TRACE SERIES · REPORT 2

# TRACE Commercial: Adapting Wall Recovery to Commercial Floor Plans

Kyle Rossignol · Foreman AI Research · Colorado, USA · hello@foremanai.co · September 2026

ABSTRACT

TRACE (Tokenized Representation of Architectural CAD Elements) is Foreman AI's wall-recovery model. It turns the vector CAD elements inside a PDF floor plan into line tokens, learns a representation for each token, decides which tokens are walls and which belong to the same wall, and emits measured wall runs in exact PDF coordinates. Its learned stages were trained on hand-traced house plans, and on commercial sheets it misses walls and confuses exterior with interior. We traced 25 commercial sheet pieces (22,828 LF of walls) from our corpus of nearly 98,000 real engineered plans (about 3.6 million pages), split by project into 20 training pieces and 5 held-out sheets, and adapted TRACE's pixel wall model, RoomNet, by fine-tuning from its production weights with commercial tiles sampled ten times as often. Held-out commercial pixel F1 rises from 0.889 to 0.943, driven by recall (0.859 to 0.971); house validation F1 falls from 0.914 to 0.900, so we keep separate house and commercial models. Through the complete production pipeline, found wall length on the held-out sheets rises from 77.4% to 93.0% of 6,482 LF and missed length falls from 1,466 to 455 LF, while invented length rises from 447 to 746 LF. Exterior/interior accuracy on matched walls rises from 78.7% to 84.1%. Adding traced sheets beyond thirteen pieces no longer moves the pixel metric. Retraining the downstream line picker on inputs regenerated through the production code path lowers its held-out token loss by 12% but leaves full-pipeline results unchanged, which localizes the remaining error to the stages that admit invented walls rather than to wall detection.

*construction takeoff · floor-plan understanding · wall recovery · vector geometry · domain adaptation · fine-tuning*

## 1. Introduction

A takeoff starts with wall length. Stud counts, plates, sheathing, insulation, gypsum board and several finishes are derived directly from recovered wall geometry, so a wall model that is wrong in a systematic way propagates that error into every downstream quantity [2]. TRACE recovers walls from the vector primitives already present in a construction PDF [1]. Deterministic geometry proposes candidate wall lines; learned models decide which candidates are walls, where they end, which of them describe the same physical wall, and which side of the building each run faces.

TRACE was developed on residential permit drawings. RoomNet, its pixel wall model, was trained on 76 hand-traced house sheets, and the line picker (QuantityNet) on 71 of them with 20 held out for validation [1]. On those populations it performs well. Commercial drawings differ in ways the house corpus rarely shows: poché-filled masonry, double-line demising and rated walls, dense casework, column grids, keynote tags, hatch fields, and scales from 1/8 in. to 3/8 in. per foot. In production these sheets failed in three visible ways: long runs traced as fragments, whole walls dropped, and interior partitions labelled exterior.

Most floor-plan work starts from raster images and predicts walls, rooms or junctions as pixels or graphs [6, 7]; blueprint vectorization has also been attempted directly on high-resolution drawings [8]. TRACE instead keeps the PDF's own vector primitives and uses learning only to verify them [1]. Adapting a trained network to a small new domain by fine-tuning is standard practice [9], with the known risk of forgetting the source domain [10]; both effects appear below.

This paper reports the first adaptation of TRACE to commercial plans. It deliberately replaces an earlier attempt to generate commercial training data with hand-written geometric rules, which did not transfer across drafting styles (Sec. 3.3). The contributions are:

- a commercial gold set of 25 traced sheet pieces with checked scales and exterior, interior and opening classes, split by project;
- a fine-tuning recipe for RoomNet with ablations over data size, class weight, initialization, scale augmentation and seed;
- a full-pipeline evaluation on held-out sheets that runs the production client and runtime code in a library environment matched to the production worker; and
- a production-path data builder for the downstream line picker and exterior/interior heads, with a negative result that localizes the remaining error.

## 2. The TRACE model

The name describes the method. TRACE never draws walls from pixels. It first turns a sheet into *tokens*: each token is one candidate wall line recovered from the PDF's own vector

CAD elements, such as parallel face pairs, filled wall bodies, stroke envelopes and bounded continuations. A token then receives a learned representation, and every output coordinate is inherited from a token rather than from a raster (Fig. 1).

