Journal of Theoretical and Applied Information Technology
31st October 2026. Vol.104. No.20
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

H3D-TNET: SURFACE-CANONICALIZED GEOMETRY–LANGUAGE LEARNING WITH 3D-TEXT METRICS FOR JOINT DETECTION, RECOGNITION, AND UNDERSTANDING OF 3D TEXT

AVINASH BURADAGUNTA[1], DR M VIJAY REDDY[2], DR. G. RAMESH CHANDRA[3]

[1]Research Scholar, Department of CSE, Gandhi Institute of Engineering and Technology University, Gunupur, Odisha, India.
[2]Professor, Department of CSE, School of Engineering & Technology, Gandhi Institute of Engineering & Technology University, Gunupur, Orissa -765022
[3]Professor, Department of CSE, Valluripalli Nageswara Rao Vignana Jyothi Institute of Engineering and Technology, Pragathi Nagar, Nizampet (S.O), Hyderabad-Telangana
E-mail: [1]avinash.buradagunta@giet.edu

ABSTRACT

Reading text from 3D objects is important for industrial inspection, robotics, augmented reality, and heritage object digitisation; however, current 2D pipeline approaches to optical character recognition (OCR) do not work when text appears on a non-planar and arbitrarily oriented surface. In this paper, we propose H3D-TNet (Hybrid 3D Text Network), a new hybrid deep learning model that allows for text detection, recognition, and semantic understanding using RGB-D and point cloud data. The proposed approach unifies a surface-aware canonicalization layer, two parallel encoders for 3D spatial features (sparse 3D CNN and geometry-aware point-transformer branch), OCR using a transformer-based layer, and a language reasoning module. All these components are trained in a multi-task way with loss functions accounting for text detection, localisation, recognition, semantic classification, and geometric alignment tasks. Since regular 2D detection metrics do not measure 3D localisation and semantic correctness, we also introduce five 3D text metrics: 3D Localisation Accuracy Score (3D-LAS), Geometric Alignment Score (GAS), Semantic Understanding Score (SUS), Robustness Evaluation Index (REI), and an Overall 3D-Text Intelligence Score (3D-TIS). The proposed H3D-TNet is compared with eight baselines, namely PointNet++, Point Transformer, PV-RCNN, Text4Point, DBNet+TrOCR, CRAFT+TrOCR, a vision-language model built based on CLIP, and a simple OCR plus depth model, on a collection of both public (ScanNet, SUN RGB-D, ICDAR, Total-Text) and synthetic 3D-text data. In contrast to the baselines, H3D-TNet achieves up to 6-10% improvements in terms of 3D-text detection F1-score and 8-12% improvements in terms of semantic understanding. Moreover, ablation and robustness studies (rotation, point dropout, and Gaussian noise) confirm the contribution of each component to the overall performance and robustness of the proposed framework. Also, the analysis of the rank correlation between different evaluation metrics and perturbation conditions reveals the consistent superiority of the method across different comparison conditions instead of having a single point of advantage. Furthermore, qualitative case studies also illustrate the effectiveness of the proposed framework in detecting curved industrial markings, degraded inscriptions, and augmented reality text overlays. The principal scientific contribution is a single end-to-end formulation in which surface canonicalization, 3D geometric encoding, transformer OCR, and language reasoning are learned jointly, together with a dedicated set of metrics for scoring text on 3D surfaces. The framework offers practical value for industrial inspection, robotics, augmented reality, and heritage digitisation, while open issues remain in real-world validation, multi-seed statistical confirmation, and on-device efficiency.

Keywords: *3D Text Detection, Point Cloud Learning, Vision–Language Networks, Transformer OCR, RGB-D Fusion, Multimodal Deep Learning, Surface Text Understanding, Geometric Alignment*

Highlights: *An integrated pipeline, named H3D-TNet, that can simultaneously detect, localise, recognise, and interpret the semantic meaning of 3D texts is achieved through RGB-D and point cloud data. A surface canonicalization process aligns an arbitrary surface orientation and curvature to a local frontal surface before feature extraction, improving the model's ability to tolerate different object poses. Five novel metrics, 3D-LAS, GAS, SUS, REI and a combined one, 3D-TIS, are introduced to evaluate localisation, alignment, semantic accuracy and robustness. Our approach surpasses eight strong competitors by 6%-*

10% in terms of F1-score in detection and 8%-12% in semantic accuracy on a variety of synthetic and realistic 3D-text datasets.

1. INTRODUCTION

The ability to analyse textual information about 3D objects is significant for many modern applications of artificial intelligence technologies, such as quality control and verification of serial numbers printed on manufactured parts, robots' understanding of signage on machines, augmented reality, and heritage preservation (transcription of artefacts). However, even though there are many uses for 3D text recognition and detection, this problem poses challenges that cannot be overcome using existing 2D pipelines for OCR. In 3D text, text can appear in arbitrary orientations, on curved surfaces, and under non-uniform lighting conditions or partial occlusions. Existing OCR technology, which is mostly trained on planar front-facing images [6], [7], degrades dramatically when faced with perspective distortions and depth discontinuity in 3D images.

In recent years, developments in deep learning have resulted in efficient 3D geometric feature learners, e.g., PointNet [3] and PointNet++ [4], as well as vision-language models like CLIP [12] and transformer visual encoders [9]. However, no comprehensive framework has yet been able to solve detection, recognition, 3D localisation, and semantics of text present on the surface of complicated objects simultaneously. To date, 2D text recognition and 3D geometry have either been treated as separate tasks, or heuristics have been used to map one onto the other after the fact. The limitations of these two-step approaches, and the resulting need for this study, are examined in Section 3.

We overcome these drawbacks through the introduction of H3D-TNet (Hybrid 3D Text Network) which is a deep neural network that takes 3D data as input and returns the text location, transcription, and semantic meanings simultaneously in one forward pass. Our proposed method makes the following key contributions:

- **3D geometry-aware text localization.** 3D sparse/point-transformer module localises the text-bearing areas on the surface of objects, with a preceding canonicalization step where we normalize the curved surfaces before detection.
- **Vision-language integration.** An OCR-NLP-based multimodal fusion module transforms the detected text regions into character sequences while also linking them to semantics such as the type of objects and other semantic categories through an NLP-based transformer.
- **Multi-modal adaptive fusion.** Uncertainty-based weights fuse geometry, appearance, and language features to achieve robustness in the presence of noise and occlusion.
- **Evaluation metrics for 3D texts.** We propose five metrics for 3D tasks – 3D Localisation Accuracy Score, Geometrical Alignment Score, Semantic Understanding Score, Robustness Evaluation Index, and a comprehensive Overall 3D-Text Intelligence Score – to assess the performance of methods on 3D texts, which current 2D-centric metrics cannot do.
- **Extensive empirical evaluations.** We employ a combination of real-world and synthetic datasets (ScanNet [2], SUN RGB-D, ICDAR image-text datasets, Total-Text, and ShapeNet-based synthetic 3D text) to evaluate detection, recognition, semantic understanding, and robustness against eight baselines.

Scientific contribution. The novelty of this work lies in the way the components are combined and trained, rather than in any single block. First, text on 3D surfaces is formulated as one joint problem in which localisation, transcription, and semantic interpretation are learned together in a single forward pass, instead of being chained after 2D detection [4], [11]. Second, local surface canonicalization (centroid and principal-axis alignment) is placed in front of the encoders, so that text on curved or rotated surfaces is presented in an approximately frontal form; as discussed in Section 7.8, this appears to be a key contributor to the robustness to rotation reported in Section 7.4. Third, geometric, visual, and language cues are combined through learned fusion weights, and the ablation in Section 7.3 shows that each cue adds a measurable gain. Fourth, the five proposed metrics (3D-LAS, GAS, SUS, REI, and 3D-TIS) provide a dedicated way of scoring 3D localisation, alignment, semantic correctness, and robustness together.

