorithm 1 that mitigates class imbalance in multilabel datasets by systematically oversampling underrepresented categories to achieve a more uniform distribution. The algorithm first defines the category set S and then computes the current frequency of each class. A target level T is then determined as the maximum class frequency, and a tolerance interval $[B_L, B_U]$ is specified using a predefined variance σ.

In the adjustment phase, the algorithm calculates the deficit $\alpha(s)$ for each class, denoting the additional number of samples required for the class to reach the lower bound B_L. During the refinement phase, synthetic samples were iteratively generated and added to classes with nonzero deficits until all class frequencies converged within the interval $[B_L, B_U]$. This process ensures a balanced class distribution in the final dataset, thereby reducing the adverse effects of imbalances on model training and evaluation.

Algorithm 2. SMOTE3D (coordinate + color with normal-aware option).

Require: minority-class points $\{(x_i, c_i)\}_{i=1}^{N}, k$, samples per seed m, perturbation scale ϵ

1: **for** each seed point p in minority class **do**
2: Find k-NN $\mathcal{N}_k(p)$ in 3D (xyz); estimate normal n_p (Principal Component Analysis)
3: **for** $j = 1$ to m**do**
4: Sample neighbor $q \in \mathcal{N}_k(p)$
5: $\lambda \sim \mathcal{U}(0,1)$; $x' \leftarrow (1-\lambda)x_p + \lambda x_q$ ▷ normal-aware tangential jitter (optional)
6: Sample $\delta \perp n_p, \|\delta\|_\infty \le \epsilon$; $x' \leftarrow x' + \delta$ ▷ photometric interpolation + bounded jitter
7: $c' \leftarrow \mathrm{clip}\left((1-\lambda)c_p + \lambda c_q + \mathcal{U}(-70,70), 0, 255\right)$
8: **emit** synthetic point $(x', c', \mathrm{classID})$
9: **end for**
10: **end for**

Algorithm 2 outlines *SMOTE3D*, which extends the classical SMOTE to point clouds by interpolating both the coordinates (xyz) and colors (RGB) within each minority class with an optional normal-aware tangential jitter. For each seed point in a minority class, we interpolated both the geometry and appearance by sampling a random neighbor and blending their coordinates (x_p, x_q) and colors (c_p, c_q) with a mixing weight $\lambda \sim \mathcal{U}(0,1)$. To better preserve the local surface structure, we optionally added a *normal-aware tangential jitter*, where a small perturbation δ was applied to the tangent plane of the estimated normal n_p. This prevents unrealistic distortions that could occur if perturbations are added along the surface normal.

For appearance, photometric interpolation was combined with a bounded random jitter to produce class-consistent but visually diverse RGB values. Importantly, interpolation and perturbation were conducted *within* each class to avoid cross-class contamination; neighbors were always drawn from the same class, and the synthetic points inherited the class label of the seed.

Two ablation variants were also considered for the analysis: (i) *coord-only*, where only geometric interpolation and jitter were applied, and (ii) *color-only*, where only photometric interpolation and jitter were applied. These ablations help isolate the relative contributions of the geometric and appearance-based augmentations. SMOTE3D generates realistic yet diverse samples for minority classes, thereby improving the balance and robustness of downstream training.

4 | EXPERIMENTS

We conducted experiments to evaluate the impact of different data-augmentation techniques on model performance for a 3D volumetric dataset in the digital asset domain. The models were trained on an ADA 6000 graphics processing unit (GPU), and all the experiments were executed for 100 epochs. We used three models—OctFormer, OA-CNNs, and PTv3—to compare the effects of various augmentation strategies.