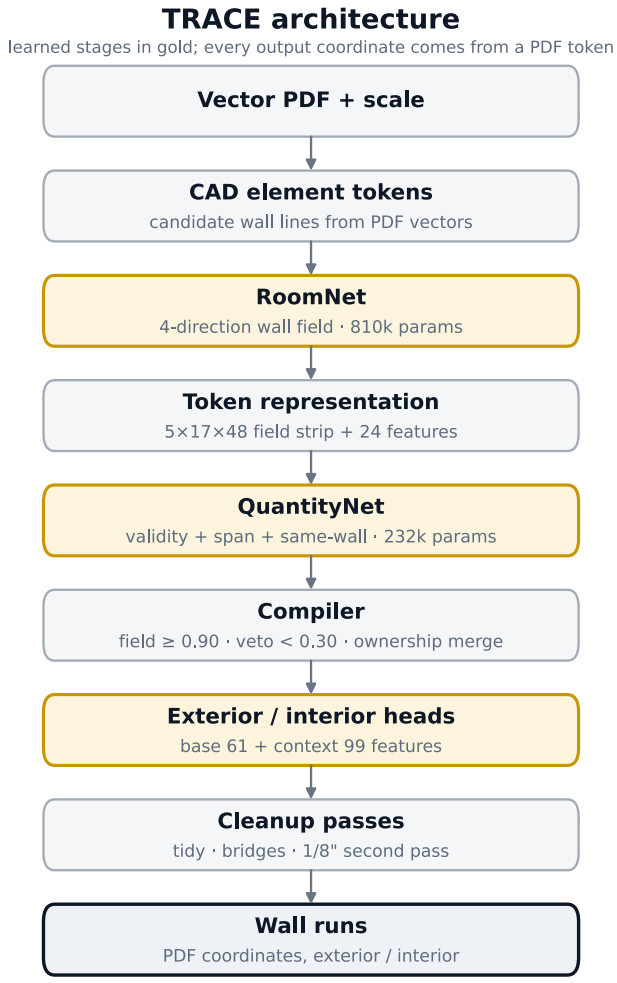


***Figure 1:*** *TRACE stages. Gold boxes are learned. The wall field from RoomNet enters every later learned stage: the line picker samples it along each token, and the exterior/interior heads compute features from it. Changing RoomNet therefore changes the inputs of every downstream model.*

**RoomNet** is a five-level U-Net [3] with encoder widths 12–24–48–96–128 and 810,096 parameters. It reads the sheet rendered in grayscale at 4 pixels per foot and emits four wall-probability channels for walls near 0°, 45°, 90° and 135°.

**Token representation.** For each token, a 5×17×48 strip is sampled along it at 8 px/ft: ink darkness plus the four RoomNet channels rotated into the token's frame, 2 ft wide and 48 samples long. Twenty-four numeric features add length, width, field statistics, face-pair support and extraction path.

**QuantityNet**, the line picker (231,588 parameters), maps the representation to a validity logit and trimmed span endpoints. A pair head receives two token embeddings and eight pair features and predicts whether overlapping parallel tokens are the same physical wall.

**Compiler and heads.** The compiler keeps RoomNet intervals at or above 0.90, vetoes tokens the line picker scores below 0.30, and merges tokens into runs by ownership. Two small MLP heads then label each run: a base head on 61 geometric and field features and a residual context head that adds 38 room-context features. Cleanup passes (tidy, room bridges, outline snap and a 1.5× second pass on 1/8 in. sheets) follow.

Because the wall field enters the line picker's strips and the heads' features, a new RoomNet changes the inputs of every later learned stage. This coupling determines the experimental order in Sec. 4.

## 3. Data

### *3.1 Commercial gold*

Sheets were drawn from Foreman AI's corpus of nearly 98,000 real engineered plans (about 3.6 million pages), choosing commercial floor plans across building types and drafting styles: clinics, a surgery center, restaurants, office and retail tenant improvements, a bank branch and a winery. Walls were traced as centrelines tagged exterior, interior or bearing, with openings drawn as spans along the wall line.

Scale was checked sheet by sheet against the printed scale nearest the traced drawing, with written dimensions used for ambiguous cases; a wrong scale silently halves or doubles every quantity measured in feet. One sheet held a 1/8 in. plan and a 1/4 in. enlarged core, so it was split into two cropped pieces with separate scales. Every piece was traced or reviewed by hand. The five held-out sheets come from five projects, none shared with a training piece (Fig. 2, Table 1).


***Figure 2:*** *Commercial gold and evaluation populations. The held-out sheets are split by project; the house gold and its splits are unchanged from the production system.*

### *3.2 House gold*

The house corpus is unchanged: 112 traced sheets, of which the line picker used 71 for training and 20 for validation, plus 8 external evaluation sheets [1]. The same 20 validation sheets serve as the house check in every experiment, so each result reports both what commercial gained and what houses lost.

### *3.3 A negative result: rule-based gold*

Before training, we attempted to generate commercial labels automatically with hand-written geometric rules for vector wall pairs, door arcs, outlines, symbol filters and alignment checks. Each new drafting style broke rules written for the previous one: symbol filters deleted real wall stubs, alignment checks rejected whole sheets, and gap bridges drew diagonals through rooms. The process converged only on plans the rules already fit, which would have taught the model the rules' blind spots. We abandoned it in favour of human tracing. The episode is the practical argument for learned verification in [1]: rules do not transfer across drafting styles; traced examples do.