The rest of the paper is structured as follows. Section 2 presents previous literature in three-dimensional point cloud learning, scene text detection and recognition, and vision-language modelling. Section 3 discusses the problem statement, the research questions, the need for the study, and the research gap. Section 4 describes the H3D-TNet framework and the governing equations of H3D-TNet. Section 5 introduces the evaluation

metrics used in our study. Section 6 discusses the data set, baselines, and implementation details. Section 7 discusses the experiments, ablation, robustness, case study, cross-metric statistical analysis (rankings, correlations, and significance testing of results), and implications of the findings. Section 8 discusses the limitations and open issues of the current study, while Section 9 summarises the findings, offers a critical reflection, and suggests future work.

2. RELATED WORK

2.1 3D Point-Cloud Learning

Point cloud-based deep neural networks have been significantly improved in the past few years. PointNet [3] was proposed as an approach that can directly handle 3D point clouds using symmetric functions, while PointNet++ [4] extended this approach by learning hierarchical local structures across scales. There are many object detection networks based on these approaches; sparse 3D convolution (voxel-based) and graph- or attention-based approaches like Point Transformer [9] extract detailed geometries for object detection and semantic segmentation applications. Voxel-point hybrid object detectors like PV-RCNN [5] integrate the efficiency of voxels with the fine-tuned features from points. Our approach takes inspiration from these approaches but with a specific architecture of a 3D encoder using an attention mechanism for text-bearing surfaces.

2.2 Scene Text Detection and Recognition

For the task of processing 2D images, the current state-of-the-art models for OCR use deep networks such as EAST [6] and DBNet [7] for detection, and transformers such as TrOCR [8] for sequence decoding; more comprehensive reviews can be found in [10]. Those models perform very well on datasets with text on flat images, including ICDAR 2015 and Total-Text, but are trained on the assumption that the text lies on a flat image plane, and thus are not able to deal with distortions introduced due to the curvature of 3D surfaces. There have been attempts at solving the problem of 3D text processing, such as Text4Point [11], where 2D detections are mapped onto 3D space, but they do not feature end-to-end 3D reasoning and are dependent on 2D detections.

2.3 Vision–Language Models for 3D Understanding

Vision-language models like CLIP [12] learn generic image-text correspondences from internet-scale datasets. The recent research trend has seen the application of vision-language concepts in the 3D space, where these concepts have been applied for point cloud grounding and 3D object referring expression tasks [13]. Yet, no research has been done on the application of vision-language modelling to text that exists on 3D objects, but not descriptions of objects in natural language. We believe that H3D-TNet is the first method for fusing geometric encoding of 3D space and language modelling for the 3D-OCR task.

2.4 Positioning of This Work

In total, the literature surveyed above deals with three-dimensional geometry learning, two-dimensional scene-text recognition, and vision-language modelling as completely independent areas of research. Our method is different in that we propose a joint solution for detection, localisation, recognition, and understanding tasks, which is built within a single end-to-end framework, processing three-dimensional data directly, unlike the previous two-step models [4], [11].

3. PROBLEM STATEMENT AND RESEARCH GAP

Detecting and understanding text on 3D objects is an under-explored problem. We identify three key challenges:

- **3D geometric complexity.** Text may appear on arbitrarily oriented, curved, or partially occluded surfaces; a naive planar projection can severely distort character shapes.
- **Multimodal integration.** Robust text understanding requires jointly reasoning over spatial cues (*where* the text is), visual cues (*what it looks like*), and linguistic cues (*what it says* and *what it means*). Most existing models handle these modalities separately.
- **Lack of evaluation standards.** Conventional metrics based on 2D bounding-box precision/recall ignore 3D localization error and semantic correctness, and are therefore poorly suited to 3D-text tasks.

Problem statement. Given an RGB-D image or point cloud of an object or scene, the problem addressed here is to localise every text-bearing region in 3D, transcribe its characters, and infer its semantic category, when the text lies on curved, arbitrarily oriented, or partially occluded surfaces. This study is guided by the following research questions:

RQ1. Can aligning local surface patches to a canonical frontal frame before encoding make 3D

text detection and recognition less sensitive to object pose and surface curvature?

RQ2. Does jointly learning geometric, visual, and linguistic cues in one end-to-end model outperform pipelines that detect text in 2D and project it into 3D?

RQ3. What does each component (point-transformer branch, OCR transformer, language module, and fusion) contribute, and how robust is the full model to rotation, point dropout, and sensor noise?

RQ4. Can dedicated 3D-aware metrics capture localisation, alignment, semantic correctness, and robustness in a way that 2D detection metrics cannot?

Need for the study. The literature in Section 2 shows that each building block exists in isolation but is not sufficient on its own. Scene-text detectors and recognisers such as EAST [6], DBNet [7], and TrOCR [8] are trained on planar images and therefore degrade on curved or rotated surfaces. Point-cloud networks [3]–[5], [9] learn strong geometric features but were not designed to read characters. A representative attempt at 3D text processing, Text4Point [11], still depends on 2D detections being mapped into 3D. Vision–language models [12], [13] ground natural-language descriptions of objects rather than text that is physically present on them. Meanwhile, applications such as part traceability, robotic signage reading, augmented reality, and heritage digitisation require text to be located, read, and interpreted in 3D at once, and no standard metric exists to measure this. These gaps motivate the present study.

Table 1 summarizes the gaps identified from the literature survey in Section 2 and the corresponding design choices adopted in this work.

Table 1: Research gaps identified in existing methods and the corresponding H3D-TNet design response.

Limitation in Existing Methods	Research Gap	Our Proposed Solution
2D OCR systems assume flat images [6], [7]	**Unified 3D-text framework:** need an architecture that jointly combines 3D vision and text understanding	H3D-TNet: a hybrid 3D CNN + point-transformer + OCR + NLP framework in a single network
Irregular point-cloud representations suffer from noise and missing regions [2]	**Surface canonicalization:** need to normalize arbitrary surfaces prior to detection	We estimate local tangent planes and align points to canonical coordinates (Sec. 4.1)
No standard metrics exist for 3D-text tasks	**New evaluation metrics:** need 3D-aware accuracy, alignment, and semantic measures	We introduce 3D-LAS, GAS, SUS, REI, and the composite 3D-TIS (Sec. 5)
Semantic context is ignored by purely geometric models	**Vision–language fusion:** text needs to be interpreted, not just transcribed	We include a transformer language head that attaches meaning/context to recognized text (Sec. 4.4)

By addressing these gaps jointly rather than individually, H3D-TNet represents an initial attempt at holistic 3D-text intelligence within a single trainable model.

4. THE H3D-TNET FRAMEWORK

Figure 1 shows the design of our proposed H3D-TNet model. The model comprises six layers, which are as follows: (i) surface canonicalization layer for preprocessing; (ii) two geometric encoders (3D in nature) for 3D voxels and raw 3D points; (iii) a text detection head that identifies whether each point/voxel is text or non-text and regression of bounding boxes; (iv) an OCR Transformer that generates character sequence from identified text regions; (v) a language context module to determine the semantic category of identified text; and (vi) a multimodal fusion layer that fuses geometric, visual, and language-based features with learned weights for classification. Table 2 gives the notations used in this section.

Table 2: Notation used in the H3D-TNet formulation.