4.1 | Dataset

We captured 2D videos of mannequins wearing three to five items per scene. Data were collected by generating 3D scenes from 2D videos. The data acquisition pipeline is described in Appendix S1. The objects were divided into five categories: hats, tops, bottoms, shoes, and bags. The counts for each object in the 16 classes are shown in Figure 2. These products were selected for their versatility and everyday use, with a focus on high-selling items in the fashion market. As a result of the data collection, 628 video datasets were generated using an iPhone 14 Pro, with mannequin-wearing combinations comprising three to five items. The 2D video datasets were converted into 3D data and stored as point clouds. A

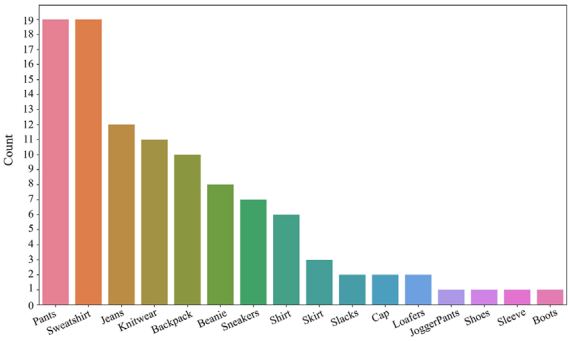

FIGURE 2 Number of unique objects per class in descending order of frequency.

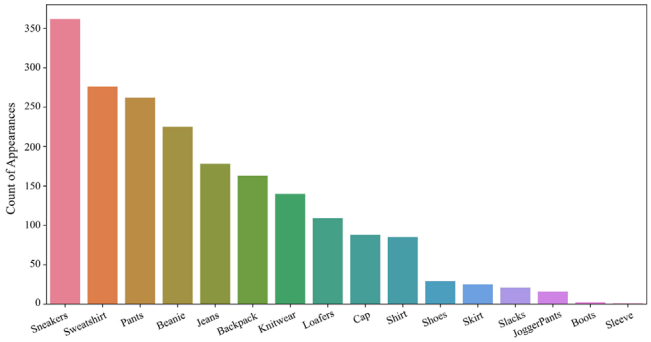

FIGURE 4 Counts of class appearances in a multiclass scene.

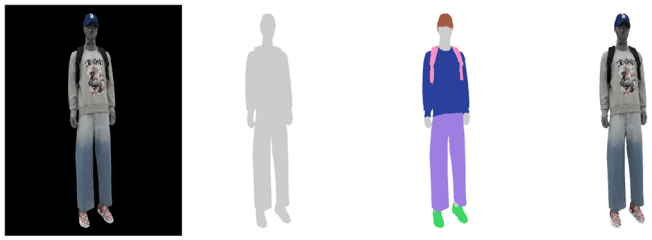

FIGURE 3 Collected data examples. The scale bar in the images represents 10 cm.

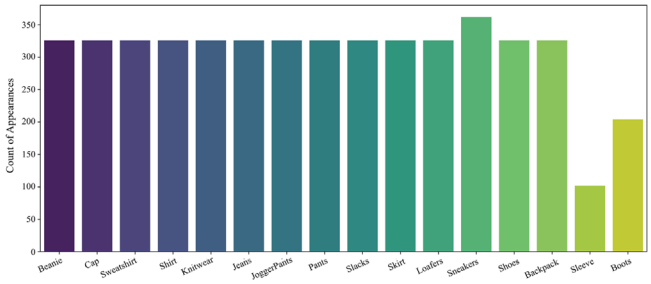

FIGURE 5 Number of assets after the application of the proposed data-augmentation method.

visualization example of the 628 datasets is shown in Figure 3, and the class distribution for each 3D scene is also depicted in Figure 4.

The collected data constituted an imbalanced dataset, with each scene containing three to five objects, posing a multiclass problem. This setting motivated class-balanced augmentation. To set a baseline, we trained each model on a minimally augmented dataset, where no significant augmentations were applied beyond basic preprocessing. The baseline performance served as a reference point for evaluating the effectiveness of different augmentation techniques.

Upon applying the proposed 3D multilabel dataset augmentation plan, the distribution of assets across scenes became more balanced, as shown in Figure 5. For assets such as boots and jogger pants, which exist in only one instance each, data augmentation was limited and thus did not achieve the same level of augmentation as those in other assets.

Ethical considerations during data collection ensured compliance with privacy regulations. Specifically, fashion items in the dataset did not feature visible trademarks, and all items were purchased individually. Additionally, individuals who acquired 2D data and participated in the tagging process were compensated in accordance with the appropriate wage standards in Korea.