***Figure 3:*** *Four training examples as the trainers receive them: traced centrelines over the faded sheet. Openings are spans on the wall line, so a wall's gross length runs through its doors.*

***Table 1:*** *Commercial gold. Walls and openings are traced segments; LF are linear feet at the checked scale.*

| Sheet | Split | Scale | Walls | Open. | Ext. LF | Int. LF | Open. LF |
|---|---|---|---|---|---|---|---|
| T01 | train | 3/16" | 110 | 33 | 392 | 310 | 75 |
| T02 | train | 1/8" | 116 | 30 | 505 | 814 | 95 |
| T03 | train | 1/8" | 191 | 53 | 680 | 1,436 | 148 |
| T04 (bank branch) | train | 1/4" | 181 | 19 | 240 | 296 | 65 |
| T05 (bookstore) | train | 1/4" | 195 | 5 | 146 | 203 | 20 |
| T06 (office, existing/demo) | train | 1/8" | 171 | 76 | 578 | 1,008 | 184 |
| T07 (plan) | train | 1/8" | 128 | 36 | 507 | 1,020 | 81 |
| T08 (enlarged core) | train | 1/4" | 52 | 17 | 171 | 172 | 30 |
| T09 | train | 1/4" | 18 | 8 | 121 | 60 | 17 |
| T10 (surgery center) | train | 1/8" | 314 | 77 | 608 | 1,744 | 241 |
| T11 | train | 1/8" | 54 | 25 | 372 | 225 | 54 |
| T12 | train | 3/8" | 45 | 16 | 134 | 118 | 30 |
| T13 | train | 1/4" | 40 | 13 | 220 | 184 | 25 |
| T14 | train | 1/4" | 63 | 26 | 196 | 271 | 62 |
| T15 | train | 1/4" | 138 | 34 | 292 | 459 | 68 |

| Sheet | Split | Scale | Walls | Open. | Ext. LF | Int. LF | Open. LF |
|---|---|---|---|---|---|---|---|
| T16 | train | 1/4" | 25 | 6 | 163 | 159 | 14 |
| T17 (plasma center) | train | 1/8" | 130 | 45 | 399 | 798 | 110 |
| T18 | train | 1/4" | 69 | 14 | 252 | 187 | 32 |
| T19 | train | 1/4" | 101 | 34 | 232 | 319 | 86 |
| T20 | train | 3/8" | 36 | 17 | 186 | 167 | 46 |
| *H1 (urgent-care clinic)* | held out | 1/8" | 161 | 61 | 481 | 1,182 | 145 |
| *H2 (winery)* | held out | 1/8" | 312 | 27 | 417 | 382 | 107 |
| *H3 (restaurant)* | held out | 1/4" | 143 | 15 | 179 | 181 | 70 |
| *H4 (angled plan)* | held out | 3/16" | 105 | 55 | 452 | 682 | 129 |
| *H5 (cardiology clinic)* | held out | 3/16" | 347 | 120 | 584 | 1,947 | 277 |
| **Training total (20 pieces)** | | | **2,177** | **584** | **6,394** | **9,948** | **1,484** |
| **Held-out total (5 sheets)** | | | **1,068** | **278** | **2,113** | **4,374** | **728** |

# 4. Method

## *4.1 RoomNet fine-tuning*

Training uses the A/B harness that produced the production house model. Sheets are rendered at 4 px/ft and traced centrelines are rasterised five pixels thick into four direction masks (Fig. 4). Random 512×512 tiles are drawn in batches of 8. The selected recipe initialises from production weights and trains for 1,500 steps with AdamW [4] at $2\times10^{-4}$ and cosine decay; tiles from commercial sheets are sampled ten times as often as house tiles.

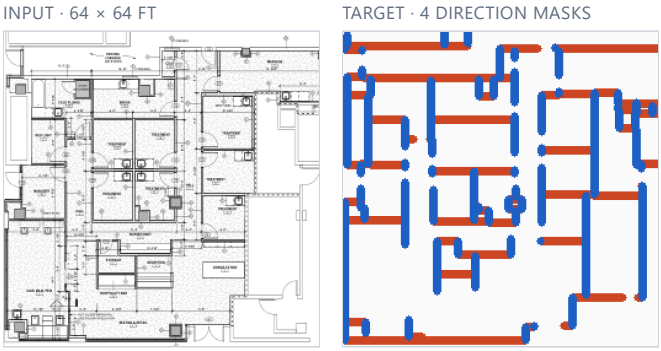


***Figure 4:*** *One actual RoomNet training tile (T15, densest 256-pixel window at 4 px/ft, shown 2×). Orange = 0°, blue = 90° wall masks rasterised 5 px thick from the traced centrelines. Door gaps and fixtures in the input carry no label.*

## *4.2 Evaluation*

**Pixel wall F1 at 1 ft** compares RoomNet's thresholded field with the gold masks, counting a pixel as matched within one foot, in the spirit of buffer-based positional accuracy [5]. It isolates the wall model but ignores every later stage.