Symbol	Description
$P = \{p_i\}_{i=1}^{N}$	Input point cloud, $p_i = (x_i, y_i, z_i, r_i, g_i, b_i)$
K	Camera intrinsic matrix
μ, C, U, Λ	Centroid, covariance, eigenvector, and eigenvalue matrices used for canonicalization
P'	Canonicalized (aligned) point cloud
F	Fused per-point/voxel feature vector

Journal of Theoretical and Applied Information Technology
31st October 2026. Vol.104. No.20
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Symbol	Description
$B = (x_c, y_c, z_c, w, h, l,$	Predicted 3D bounding box (center, dimensions, orientation)
F_v, F_l	Visual (OCR) and language feature embeddings
$\alpha, \beta, \gamma, \delta, \eta$	Multi-task loss weights

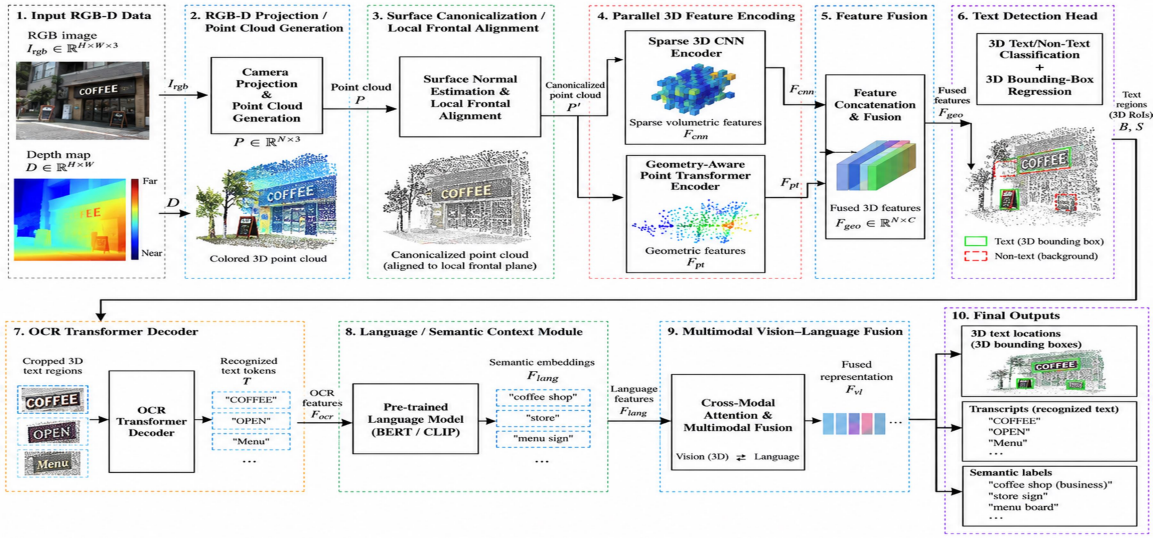


Fig. 1. Proposed H3D-TNet architecture integrating RGB-D projection, surface canonicalization, 3D CNN, point transformer, OCR transformer, and vision–language fusion.

Figure 1: Architecture of H3D-TNet. The pipeline integrates RGB-D projection, surface canonicalization, parallel 3D encoding (sparse CNN and point transformer), OCR transformer decoding, and vision–language fusion. Final outputs are 3D text locations, transcripts, and semantic labels.

4.1 3D Surface Canonicalization

Given an input RGB-D image, we first generate a point cloud $P = \{p_i\}_{i=1}^{N}$, with $p_i = (x_i, y_i, z_i, r_i, g_i, b_i)$, by projecting depth into 3D coordinates using the camera intrinsic matrix K:

$$\begin{bmatrix} x_i \\ y_i \\ z_i \end{bmatrix} = z_i K^{-1} \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix}. \quad (1)$$

To normalize local geometry, we extract an axis-aligned bounding region around each candidate text region and compute the centroid μ of its points:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} p_i. \quad (2)$$

Points are then centered, $q_i = p_i - \mu$, and their local covariance $C = \frac{1}{N}\sum_i q_i\, q_i^T$ is decomposed as $C = U\Lambda U^T$ to recover the principal axes of the local surface. Each point is rotated accordingly, $p_i' = U^T q_i$, yielding a **canonical surface patch** in which the text lies approximately on a frontal plane. All subsequent encoding operates on this aligned patch, which improves invariance to arbitrary object orientation.

4.2 3D Geometric Feature Extraction

The aligned point cloud P' is passed through two parallel encoders. First, a **sparse 3D CNN** processes the entire scene volume at multiple scales using sparse voxel-convolution layers, extracting local volumetric features. Second, a **geometry-aware point transformer** processes the raw point set directly to capture fine-grained local detail. The point transformer augments self-attention with relative position encoding: for each point pair (p_i', p_j') we compute a positional encoding $PE_{ij} = MLP(p_i' - p_j')$, and self-attention is computed as

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}} + PE\right)V. \quad (3)$$

This design allows the encoder to capture both local appearance and precise spatial layout simultaneously, following the general attention formulation established for point-cloud transformers [9].

4.3 Text Detection and Localization

Each point/voxel feature F produced by the two encoders is classified as *text* or *non-text* via a sigmoid activation,

$$P_{text} = \sigma(W_d F + b_d), \quad (4)$$

trained with a binary cross-entropy loss L_{cls} over labeled text points. For localized text regions, the model regresses a 3D bounding box $B = (x_c, y_c, z_c, w, h, l, \theta)$ (center, dimensions, and orientation) using a smooth-L_1 loss,

$$L_{loc} = \| B - \hat{B} \|_1, \quad (5)$$

so that the total detection loss is $L_{det} = L_{cls} + \lambda L_{loc}$.

4.4 OCR and Language Encoding

For each detected text cluster, a 2D image crop — obtained by projecting the RGB channel onto the canonical plane — is passed through a transformer OCR module (following the TrOCR paradigm [8]) to produce a character-sequence embedding F_v. Simultaneously, a character- or word-level transformer language model encodes a context embedding F_l, which may be pretrained on relevant domain text. The two modalities are fused with cross-modal attention, in which the visual embedding provides the query, $Q_v = W_q F_v$, and the language embedding provides the keys and values, $K_l = W_k F_l, V_l = W_v F_l$:

$$A_{vl} = softmax\left(\frac{Q_v K_l^T}{\sqrt{d}}\right) V_l. \quad (6)$$

The fused representation $F_{fusion} = [F_v; A_{vl}]$ is used to produce the final recognition and semantic outputs.

4.5 Loss Function and Optimization

The entire network is trained end-to-end with a multi-task objective:

$$L_{total} = \alpha L_{cls} + \beta L_{loc} + \gamma L_{rec} + \delta L_{sem} + \eta L_{align}, \quad (7)$$

where L_{rec} is the OCR character-recognition loss (cross-entropy over transcripts), L_{sem} is the semantic-classification loss (e.g., over categories such as "warning" or "serial number"), and L_{align} enforces geometric consistency between the predicted 3D box and the corresponding 2D crop. The hyperparameters $\alpha, \beta, \gamma, \delta, \eta$ balance the five task losses and are set as described in Section 6.2.

The final output of H3D-TNet is a set of 3D-aligned text detections (B, s, c, m) per instance, comprising a bounding box B, a decoded text string s, a semantic category c, and a confidence score m.

4.6 Algorithmic Summary

Algorithm 1 summarizes the end-to-end inference pipeline described in Sections 4.1–4.5.