4.2 | Evaluation metrics

To evaluate the semantic segmentation performance on the point-cloud dataset, we employed standard evaluation metrics for 3D object detection and segmentation, including accuracy and intersection over union (IoU). These metrics are defined as follows.

- *Accuracy*: The ratio of correctly predicted instances (true positives and negatives) to the total number of instances. This ratio quantifies the overall correctness of the model's predictions [19].
- *IoU*: The ratio of the overlapping area between the predicted segmentation and the ground truth to their union area. This metric is extensively used for evaluating the accuracy of segmentation tasks, particularly in 3D point-cloud data, as it quantifies how well the predicted object boundaries match the actual objects [20].
- *Average precision (AP)*: The precision–recall trade-off that calculates the area under the precision–recall curve. This marker is particularly effective in evaluating segmentation models, especially when class imbalance is present, by providing a robust measure of model performance across all recall levels. AP evaluates how well the predicted boundaries of an object match the ground truth [21].

These metrics provide a balanced evaluation of both classification and segmentation performances across all the object classes in a scene. The use of IoU and AP is particularly effective in 3D object detection and segmentation tasks, ensuring a robust evaluation of spatial precision and the model's ability to localize and classify objects within point-cloud data [22–24].

4.3 | Experiment

4.3.1 | Training details

Each model (OA-CNNs, PTv3, and OctFormer) was trained for 100 epochs using an ADA 6000 GPU. We compared four different augmentation techniques to assess their impact on the model performance:

- *Weighted dataset (baseline)*: This setting serves as our baseline and is trained on the original dataset using conventional data-augmentation techniques commonly adopted in vision research [14]. These augmentations increase training diversity and improve robustness to geometric variations. Specifically,
  - *Flip: 3D objects are horizontally flipped to encourage invariance to mirrored configurations.*
  - *Rotation: Random rotations are applied to expose the model to diverse object orientations, which is essential for 3D recognition.*
  - *Scaling: Objects are randomly scaled to simulate size variations and enhance robustness to changes in object dimensions.*

  In addition to geometric augmentations, class-balanced weighting is applied during training to mitigate class imbalance by assigning higher importance to underrepresented categories.
- *Color augmentation*: The dataset applies consistent color shifts across all points within the same class to increase visual diversity. Specifically, color augmentation involves uniformly shifting the RGB values of all the points within a given class. This method introduces realistic variability while maintaining the overall color scheme for each class, allowing the model to handle better variations in lighting conditions and different shades of the same color. For example, accessories such as bags and shoes may appear in slightly different colors in real-world environments, and this augmentation effectively emulates these types of variations.
- *Coordinate augmentation*: Coordinate augmentation was applied by randomly shifting each point's x, y, and z coordinates within a small range, typically defined by a uniform distribution. This approach emulates slight variations in positioning or object structures, which can occur because of factors, such as imperfect camera alignment or the natural movement of fashion items. Introducing these slight coordinate perturbations makes the model more robust to positional changes, thereby enhancing its ability to segment accurately objects from real-world scenes.
- *Combined augmentation*: Both color and coordinate augmentations were applied simultaneously to introduce variability in both dimensions, namely, visual appearance and spatial location. This combined augmentation ensures that the model learns to generalize across different instances in which both the color and positioning may vary. For example, a handbag can be placed slightly differently and have a different shade, and these types of combined augmentations help the model to learn effectively these variations. This approach has proven particularly useful for handling multiclass datasets where item colors and positions may differ considerably.

4.3.2 | Results and discussion

The experimental results presented in Tables 1 and 2 highlight the major influences of the augmentation strategy and model architecture on the segmentation performance of 3D fashion data. The results indicate that multiple augmentations achieved optimal segmentation performance.

As shown in Table 1, the simultaneous application of color and coordinate augmentation achieved the highest AP across different fashion categories. Improved performance was observed in the sweatshirt and shirt classes when both augmentations were applied. Classes that initially exhibited low performance, including loafers and shoes, also benefited substantially from the combined approach, with their mAP values improving from 0.72 to 0.90 and from 0.75 to 0.91, respectively.