**Full-pipeline wall length.** Each held-out sheet runs through the production client and runtime (v16.2.4), with its PDF library and software environment matched package-for-package to the production worker. Gold and traced walls are drawn as one-pixel centrelines at 8 px/ft with openings excluded. With $G$ the gold centrelines, $T$ the traced centrelines, $L(\cdot)$ length and $N_1(\cdot)$ the 1-ft neighbourhood:

$$found = L(G \cap N_1(T)), \quad missed = L(G) - found,$$
$$invented = L(T \setminus N_1(G)).$$

**Exterior/interior accuracy** samples every traced wall each half foot and compares each sample within 1 ft of a gold wall with that wall's class, weighted by length.

The production sheet gate refuses two of the five held-out sheets, and the manual box fallback refuses a third because a hatch path crosses the box edge. To measure the models rather than the gate, every held-out run uses a box around the traced drawing whose redaction keeps every path that reaches it. The box stands in for a gate that accepts commercial plans.

## *4.3 Line-picker data on the production path*

The production line picker never saw fields from the commercial RoomNet. We therefore regenerated its training data by running the production runtime's own stages on all 116 sheets: vector extraction, the 4 px/ft render, the RoomNet field and the QuantityNet input packer, each sheet producing a folder laid out exactly like a production cache. Targets reuse the original label code unchanged. A token is a real wall if at least half its length lies within 1 ft and 6° of a gold wall axis; endpoints are trimmed to the covered span; and two real tokens are the same wall if they fall on the same gold axis. On commercial sheets only tokens inside the traced drawing (gold box + 4 ft) are kept, since the rest of the sheet was never traced. Fifteen house sheets had frozen extractions holding tokens for several trial scales at once (for example 9, 13.5 and 18 pt/ft); production extracts at one calibrated scale, so those sheets were re-extracted at their gold scale. Figures 5 and 6 show actual training tokens.

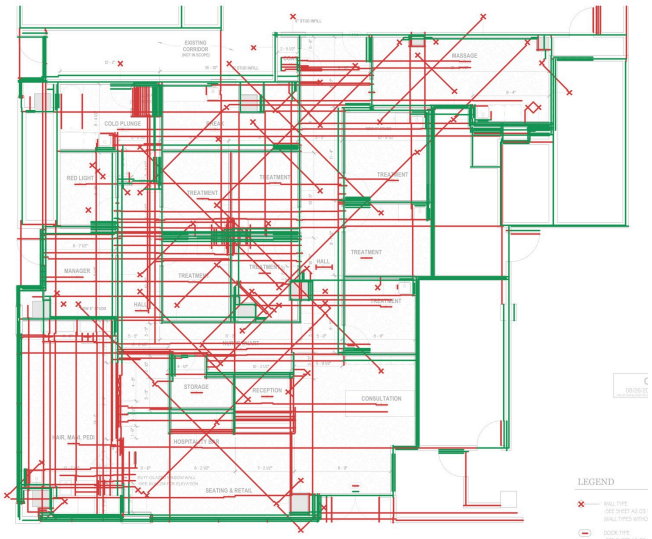

***Figure 5:*** *Every token inside the traced drawing of T15, coloured by its training label: green = wall, red = not a wall. Many tokens are the two faces of one wall; the same-wall pair head learns to merge them.*

We compare fine-tuning from the production line picker, which keeps its input normalisation, with training from scratch under the original recipe, sampling commercial tokens with weight 1× or 3×. Both use batch 192 and the original losses: length-weighted validity cross-entropy, span regression weighted 12×, and balanced same-wall cross-entropy. The training set holds about 107,000 tokens, 23,000 of them commercial.