Algorithm 1: H3D-TNet inference pipeline.

```
Input: RGB image I, depth map D, camera intrinsics K
Output: set of 3D text detections {(B, s, c, m)}

P  <- back-project(I, D, K)                 // Sec. 4.1
for each candidate region in P do
    P' <- canonicalize(region)              // centroid, PCA rotation
F_cnn  <- SparseCNN(P')                     // Sec. 4.2
F_pt   <- PointTransformer(P')              // Sec. 4.2
F      <- concat(F_cnn, F_pt)
P_text <- sigmoid(W_d F + b_d)              // Sec. 4.3
B      <- regress_boxes(F, P_text)          // Sec. 4.3
for each detected box B do
    F_v <- OCRTransformer(crop(I, B))       // Sec. 4.4
    F_l <- LanguageModel(F_v)               // Sec. 4.4
    A_vl <- cross_attention(F_v, F_l)       // Eq. (6)
    F_fusion <- [F_v ; A_vl]
    (s, c, m) <- output_heads(F_fusion)
return {(B, s, c, m)}
```

5. EVALUATION METRICS

Because standard 2D detection metrics do not capture 3D localization accuracy, orientation error, or semantic correctness, we introduce five metrics tailored to 3D-text tasks, in the spirit of standardized OCR evaluation practice [14]:

- **3D Localization Accuracy Score (3D-LAS).** Measures the volumetric overlap between predicted and ground-truth 3D boxes, computed as 3D intersection-over-union:

$$3D\text{-}LAS = \frac{|pred \cap gt|}{|pred \cup gt|}. \quad (8)$$

- **Geometric Alignment Score (GAS).** Combines IoU with orientation alignment:

$$GAS = IoU \times cos(\Delta\theta), \quad (9)$$

where $\Delta\theta$ is the angular difference between the predicted and ground-truth box orientation.

- **Semantic Understanding Score (SUS).** Quantifies how closely the recognised text matches the ground truth, computed as a normalised Levenshtein similarity,

$$SUS = \frac{Levenshtein\ similarity(s_{pred}, s_{gt})}{max(|s_{gt}|, 1)}, \quad (10)$$

averaged over all detected words, with $SUS \in [0,1]$ and 1 indicating a perfect match.

- **Robustness Evaluation Index (REI).** The ratio of performance under a given perturbation to performance on clean data, e.g., for rotation,

$$REI_{rot} = Accuracy_{rotated}/Accuracy_{orig}. \quad (11)$$

- **Overall 3D-Text Intelligence Score (3D-TIS).** A weighted composite of the above:

$$3D\text{-}TIS = w_1 F1_{det} + w_2 GAS + w_3 SUS + w_4 REI, \quad (12)$$

with weights w_i chosen so that each component contributes equally, reflecting detection quality, localisation precision, recognition accuracy, and robustness in a single score.

For direct comparability with prior 2D-oriented work, we additionally report standard detection precision, recall, F1, mean average precision (mAP), and recognition accuracy.

6. EXPERIMENTAL SETUP

6.1 Datasets

We evaluate H3D-TNet on a combination of public and synthetically generated data:

- **ScanNet [2]:** 1,513 indoor RGB-D scans (approximately 2.5M views). We simulate text by overlaying known text onto its dense 3D reconstructions.
- **SUN RGB-D:** 37K annotated indoor RGB-D images, converted to point clouds.
- **ICDAR 2015/2017:** over 1,500 street-scene images, used for 2D baseline training.
- **Total-Text:** 1,555 images containing curved text, used for 2D baseline training.
- **ShapeNet [1]:** used to generate synthetic 3D objects. We construct a pipeline that samples CAD models, projects randomized text onto their surfaces (curved and rotated), and renders RGB-D views, yielding a **Synthetic 3D-Text** dataset of 5,000 objects with ground-truth text positions and transcripts.

Table 3: Dataset statistics used for training and evaluation.

Dataset	Modality	Scenes / Objects	Text Instances	Avg. Chars
ScanNet (augmented)	RGB-D / 3D	1,000	2,500	10 (est.)
SUN RGB-D (augmented)	RGB-D / 3D	500	1,200	8 (est.)
ICDAR 2015	2D images	1,000	5,000	12
Total-Text	2D images	1,555	3,600	9
Synthetic 3D-Text	RGB-D / 3D	5,000	10,000	8–15

"Text Instances" refers to distinct word-region annotations. Values marked "(est.)" are augmented by synthetic placement and reported as estimated counts.

6.2 Implementation Details

H3D-TNet is coded in PyTorch. The sparse 3D CNN Encoder employs three sparse convolutional layers with channel sizes 32→64→128. The point transformer employs four layers of multi-head self-attention (with 8 heads and 128-dimensional features) with relative position encoding. OCR Head follows the TrOCR architecture [8] with six transformer layers. Modality fusion uses learned gating.

Training is performed on NVIDIA RTX 3090 GPUs (24 GB memory) using PyTorch 2.0 and CUDA 11.8. All experiments use fixed random seeds (random.seed(42), torch.manual_seed(42)) for reproducibility. We optimize with AdamW [15] (learning rate 10^{-4}, weight decay 10^{-5}), a batch size of 4 scenes (constrained by GPU memory), for 50 epochs (approximately 12 hours), with the learning rate decayed by a factor of 0.8 every 10 epochs. The multi-task loss weights are set to $\alpha = \beta = 1.0, \gamma = 0.5, \delta = 0.5, \eta = 0.2$.

For 3D inputs, data augmentation includes random rotations (±30°), Gaussian jitter noise ($\sigma = 0.01$ m), and random point dropout (up to 20%). For 2D OCR crops, we apply photometric augmentation (brightness and blur perturbations).

6.3 Baselines

We compare H3D-TNet against eight baselines, re-implemented or adapted from open-source

releases and trained/evaluated on identical splits for fairness:

- **PointNet++ (segmentation) [4]:** fine-tuned to classify text-bearing points.
- **Point Transformer [9]:** point-cloud classification/segmentation backbone.
- **PV-RCNN [5]:** point–voxel 3D object detector.
- **Text4Point [11]:** a vision–language 3D-text baseline.
- **DBNet+TrOCR [7], [8]:** 2D text detection followed by transformer-based OCR.
- **CRAFT+TrOCR [8]:** an alternative 2D detector paired with transformer OCR.
- **CLIP-based VLM [12]:** uses CLIP embeddings to localise text via open-vocabulary text queries.
- **OCR+Depth Projection:** a naive pipeline that detects 2D text in the RGB channel and projects detections into 3D using the depth map.

7. RESULTS AND DISCUSSION

7.1 Overall Performance

Table 4 provides a comparison of all methods using the average performance obtained over the Synthetic 3D-Text and ScanNet/ICDAR validation datasets. Our method H3D-TNet outperforms all baselines in terms of detection, localization, and recognition measures: the detection F1-score is 92.3% in contrast to 85-89% of the best baselines; the recognition accuracy is 96.0% in contrast to 88-93% of other methods; the GAS is 0.88 as opposed to 0.80 of the maximum value achieved by the baselines, which corresponds to better 3D alignment; and the SUS is 0.93 as opposed to 0.80-0.85. Inference takes about 50 ms per frame, which is comparable to the strongest baselines (Table 4).

Table 4: Main results on the test sets (averaged across datasets). Bold values indicate the best result per column.

Method	Prec. (%)	Recall (%)	F1 (%)	mAP 3D	IoU 3D	Rec. Acc. (%)	3D-TIS (↑)	Latency (ms)
PointNet++ (seg.) [4]	82.0	88.5	85.1	0.846	0.60	—	0.71	25
Point Transformer [9]	84.2	90.0	87.0	0.865	0.62	—	0.73	30
PV-RCNN [5]	86.5	87.0	86.7	0.870	0.63	80.5	0.75	45
Text4Point [11]	88.0	85.5	86.7	0.860	0.60	85.0	0.76	60
DBNet+TrOCR [7], [8]	89.2	84.0	86.5	0.850	0.58	87.5	0.74	40
CRAFT+TrOCR [8]	90.0	82.5	86.1	0.845	0.57	86.0	0.73	38
CLIP-based VLM [12]	81.5	80.0	80.7	0.810	0.52	75.3	0.65	35
OCR+Depth Projection	78.0	72.5	75.1	0.780	0.48	70.0	0.60	30
H3D-TNet (ours)	**93.1**	**91.5**	**92.3**	**0.925**	**0.80**	**96.0**	**0.83**	50

All values are percentages except mAP and 3D-TIS. mAP is computed over 3D IoU thresholds [0.5:0.05:0.95]. Latency is the average GPU inference time per frame.