Color augmentation proved particularly effective for categories where visual appearance was the dominant discriminative feature, such as in knitwear and caps. Conversely, coordinate augmentation contributed to performance gains in categories where spatial deformation played a critical role, such as jeans and pants. However, the integration of both augmentation methods consistently produced the most balanced and robust results across the diverse item types. These findings demonstrate that identifying both photometric and geometric variations is vital for constructing generalizable representations of 3D semantic segmentation tasks.

Table 2 presents a comparative analysis of the three segmentation models under different augmentation conditions. Among the evaluated models, the convolution-based OA-CNNs achieved the highest mAP score (0.96

TABLE 1 Validation of different data-augmentation methods for various fashion classes.

	Weighted dataset[a]			Color aug.[b]			Coord. aug.[c]			Color + Coord. aug.[d]		
Class	**mIoU**	**Acc.**	**mAP**	**mIoU**	**Acc.**	**mAP**	**mIoU**	**Acc.**	**mAP**	**mIoU**	**Acc.**	**mAP**
Backpack	0.74	0.86	**0.95**	0.82	0.88	0.94	0.83	0.88	0.94	**0.92**	**0.90**	**0.95**
Beanie	0.67	0.92	0.92	0.96	0.97	0.89	**0.98**	0.95	0.89	**0.98**	0.94	**0.95**
Boots	0.00	0.00	**0.97**	0.97	**0.99**	0.95	**0.98**	0.96	0.96	0.88	0.91	0.95
Cap	0.86	1.00	**0.98**	0.97	**0.99**	0.94	**0.98**	0.96	0.94	0.96	0.95	0.96
Jeans	0.56	0.94	0.96	0.95	**0.98**	0.96	**0.96**	0.91	0.96	**0.98**	0.93	**0.98**
Sweatshirt	0.34	0.38	0.96	**0.94**	**0.97**	0.94	0.83	0.85	0.94	0.90	0.88	**0.98**
Shirt	0.70	0.71	0.96	**0.95**	**0.98**	0.96	0.86	0.88	0.96	0.91	0.90	**0.97**
Knitwear	0.40	0.89	0.92	**0.91**	**0.96**	0.95	0.79	0.83	0.95	0.88	0.87	**0.97**
Jogger Pants	0.34	1.00	**0.99**	**0.98**	**0.99**	0.98	0.85	0.86	0.98	0.92	0.89	**0.99**
Pants	0.42	0.43	**0.99**	**0.96**	**0.97**	0.97	0.87	0.87	0.97	0.93	0.91	**0.99**
Slacks	0.92	1.00	0.99	**0.98**	**0.99**	0.98	0.88	0.88	0.98	0.94	0.92	**0.99**
Skirt	0.70	1.00	**0.98**	**0.95**	**0.97**	0.97	0.89	0.89	0.97	0.95	0.93	**0.98**
Loafers	0.56	0.80	0.72	0.73	0.82	0.56	0.90	0.90	0.56	**0.96**	**0.94**	**0.90**
Sneakers	0.84	0.93	0.92	0.91	0.95	0.88	0.91	0.91	0.88	**0.97**	**0.95**	**0.95**
Shoes	0.57	**0.99**	0.75	0.87	0.97	0.79	0.92	0.92	0.78	**0.98**	0.96	**0.91**
Sleeve	0.00	0.00	**0.98**	0.97	0.98	0.98	0.93	0.93	0.98	**0.99**	**0.97**	**0.98**

Note: The table shows the results for different augmentation methods (weighted dataset, color augmentation, coordinate augmentation, and combined augmentation) for each object class using OA-CNNs [17]. Bold values indicate the best performance for each metric within the corresponding comparison group. Tied best values are also shown in bold.

Abbreviations: Acc., accuracy; aug., augmentation; Coord., coordinate; mAP, mean average precision; mIoU, mean intersection over union; OA-CNN, omniadaptive sparse convolutional neural network.

[a]The weighted dataset baseline applies class-balanced sampling during training without geometric or color perturbations.