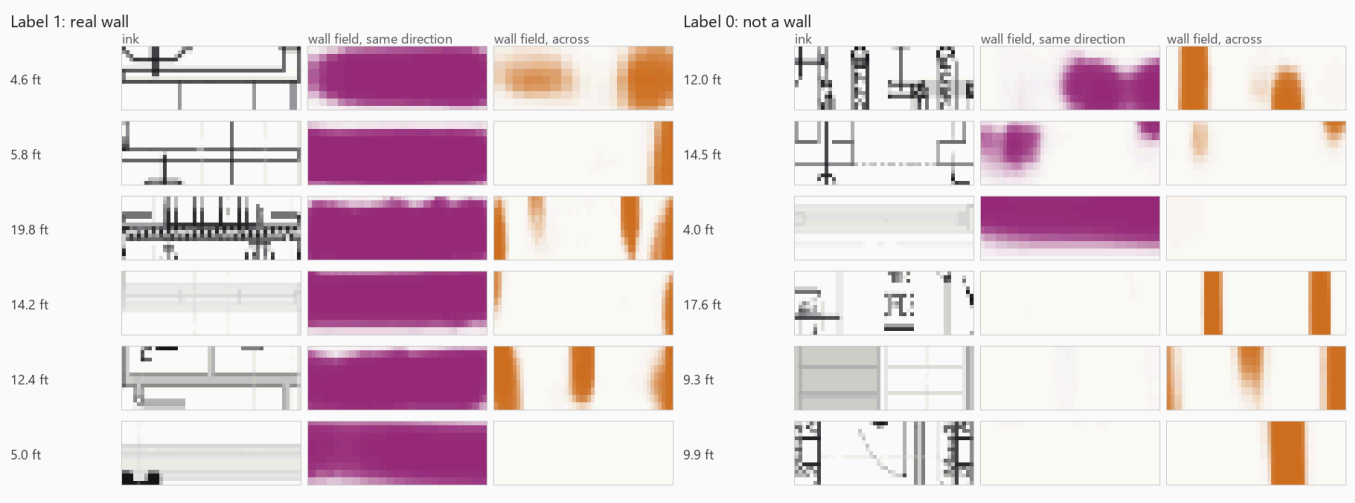


***Figure 6:*** *Twelve actual line-picker training tokens from T10, drawn at random from tokens at least 4 ft long. Each row shows three of the five input channels: ink along the token, RoomNet's field in the token's own direction, and the field across it. Left: tokens labelled wall. Right: tokens labelled not-wall. The third not-wall token has a strong field but is a light parallel line, the case the line picker exists to veto.*

## 5. Results

### *5.1 RoomNet*

The production model's commercial errors are mostly misses: precision 0.922, recall 0.859. The selected fine-tune holds precision at 0.917 and raises recall to 0.971. Table 2 and Fig. 7 list every run on identical populations.

- **Commercial data helps immediately.** Eight traced pieces lift held-out commercial F1 from about 0.89 to 0.93; houses-only retrains of the same architecture reach 0.880 and 0.892.
- **Fine-tuning matches training from scratch at a fifth of the compute** (0.928 versus 0.926 with eight pieces) and reaches 0.939–0.943 with thirteen.
- **More pieces help, then saturate.** From 8 to 13 pieces gains about a point; from 13 to 20 gains nothing measurable (0.935 and 0.940 over two seeds versus 0.943).
- **Weighting barely matters**: commercial ×3, ×6 and ×10 land within 0.004.
- **Houses pay a small price.** Every mix loses 1–1.7 points on the house set; commercial-only training keeps commercial F1 (0.935) but collapses houses to 0.740, the familiar forgetting of fine-tuned networks [10].

- **Scale augmentation hurts**: randomly rescaling tiles to imitate 1/8 in. sheets lowered commercial F1 from 0.918 to 0.904 on the earlier four-sheet test set.

Seed-to-seed spread on the commercial set is about half a point, so smaller differences are not interpreted.

***Table 2:*** *RoomNet runs, pixel wall F1 at 1 ft, all scored on the same five held-out commercial sheets, the 20 house validation sheets and the 8 external house sheets. Selected run in bold.*

| Run | Training data | Init | Comm. × | Steps | Comm. F1 | Houses F1 | External-8 F1 |
|---|---|---|---|---|---|---|---|
| Production RoomNet | 76 house sheets | – | – | – | 0.889 | 0.914 | 0.868 |
| Houses only, retrained | houses | scratch | – | 3,000 | 0.892 | 0.915 | 0.925 |
| Houses only, other seed | houses | scratch | – | 3,000 | 0.880 | 0.910 | 0.891 |
| First commercial mix | houses + 8 | scratch | ×6 | 3,000 | 0.926 | 0.906 | 0.921 |
| Fine-tune | houses + 8 | production | ×3 | 1,500 | 0.928 | 0.908 | 0.919 |
| Fine-tune | houses + 8 | production | ×6 | 1,500 | 0.933 | 0.900 | 0.914 |
| Fine-tune | houses + 13 | production | ×3 | 1,500 | 0.940 | 0.901 | 0.896 |
| Fine-tune | houses + 13 | production | ×6 | 1,500 | 0.939 | 0.898 | 0.894 |
| Fine-tune, seed 43 | houses + 13 | production | ×6 | 1,500 | 0.940 | 0.898 | 0.894 |
| **Fine-tune (selected)** | **houses + 13** | **production** | **×10** | **1,500** | **0.943** | **0.900** | **0.894** |
| Fine-tune, 2× longer | houses + 13 | production | ×6 | 3,000 | 0.941 | 0.899 | 0.904 |
| Commercial only | 13, no houses | production | – | 1,500 | 0.935 | 0.740 | 0.786 |
| Fine-tune | houses + 20 | production | ×10 | 1,500 | 0.935 | 0.897 | 0.882 |
| Fine-tune, seed 43 | houses + 20 | production | ×10 | 1,500 | 0.940 | 0.897 | 0.894 |