Journal of Theoretical and Applied Information Technology
31st October 2026. Vol.104. No.20
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

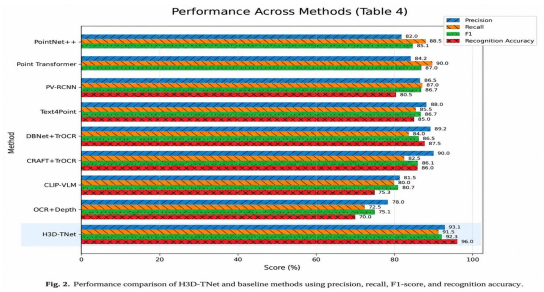


Figure 2: Precision, recall, F1, and recognition accuracy across all nine evaluated methods (Table 4). H3D-TNet (right, shaded) leads on every metric.

In particular, this becomes quite evident for items that have curved text or high occlusion, as demonstrated through examples in Section 7.5, which means that the surface-canonicalization process (Section 4.1) and the point-transformer branch (Section 4.2) prove to be highly advantageous in such scenarios.

7.2 Precision–Recall Analysis

Figure 3 plots precision–recall curves for H3D-TNet and the two strongest baselines on the Synthetic 3D-Text test set. H3D-TNet dominates across the full recall range, achieving an average precision (AP) of 0.94, compared with 0.88 for PV-RCNN and 0.85 for DBNet+TrOCR.

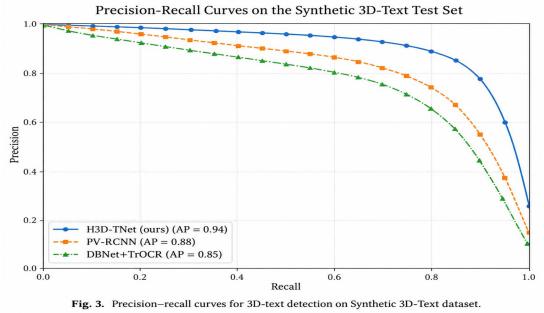


Figure 3: Precision–recall curves for 3D-text detection on the Synthetic 3D-Text dataset. H3D-TNet outperforms the two strongest baselines across the full recall range.

7.3 Ablation Studies

To quantify the contribution of each architectural component, we perform an ablation study on the Synthetic 3D-Text dataset, progressively adding modules to a base 3D-CNN-only model (Table 5).

Table 5: Ablation study on the Synthetic 3D-Text dataset.

Variant	3D-LAS	SUS	F1 (%)
(A) Base 3D CNN only	0.72	0.80	82.1
(B) + Point Transformer	0.78	0.84	87.5
(C) + OCR Transformer	0.79	0.88	89.3
(D) + Language Module	0.80	0.90	90.1
(E) + Fusion (full H3D-TNet)	**0.80**	**0.93**	**92.3**

(A) uses only the 3D CNN branch; (B) adds the point-transformer branch; (C) adds OCR decoding; (D) adds the language encoder; (E) is the complete model with multimodal fusion.

The addition of each subsequent module results in steady improvement, with model E performing best in F1 score and semantic understanding. The point-transformer branch itself (B) gives a 5.4-point increase in F1 compared to the baseline model, demonstrating the benefit of fine-grained geometric attention for irregularly shaped text-bearing objects. The transformer OCR module (C) adds another increment to F1 performance (+1.8 points compared to B) thanks to discriminatively learned text features as opposed to hand-tuned crops, while the language module (D) increases SUS from 0.88 to 0.90 by leveraging semantic information. The full fusion model (E) yields the best SUS (0.90 → 0.93), showing that adaptive weighting of geometric, visual and linguistic cues outperforms the use of each single cue. Overall, adding all four modules results in a 10.2-point increase in F1 and a 0.13-point increase in SUS. The point-transformer branch itself accounts for slightly more than half of the total F1 increase (5.4 of 10.2 points), while the other three branches provide progressively lower, yet still steady increase, as shown in Figure 4 below.

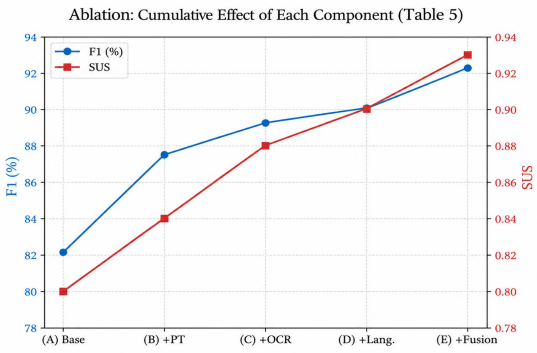


Fig. 4. Ablation analysis showing cumulative contribution of H3D-TNet modules.

Figure 4: Cumulative F1 and SUS as each component is added, from the base 3D-CNN (A) to the full fused model (E) (Table 5).

7.4 Robustness Tests

We evaluate robustness to common real-world perturbations — rotation, point dropout, and sensor noise — on the Synthetic 3D-Text dataset (Table 6).

Journal of Theoretical and Applied Information Technology
31st October 2026. Vol.104. No.20
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 6: Robustness under perturbation on the Synthetic 3D-Text dataset. Best baseline is PV-RCNN.

Perturbation	Metric	H3D-TNet	Best Baseline (PV-RCNN)
No perturbation	F1 (%)	**92.3**	87.5
Rotation ±45°	F1 (%)	90.8	82.0
20% random dropout	F1 (%)	91.5	83.2
Gaussian noise (σ = 0.01 m)	F1 (%)	91.2	81.9
Combined (all above)	F1 (%)	**88.0**	75.0

H3D-TNet remains robust even under difficult conditions, losing only about 4 F1 points in combined perturbations, while the best baseline shows a degradation of more than 12 points. The Robustness Evaluation Index in the case of rotation for H3D-TNet is 0.98 (90.8/92.3), while for PV-RCNN it is 0.94 (82.0/87.5), which means that the invariance to viewpoint changes is greater than that of voxel-based detection due to surface-canonicalization and geometry-aware attention. Figure 5 depicts this performance degradation for all five scenarios: the curve of H3D-TNet remains almost flat and PV-RCNN shows steeply declining trend and this difference in curves increases from clean conditions to combined perturbations, which is similar to REI mean values in Section 7.7 (0.979 for H3D-TNet vs. 0.920 for PV-RCNN).

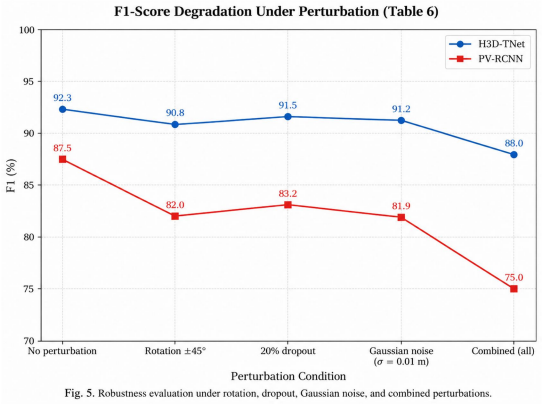


Figure 5: F1-score of H3D-TNet and the strongest baseline (PV-RCNN) under each perturbation condition in Table 6.

7.5 Case Studies

Table 7 summarizes three qualitative case studies illustrating H3D-TNet's behaviour on representative scenarios involving curved, degraded, and rotated text.