[b]Augmentation strategies applied to the original training set *without* any class reweighting or resampling.

[c]Augmentation strategies applied to the original training set *without* any class reweighting or resampling.

[d]The final augmented training set, where both augmentation strategies are jointly applied without additional weighting.

TABLE 2 Validation of different data-augmentation methods across multiple 3D semantic segmentation models.

	OctFormer [16]			OA-CNNs [17]			PTv3 [18]		
Aug. method	**mIoU**	**Accuracy**	**mAP**	**mIoU**	**Accuracy**	**mAP**	**mIoU**	**Accuracy**	**mAP**
Weighted dataset	0.45	0.65	0.87	0.54	0.74	0.93	**0.86**	**0.87**	0.90
Color aug.	0.70	0.85	0.90	**0.93**	**0.96**	0.91	0.91	0.95	0.97
Coord. aug.	0.67	0.74	0.72	**0.90**	**0.94**	0.91	0.89	0.92	0.93
Color + coord. aug.	0.69	0.86	0.89	**0.94**	0.92	0.96	0.92	**0.95**	0.98

Note: We evaluated the effectiveness of various augmentation techniques (weighted dataset, color augmentation, coordinate augmentation, and combined augmentation) using three models, namely, OctFormer [16], OA-CNNs [17], and PTv3 [18]. This evaluation identifies the most effective augmentation strategy to improve model performance across the validation and test sets. Bold values indicate the best performance for each metric within the corresponding comparison group. Tied best values are also shown in bold.

when trained using the combined augmentation strategy). This superior performance is attributed to the ability of the model to exploit the spatial regularity and local consistency inherent in the dataset, which was constructed from mannequins in stable, uniform poses.

PTv3, which leverages serialized attention mechanisms, exhibited a strong baseline performance with the use of the Weighted dataset method and reached an mAP value of 0.98 with the use of the combined augmentation method. This result suggests that PTv3 possesses a strong generalization capability, particularly in scenarios with limited or no augmentation. However, its performance in categories with complex local structures is slightly lower than that of OA-CNNs, indicating that fine-grained local

detail extraction remains a challenge for transformer-based models.

OctFormer, which utilizes a hierarchical octree-based attention mechanism, consistently exhibited the lowest performance across all metrics. This result highlighted its limited suitability for dense and highly structured datasets, particularly for those with limited pose variability or occlusion diversity.

While Mix3D [1] enhances scene-level contextual generalization by mixing complete 3D environments, its augmentation principle primarily targets the global context rather than the intraclass diversity. By contrast, our proposed SMOTE3D-based approach directly synthesizes minority-class samples within both geometric (*xyz*) and photometric (RGB) spaces, explicitly addressing class imbalance. Unlike Mix3D's global scene mixing, which primarily benefits from rigid indoor environments, our approach operates at the point level. This distinction is important in the fashion domain, where local color-shape consistency and deformable structures dominate segmentation accuracy. Consequently, our augmentations yielded stronger improvements in the class-level F_1 and mAP outcomes, especially for underrepresented items (for example, loafers, and shoes), where Mix3D provided limited gains.

These results suggested that the dataset's structural characteristics strongly influenced the effectiveness of the segmentation model. Models with inductive biases that align with the data's spatial properties tend to outperform more general-purpose architectures, especially when data variability is constrained.

5 | CONCLUSION

This study demonstrated that geometry- and color-aware synthesis effectively mitigates intraclass imbalance in deformable fashion assets, providing a practical augmentation strategy for 3D fashion segmentation. Beyond benchmark evaluations, the proposed method has been integrated into a production-level avatar fashion system in ZEPETO, where it exhibits stable behavior in conjunction with the use of real-world data distributions. The proposed SMOTE3D framework, along with full training scripts, will be publicly released to facilitate reproducibility and encourage further research to promote 3D fashion understanding.