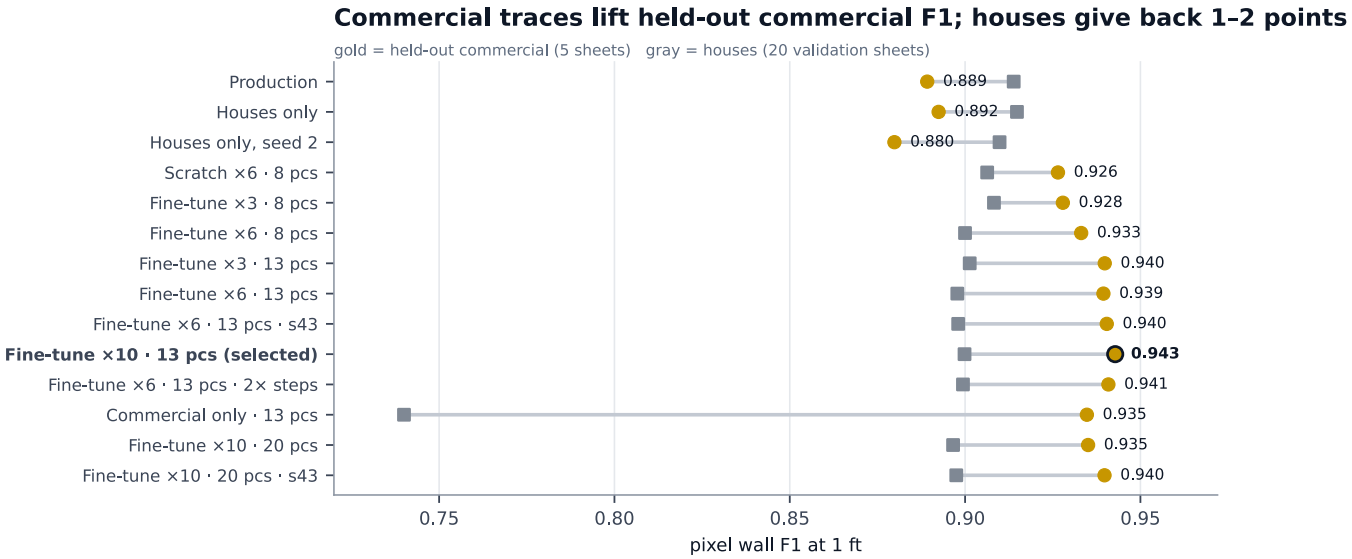


***Figure 7:*** *RoomNet ablations on identical held-out populations (Table 2). Every run that includes commercial traces gains on commercial sheets; every such run gives back 1–1.7 points on houses, and training on commercial sheets alone collapses house accuracy.*

### *5.2 Full pipeline*

Pixel F1 hides what a takeoff sees. With every stage except RoomNet left at production, found wall length on the held-out sheets rises from 77.4% to 93.0% and missed length falls by 69%, from 1,466 to 455 LF (Table 3, Figs. 8–9). Found length rises on all five sheets; the cardiology clinic moves from 69% to 94%. The cost is invented length, up from 447 to 746 LF, concentrated on casework, fixture outlines, grid stubs and dashed egress paths (Fig. 10). Figure 11 shows the mechanism: the production field fades on thin exterior walls and long corridor walls, and the commercial field does not.

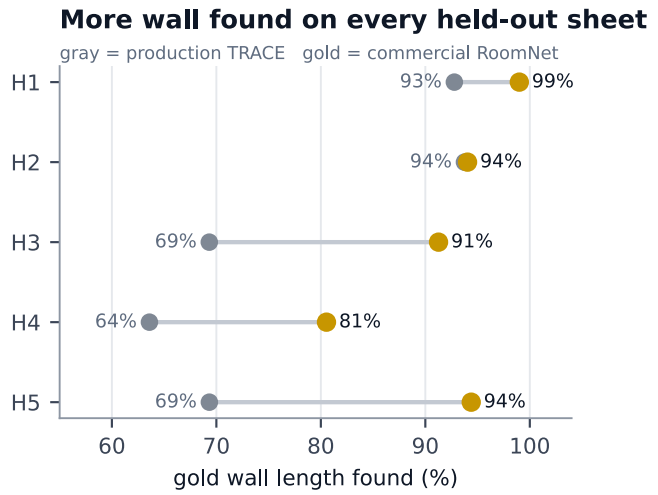


***Figure 8:*** *Gold wall length found per held-out sheet, full pipeline. Production (gray) versus the commercial RoomNet (gold); all other stages are production.*