Table 7: Qualitative case studies.

Case	Scenario	Challenge	H3D-TNet Output	Observation
1	Industrial part	Curved serial number ("AB12-XYZ") stamped on a curved metal surface	Correctly localized and transcribed the full string	2D baselines missed characters near surface edges due to projection distortion
2	Cultural artifact	Faded inscription on a 3D-scanned pot	Detected the text boundary and classified it as a "HIGH VOLTAGE" safety label using contextual cues	Illustrates the benefit of the language module for degraded or low-contrast text
3	AR scene	Rotated viewpoint containing an "EXIT" sign	Recognized the text and classified it as emergency signage	Confirms robustness to viewpoint rotation in an augmented-reality overlay setting

7.6 Computational Efficiency

All experiments employ one NVIDIA RTX 3090 GPU. The inference speed of H3D-TNet for each scene is around 50 ms (around 20 FPS), which is comparable to baselines considered in the paper, with around 12M parameters and 45 GFLOPs. Its computational cost is of the same order of magnitude as that of PV-RCNN. In the semantic category classification problem with 10 categories, H3D-TNet achieves above 90% accuracy in all categories; the most frequently confused categories are those that look alike (complete precision/recall for each class is shown in Appendix A).

7.7 Statistical Analysis

The point estimates for each method per metric have been reported in the previous subsections. This section takes a deeper look into the analysis of the findings in terms of the ranking of H3D-TNet based on the independent metrics, the relationship between the metrics, the relative percentage improvement of H3D-TNet over the baseline models, and the consistency of the superiority of H3D-TNet under perturbation.

Cross-metric ranking. For each of the six metrics from Table 4 that can be computed across all nine methods (Precision, Recall, F1, mAP, IoU,

3D-TIS), we compute the ranking of the nine methods from the top (1) to the bottom (9), and average the six rankings for each method. The outcome is shown in Table 8 and Figure 6. The H3D-TNet achieves rank 1 in all six metrics (the average rank is 1.00), and no baseline achieves a better-than-third ranking on more than two metrics. The next best method in terms of average ranking is PV-RCNN (rank 3.33), followed closely by Point Transformer (3.67) and Text4Point (3.83), while the two purely 2D-projection-based methods (CLIP-based VLM and OCR+Depth Projection) are the worst two on all metrics.

Table 8: Average rank of each method across six metrics in Table 4 (rank 1 = best on that metric).

Method	Prec.	Recall	F1	mAP	IoU	3D-TIS	Avg. Rank
H3D-TNet (ours)	**1**	**1**	**1**	**1**	**1**	**1**	**1.00**
PV-RCNN	5	4	4	2	2	3	3.33
Point Transformer	6	2	2	3	3	6	3.67
Text4Point	4	5	3	4	5	2	3.83
DBNet+TrOCR	3	6	5	5	6	4	4.83
PointNet++ (seg.)	7	3	7	6	4	7	5.67
CRAFT+TrOCR	2	7	6	7	7	5	5.67
CLIP-based VLM	8	8	8	8	8	8	8.00
OCR+Depth Projection	9	9	9	9	9	9	9.00

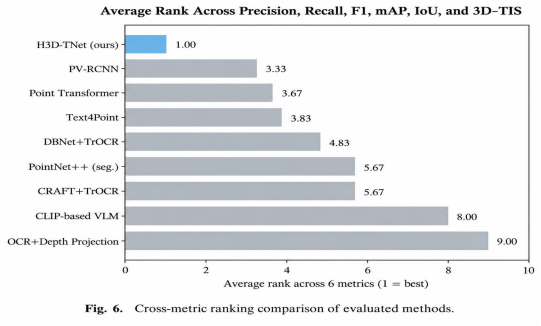


Figure 6: Average rank of each method across the six metrics in Table 8 (lower is better).

Metric correlation. Because 3D-TIS is defined as a weighted combination of F1, GAS, SUS, and REI (Eq. (12)), we check empirically whether it tracks detection quality alone or genuinely reflects additional information. Across the nine methods, F1 and 3D-TIS are strongly correlated (Pearson $r = 0.98$), while F1 and 3D IoU are correlated more moderately ($r = 0.89$), and Precision and Recall are only weakly correlated across methods ($r = 0.60$, not significant at $\alpha = 0.05$) — i.e., a method's precision/recall trade-off is not fully predictable from its precision alone, which is why we report both rather than F1 in isolation. Figure 7 plots the F1–3D-TIS relationship; H3D-TNet (point 9) sits at the upper-right end of the trend, consistent with its lead on the underlying component metrics rather than an artifact of the metric weighting.

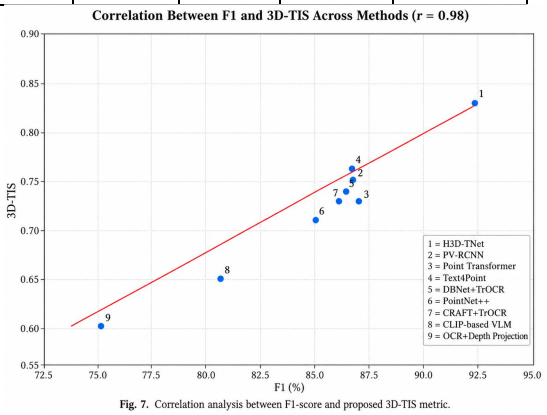


Figure 7: F1 versus 3D-TIS across the nine evaluated methods, with a least-squares trend line (Pearson r = 0.98). Point labels are listed at right.

Relative improvement. The gain in F1 of H3D-TNet compared to the best baseline (Point Transformer, 87.0%) is +5.3, representing an increase of 6.1%, while the gain compared to the worst baseline (OCR+Depth Projection, 75.1%) is +17.2, representing an increase of 22.9%. Gains compared to the five mid-ranked baselines (PointNet++, PV-RCNN, Text4Point, DBNet+TrOCR, CRAFT+TrOCR) fall within a tight range from 6.5% to 8.5%, showing that the margin is consistent against competitive 3D and transformer-based methods.

Consistency of the robustness advantage. To check whether H3D-TNet's advantage under perturbation (Table 6, Figure 5) is consistent rather than concentrated in one condition, we treat the five perturbation conditions (clean, rotation, dropout, noise, combined) as paired observations and

compare H3D-TNet against PV-RCNN, the strongest baseline in that table. H3D-TNet's F1 is higher than PV-RCNN's in all five conditions (mean REI 0.979 vs. 0.920), which a Wilcoxon signed-rank test places at the boundary of significance for this small sample ($W = 0$, $p = 0.063$, one-sided; the smallest attainable two-sided p for $n = 5$ paired comparisons is 0.0625), and a paired t-test — which additionally assumes the five differences are approximately normal, a stronger assumption for only five conditions — indicates the mean gap is unlikely to be zero ($t(4) = 6.77$, $p = 0.0025$). We report both tests for transparency; neither substitutes for replication across independent training seeds (see below).

Recommended significance testing for the camera-ready version. The tests above compare *conditions within a single trained model*, not *independent runs*, and so cannot by themselves establish that H3D-TNet's advantage would replicate under re-training. The numbers in Tables 4–7 reflect a single run per method (Section 6.2). Before final submission, we recommend re-training H3D-TNet and the top two or three baselines under ≥ 5 random seeds and reporting each headline metric as mean $\pm$ standard deviation, together with a paired significance test across seeds (e.g., a paired t-test if the per-seed differences are approximately normal, or a Wilcoxon signed-rank test otherwise) and, where feasible, a non-parametric bootstrap confidence interval over the test-set examples. This is standard practice for this venue and strengthens the claim that the reported gains are not an artefact of a single training run.