5.1 | Limitations and future work

Despite its effectiveness, this study has several limitations. First, quantitative evaluations were restricted to a single dataset due to the lack of public benchmarks supporting dense 3D semantic segmentation with authentic, structured garment occlusions (unlike synthetic settings in previous studies [2]). Second, the dataset is limited to stable mannequin poses and lacks complex materials (for example, highly reflective or transparent fabrics), restricting deformation diversity. Finally, while integration into the ZEPETO avatar system confirms compatibility with commercial pipelines, it currently excludes physics-based dynamic cloth simulations. Future work will extend SMOTE3D to incorporate articulated human motions, diverse material properties, and physically grounded garment interactions, alongside cross-domain validation across external datasets and avatar platforms.

ACKNOWLEDGMENTS

This work was supported by an Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (RS-2023-00215700, Trustworthy Metaverse: Blockchain-enabled Convergence Research).

CONFLICT OF INTEREST STATEMENT

The authors declare that there are no conflicts of interest.

DATA AVAILABILITY STATEMENT

Upon publication, we will release the complete SMOTE3D implementation framework, including the data preprocessing utilities, training and evaluation scripts, and configuration files at https://github.com/kusses/3D_Augmentation. The repository enables the full reproduction of all reported results, including the Mix3D plug-in baseline and significance tests.

ORCID

Jiyoun Lim https://orcid.org/0000-0003-4246-3081
Jeong-Woo Son https://orcid.org/0000-0003-2592-3875

REFERENCES

1. A. Nekrasov, L. Schneider, Y. Liao, and A. Geiger, *Mix3D: out-of-context data augmentation for 3D scenes*, arXiv preprint (2021). DOI 10.48550/arXiv:2110.02210
2. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, *SMOTE: synthetic minority over-sampling technique*, J. Artif. Intell. Res. **16** (2002), 321–357.
3. D. Griffiths and J. Boehm, *Weighted point cloud augmentation for neural network training data class-imbalance*, arXiv preprint (2019). DOI 10.48550/arXiv:1904.04094
4. H. Yue, Q. Wang, L. Huang, and M. Zhang, *Enhancing point cloud semantic segmentation of building interiors through diffusion-based scene-level synthesis*, Autom. Constr. **178** (2025), 106390.
5. M. Li, S. Lin, Z. Wang, Y. Shen, B. Zhang, and L. Ma, *Class-imbalanced semi-supervised learning for large-scale point cloud*