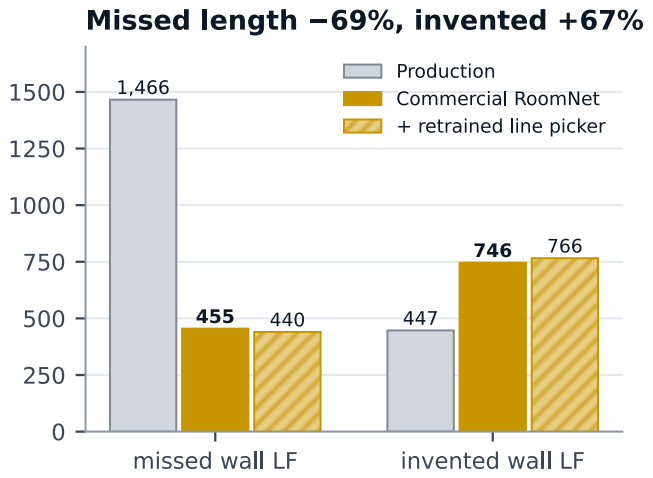


***Figure 9:*** *Missed and invented wall length summed over the five held-out sheets. The retrained line picker (hatched) leaves both essentially unchanged.*

***Table 3:*** *Full-pipeline wall length on held-out sheets (LF). Every column runs the complete production client and runtime; only the models differ.*

| Held-out sheet | Gold LF | Production: found / missed / invented | Commercial RoomNet: found / missed / invented | + line picker: found / missed / invented |
|---|---|---|---|---|
| H1 (urgent-care clinic) | 1,670 | 1,550 / 121 / 33 | 1,654 / 17 / 50 | 1,653 / 18 / 63 |
| H2 (winery) | 797 | 747 / 50 / 139 | 750 / 48 / 252 | 757 / 40 / 237 |
| H3 (restaurant) | 358 | 248 / 110 / 67 | 326 / 31 / 57 | 327 / 31 / 61 |
| H4 (angled plan) | 1,113 | 708 / 405 / 140 | 896 / 217 / 188 | 901 / 212 / 204 |
| H5 (cardiology clinic) | 2,544 | 1,764 / 780 / 68 | 2,401 / 143 / 199 | 2,404 / 140 / 201 |
| **Total** | **6,482** | **5,016 (77.4%) / 1,466 / 447** | **6,028 (93.0%) / 455 / 746** | **6,042 (93.2%) / 440 / 766** |

### *5.3 Exterior and interior*

Exterior/interior accuracy also improves overall, from 78.7% to 84.1% (Table 4, Fig. 12). Part of the gain is that the commercial model finds walls the production heads then classify correctly. Per sheet the result is mixed: two sheets (H1, H2) lose three to four points while the other three gain seven to twenty-four. In every column the dominant error is an interior partition called exterior, the natural bias of heads trained only on houses, where long straight walls near the outline are usually exterior.

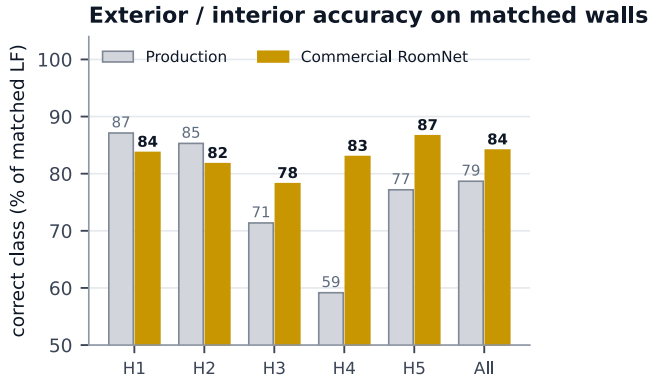


***Figure 12:*** *Exterior/interior accuracy per held-out sheet with production heads, before and after the RoomNet change.*

### *5.4 Line picker*

Table 5 and Fig. 13 score each line-picker variant on held-out tokens using the commercial field. Fine-tuning lowers commercial validity loss from 0.373 to 0.327 and keeps two to three points more real wall length at the production veto threshold; house loss also improves. Training from scratch is best at 1,250 steps and then overfits, ending worse than production on both populations by 8,000 steps. The pair head is the weak component on commercial sheets, separating same-wall from different-wall pairs only 60–67% of the time against about 76% on houses. The hard cases are walls within a foot of each other: demising walls, furring and chase walls.