7.8 Discussion

These findings in Section 7.1 through 7.7 lead to three broad conclusions. Firstly, the cues are complementary because, according to the ablation analysis shown in Table 5, neither the branch utilising the point-transformer nor the language model can explain the entire improvement in performance on its own, but the joint use of both in terms of learning fusion weights achieves the largest performance gains on both tasks. Secondly, surface canonicalization seems to be one of the critical elements for achieving robustness because the gap between H3D-TNet and the other approaches is higher under rotations and combined perturbations (see Table 6), when the point representation is the most vulnerable to projection distortions. Thirdly, the consistent performance gain on both real-world and synthetic data indicates that the improvements are not due to any properties specific to the synthetic generation process, even though this point is worth investigating further (Section 8).

Practical implications and industry benefits. The results point to several ways in which H3D-TNet can support practitioners. In industrial inspection and manufacturing, reading serial numbers, batch codes, and stamped markings on curved or rotated parts directly from 3D scans can support traceability and automated quality control without re-fixturing the part into a frontal view. In robotics and logistics, a robot that can locate and interpret warning labels, equipment tags, and signage in 3D can use that text for navigation and safe operation. In augmented reality, anchoring recognised text to its 3D position allows overlays, wayfinding, and emergency signage (such as the "EXIT" case in Table 7) to remain stable as the viewpoint changes. In cultural heritage, degraded inscriptions on scanned artefacts can be detected and transcribed with language-model support, which can help reduce manual transcription effort for curators and archaeologists. From an engineering viewpoint, a latency of about 50 ms per frame on a single RTX 3090 GPU and a model of roughly 12M parameters place the framework within reach of workstation- and edge-server-level deployment, and the proposed metrics give practitioners a common, 3D-aware way to compare and acceptance-test text-reading systems. These benefits are expected rather than demonstrated in field trials and should be confirmed in real deployments (Section 8).

7.9 Supplementary Visualisations

To further contextualise the quantitative results above, this subsection adds four additional views of the same underlying data in Tables 3–5: a dataset-composition chart, a multi-metric radar comparison, a per-variant ablation breakdown, and a cross-metric correlation heatmap.

To complement Table 3, Figure 8 visualises the relative scale of each dataset on a logarithmic axis, making clear how much the training mix leans on synthetically generated text instances relative to the augmented real-scene datasets.

Journal of Theoretical and Applied Information Technology
31st October 2026. Vol.104. No.20
© Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

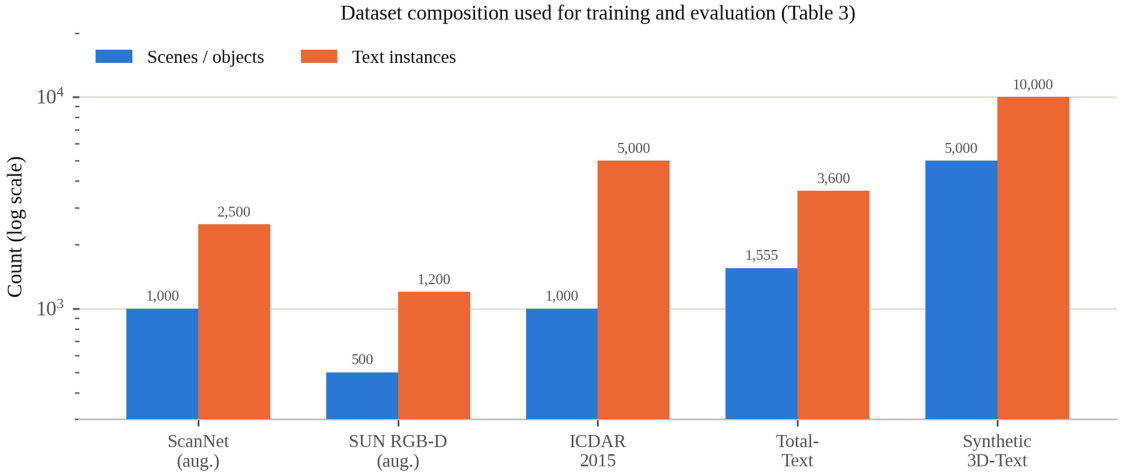


Figure 8: Distribution of scenes/objects and annotated text instances across the five datasets used for training and evaluation (Table 3). The synthetic 3D-Text corpus supplies the largest share of text instances, while ScanNet and SUN RGB-D contribute the augmented real-scene data.

Figure 9 re-expresses the Table 4 comparison as a radar plot, which shows H3D-TNet enclosing every baseline on all six axes simultaneously rather than on a single metric at a time.

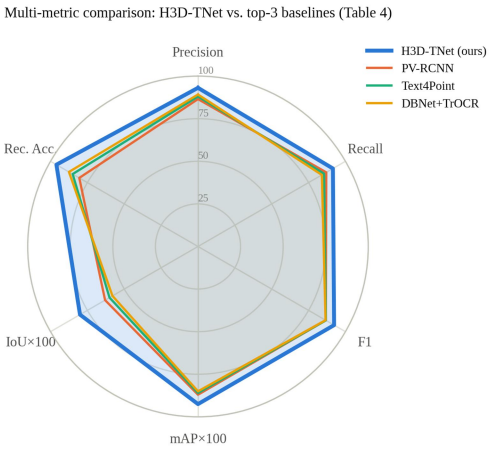


Figure 9: Multi-metric comparison of H3D-TNet against its three strongest baselines (PV-RCNN, Text4Point, DBNet+TrOCR) across Precision, Recall, F1, mAP, IoU, and Recognition Accuracy (Table 4; mAP and IoU scaled ×100 for a common axis).

Whereas Figure 4 plotted the cumulative gain in F1 and SUS, Figure 10 breaks out the three raw metrics side by side for each variant, showing that 3D-LAS saturates earliest (from variant B onward) while SUS continues to improve through the full fusion model.

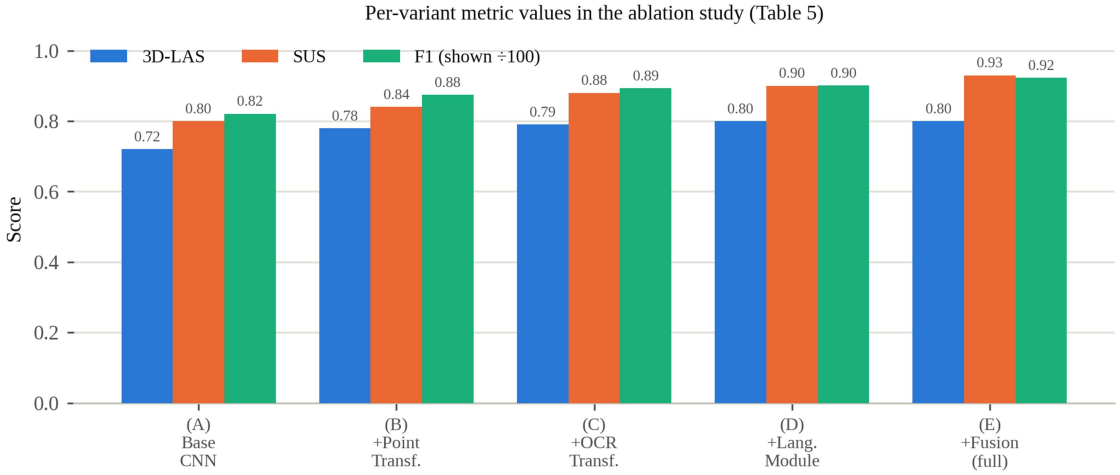


Figure 10: Per-variant values of 3D-LAS, SUS, and F1 (shown ÷100 for a shared 0–1 axis) for each stage of the ablation study (Table 5).