semantic segmentation via decoupling optimization, Pattern Recognit. **156** (2024), 110701.
6. F. Xu, Y. Xiao, J. Mei, Y. Hu, Q. Fu, and H. Shi, *Domain-adaptive point cloud semantic segmentation via knowledge-augmented deep learning*, Pattern Recognit. **172** (2025), 112528.
7. Y. Zhao and Z. Wang, *Class-balanced grouping and sampling for point cloud 3D object detection*, arXiv preprint (2021). DOI 10.48550/arXiv:2111.03516
8. S. Baek and J. T. Lee, *High-speed and precise virtual try-on with two-stage semantic segmentation and a latent consistency model for optimized diffusion processes*, ETRI J. **47** (2025), no. 5, 881–892.
9. J. Han, K. Liu, W. Li, F. Zhang, and X.-G. Xia, *Generating inverse feature space for class imbalance in point cloud semantic segmentation*, IEEE Trans. Pattern Anal. Mach. Intell. **47** (2025), 5778–5793.
10. J. Zhou and Q. Chen, *Subspace prototype guidance for mitigating class imbalance in point cloud semantic segmentation*, arXiv preprint (2023). DOI 10.48550/arXiv:2408.10537
11. M. T. Keane, E. M. Kenny, M. Temraz, D. Greene, and B. Smyth, *Twin systems for DeepCBR: a menagerie of deep learning and case-based reasoning pairings for explanation and data augmentation*, arXiv preprint (2021). DOI 10.48550/arXiv:2104.14461
12. H. Xiao, K. Rasul, and R. Vollgraf, *Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms*, arXiv preprint (2017). DOI 10.48550/arXiv:1708.07747
13. R. Velioglu, R. Chan, and B. Hammer, *FashionFail: addressing failure cases in fashion object detection and segmentation* (2024 International Joint Conference on Neural Networks (IJCNN), Yokohama, Japan), 2024, pp. 1–8.
14. Y. Chen, R. Xie, S. Yang, L. Dai, H. Sun, Y. Huo, and R. Li, *Single-View 3D garment reconstruction using neural volumetric rendering*, IEEE Access **12** (2024), 49682–49693.
15. W. S. Harrison III and F. Proctor, *Virtual Fusion: state of the art in component simulation/emulation for manufacturing*, Procedia Manuf. **1** (2015), 110–121.
16. P.-S. Wang, *OctFormer: octree-based transformers for 3D point clouds*, ACM Trans. Graph. (TOG) **42** (2023), no. 4, 1–11.
17. B. Peng, X. Wu, L. Jiang, Y. Chen, H. Zhao, Z. Tian, and J. Jia, *OA-CNNs: omni-adaptive sparse CNNs for 3D semantic segmentation* (Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA), 2024, pp. 21305–21315.
18. X. Wu, L. Jiang, P.-S. Wang, Z. Liu, X. Liu, Y. Qiao, W. Ouyang, T. He, and H. Zhao, *Point Transformer V3: simpler, faster, stronger* (Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA), 2024, pp. 4840–4851.
19. H. Shim, C. Kim, and E. Yang, *CloudFixer: test-time adaptation for 3D point clouds via diffusion-guided geometric transformation* (European Conference on Computer Vision, Milan, Italy), 2024, pp. 454–471.
20. C. Shi, Y. Zhang, B. Yang, J. Tang, Y. Ma, and S. Yang, *Part2Object: hierarchical unsupervised 3D instance segmentation* (European Conference on Computer Vision, Milan, Italy), 2024, pp. 1–18.
21. D. Rozenberszki, O. Litany, and A. Dai, *UnScene3D: unsupervised 3D instance segmentation for indoor scenes* (Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA), 2024, pp. 19957–19967.
22. H. Sheng, S. Cai, N. Zhao, B. Deng, J. Huang, X.-S. Hua, M.-J. Zhao, and G. H. Lee, *Rethinking IoU-based optimization for single-stage 3D object detection* (European Conference on Computer Vision, Tel Aviv, Israel), 2022, pp. 544–561.
23. C.-J. Tsai, H.-T. Chen, and T.-L. Liu, *Pseudo-embedding for generalized few-shot 3D segmentation* (European Conference on Computer Vision, Milan, Italy), 2024, pp. 383–400.
24. W. Zheng, W. Tang, S. Chen, L. Jiang, and C.-W. Fu, *CIA-SSD: confident IoU-aware single-stage object detector from point cloud* (Proceedings of the AAAI Conference on Artificial Intelligence), 2021, pp. 3555–3562.

AUTHOR BIOGRAPHIES

Jiyoun Lim has been at the Electronics and Telecommunications Research Institute (ETRI) since 2013. She is currently a principal researcher in the Media Intellectualization Research Section of ETRI. She received her B.S., M.S., and Ph.D. degrees in Industrial Systems and Engineering from KAIST.

Jungwoo Son has been working at the ETRI since 2013. He is currently the principal researcher in Media Intellectualization Research. He received his Ph.D. in Computer Science from Kyungpook National University in 2012.

Alex Lee is a senior researcher in Media Intellectualization Research. His interests focus on the extraction of the attributes of intelligent media. He has also researched 5G traffic analysis, spatial audio modeling, lossless audio compression, deep learning for scene representation, and programmable interactive media platforms.

Sun Joong Kim is a principal researcher in the Media Intellectualization Research Section of the ETRI.

Namkyung Lee is a leader in Media Intellectualization Research and a principal researcher at the ETRI.

Wonjoo Park received B.S. and M.S. degrees in Information and Communications Engineering from Chungnam National University, Daejeon, Republic of Korea, in 1998 and 2000, respectively. She joined the ETRI in 2000, where she is currently a principal researcher.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: J. Lim, J.-W. Son, A. Lee, S. Joong Kim, N. Lee, and W. Park, *SMOTE3D: Geometry- and color-aware volumetric data augmentation for 3D fashion asset segmentation*, ETRI Journal (2026), 1–11, DOI 10.4218/etrij.2025-0458