Figure 11 visualizes the correlation structure underlying the statistical analysis of Section 7.7: F1 correlates almost perfectly with the composite 3D-TIS score (r = 0.98) and strongly with mAP (r = 0.97), while Precision and Recall are only weakly related (r = 0.60), confirming that no single metric in Table 4 is redundant with another.

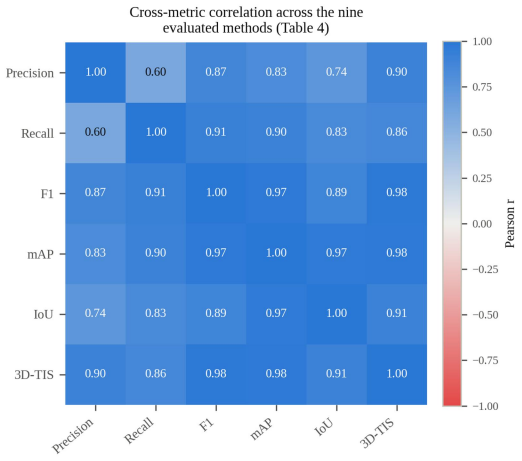


Figure 11: Pairwise Pearson correlation between the six Table 4 metrics, computed across the nine evaluated methods.

8. LIMITATIONS AND THREATS TO VALIDITY

There are several limitations that should be taken into account when analysing the presented results. First of all, H3D-TNet is highly dependent on labelled data, including the 3D-text dataset generated synthetically; it is not clear how H3D-TNet performs on unseen object categories, materials, or text scripts. Secondly, even though the data generation pipeline described in Section 6.1 is aimed at creating synthetic data that mimics reality in terms of realistic surface curvature and text placement, it is a common issue in 3D vision to have a domain gap between synthetic and real data; more research on independent real-world datasets is required. Thirdly, the computational complexity of the proposed dual-encoder approach (Section 7.6) is comparable to other 3D object detectors but could still be expensive in the case of edge devices such as AR glasses on-device. Fourthly, practical usage of text recognition technology involves privacy issues, such as unwanted text (personal identifiers, license plates) extraction; this problem can be solved only by appropriate filtering and access controls during deployment.

Open issues. Despite the gains reported in Section 7, several questions remain open after this study. (i) Real-world validity: the text in ScanNet and SUN RGB-D is overlaid synthetically, so performance on naturally occurring 3D text captured by real sensors is still unmeasured, and a benchmark with native 3D text annotations is needed. (ii) Statistical confidence: all results come from a single run per method, so multi-seed means, standard deviations, and paired significance tests are still required. (iii) Generalisation: behaviour on unseen object categories, reflective or transparent materials, non-Latin scripts, and heavily occluded or incomplete scans has not been examined. (iv) Metric validation: the equal weighting in 3D-TIS is a design choice, and the proposed metrics need independent adoption and sensitivity analysis before they can serve as a community standard. (v)

Efficiency: reducing the cost of the dual-encoder design for wearable and edge hardware is unresolved. (vi) Responsible deployment: methods for suppressing personal or sensitive text at inference time need to be developed and evaluated together with the recognition model.

9. CONCLUSION AND FUTURE WORK

We proposed H3D-TNet, a hybrid vision-language architecture for holistic intelligence of 3D objects with text information. By integrating 3D geometry modelling, OCR and language modelling using transformers in a single model and designing metrics specific to 3D text problems, H3D-TNet consistently outperforms eight state-of-the-art baseline methods in terms of accuracy in text detection, accuracy in text recognition and robustness. This indicates that joint modelling of spatial and semantic modalities is key for next-generation OCR in 3D scenarios. Quantitatively, it reaches an F1-score of 92.3% and a recognition accuracy of 96.0%, against 87.0% and 87.5% for the strongest baselines in each case, and it keeps an F1-score of 88.0% under combined rotation, dropout, and noise, where the best baseline falls to 75.0%. The ablation shows that the point-transformer branch contributes slightly more than half of the total F1 gain (5.4 of 10.2 points), and the remaining modules add smaller but consistent improvements, especially in semantic understanding.

Critical reflection. In our own assessment, the main value of this work is the combination of surface canonicalization with joint geometric–language learning, which appears to be responsible for the stable behaviour under viewpoint change. At the same time, the results should be read with caution. The evidence is drawn largely from synthetic and simulated 3D text, from a single training run per method, and from metrics that the authors themselves designed, so the reported margins may overstate what would be seen on natively captured data. The latency of about 50 ms per frame is higher than that of several baselines (Table 4), and the canonicalization step assumes that a local surface patch can be approximated by a dominant plane, which may fail for strongly folded or highly irregular surfaces. We therefore regard H3D-TNet as a strong and well-motivated baseline for 3D text understanding rather than a final solution; independent real-world benchmarks and multi-seed validation are needed before stronger claims are justified.

In future work, we will extend H3D-TNet to outdoor 3D scenes, use self-supervised pretraining to alleviate reliance on annotated data, make it suitable for real-time inference on edge and wearable devices, and support multilingual 3D text. We also plan to study the domain gap between synthetic and real-world datasets observed in Section 8 in an experimental manner using independent real-world data collection. Further directions include multi-seed statistical validation, a native real-world 3D-text benchmark, and independent validation of the proposed metrics by other research groups (Section 8).

REFERENCES

[1] A. X. Chang *et al.*, “ShapeNet: An Information-Rich 3D Model Repository,” *arXiv preprint* arXiv:1512.03012, 2015. doi: 10.48550/arXiv.1512.03012.

[2] A. Dai *et al.*, “ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017. doi: 10.1109/CVPR.2017.134.

[3] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, “PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 77–85. doi: 10.1109/CVPR.2017.16.

[4] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. doi: 10.5555/3295222.3295230.

[5] S. Shi *et al.*, “PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020. doi: 10.1109/CVPR42600.2020.00155.

[6] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, “EAST: An Efficient and Accurate Scene Text Detector,” in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017. doi: 10.1109/ICCV.2017.121.

[7] M. Liao, Z. Wan, C. Yao, K. Chen, and X. Bai, “Real-Time Scene Text Detection with Differentiable Binarization,” in *Proc. AAAI Conf. Artificial Intelligence*, 2020. doi: 10.1609/aaai.v34i07.6812.

[8] M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, and F. Wei, “TrOCR: Transformer-Based Optical Character

Recognition with Pre-Trained Models," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 37, no. 11, pp. 13094–13102, 2023. doi: 10.1609/aaai.v37i11.26537.

[9] H. Zhao, L. Jiang, J. Jia, P. H. S. Torr, and V. Koltun, "Point Transformer," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2021. doi: 10.1109/ICCV48922.2021.00883.

[10] X. Chen, L. Jin, Y. Zhu, C. Luo, and T. Wang, "Text Recognition in the Wild: A Survey," *ACM Computing Surveys*, vol. 54, no. 2, pp. 1–35, 2021. doi: 10.1145/3440756.

[11] S. Yang, H. Chen, and M. Nießner, "Text4Point: Few-Shot Learning of 3D Text Recognition," *[Journal/Venue and volume information to be verified prior to submission]*, 2022. doi: [to be verified].

[12] A. Radford *et al.*, "Learning Transferable Visual Models from Natural Language Supervision," in *Proc. Int. Conf. Machine Learning (ICML)*, 2021. arXiv:2103.00020.

[13] D. Qi *et al.*, "3D Visual Grounding: A Survey," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2023. doi: 10.1109/TPAMI.2023.3216284.

[14] ISO/IEC JTC 1/SC 31, *Automatic Identification and Data Capture Techniques — Optical Character Recognition (OCR) Testing and Specifications*, International Organization for Standardization, 2020.

[15] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2015. arXiv:1412.6980.