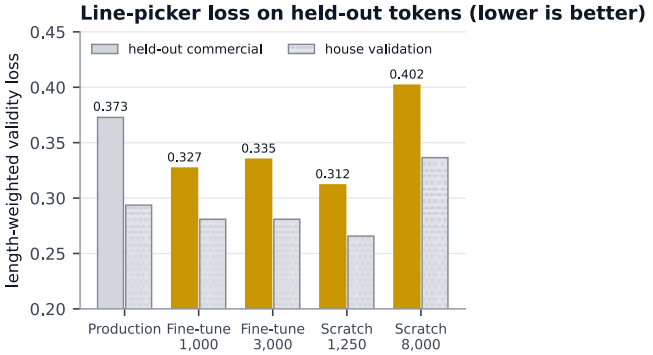


***Figure 13:*** *Line-picker validity loss on held-out tokens. Fine-tuning from production weights lowers commercial loss; training from scratch overfits by 8,000 steps.*

The token-level gain does not carry through. Run end to end, the fine-tuned line picker moves found length from 93.0% to 93.2%, missed length from 455 to 440 LF, invented length from 746 to 766 LF and exterior/interior accuracy from 84.1% to 84.4%, all inside the noise. The retrained model mainly raises its scores on real walls; its veto removes about as much non-wall length as before, and the walls the full pipeline invents pass the veto with either model.

***Table 4:*** *Exterior/interior accuracy over traced wall length within 1 ft of a gold wall; bearing walls count as interior.*

| Held-out sheet | Production: correct (LF compared) | Commercial RoomNet: correct (LF compared) | + line picker: correct (LF compared) |
|---|---|---|---|
| H1 (urgent-care clinic) | 87.1% (1,316) | 83.7% (1,445) | 83.6% (1,417) |
| H2 (winery) | 85.3% (975) | 81.7% (956) | 82.5% (947) |
| H3 (restaurant) | 71.4% (239) | 78.2% (330) | 76.4% (343) |
| H4 (angled plan) | 59.2% (686) | 83.0% (898) | 81.6% (907) |
| H5 (cardiology clinic) | 77.2% (1,627) | 86.6% (2,322) | 87.8% (2,357) |
| **All** | **78.7% (4,843)** | **84.1% (5,951)** | **84.4% (5,971)** |
| Interior called exterior (LF) | 937 | 755 | 752 |
| Exterior called interior (LF) | 95 | 190 | 181 |

***Table 5:*** *Line picker on held-out tokens (C = commercial, 5 sheets, 9,194 tokens; H = houses, 20 sheets, 27,675 tokens). Loss = length-weighted validity cross-entropy. Kept = real wall LF surviving the 0.30 veto; removed = non-wall LF vetoed; pair = balanced same-wall accuracy.*

| Line picker | C loss | C kept | C removed | C pair | H loss | H kept | H removed | H pair |
|---|---|---|---|---|---|---|---|---|
| Production line picker (new field) | 0.373 | 91.8% | 83.4% | 0.600 | 0.294 | 93.2% | 87.6% | 0.762 |
| Fine-tune ×3, lr 2e-4, step 1,000 | 0.327 | 94.2% | 82.0% | 0.641 | 0.281 | 95.3% | 85.5% | 0.773 |
| Fine-tune ×3, lr 2e-4, step 3,000 | 0.335 | 93.7% | 83.0% | 0.622 | 0.281 | 94.5% | 86.1% | 0.755 |
| Fine-tune ×1, lr 2e-4, step 3,000 | 0.335 | 93.6% | 82.6% | 0.625 | 0.281 | 94.3% | 86.8% | 0.761 |
| Fine-tune ×3, lr 5e-4, step 3,000 | 0.354 | 93.1% | 83.5% | 0.596 | 0.299 | 93.8% | 86.7% | 0.743 |
| Scratch ×3, step 1,250 | 0.312 | 94.7% | 81.7% | 0.673 | 0.266 | 95.2% | 86.2% | 0.764 |
| Scratch ×3, step 8,000 | 0.402 | 91.4% | 84.1% | 0.595 | 0.337 | 93.0% | 88.5% | 0.745 |

## 6. Discussion

The central result is that the bottleneck moved. With thirteen traced pieces the pixel model stopped improving on this test, and the full pipeline now finds 93% of commercial wall length. What remains is invented length and class errors. Because a better line picker did not remove the invented walls, either they are produced by stages it does not gate (cleanup passes, perimeter recovery, the 1/8 in. second pass), or the 0.30 veto is too permissive for commercial sheets. The next measurement should attribute each invented foot to the stage that produced it before any further training.

The house cost argues for two models. Every commercial mix gave back 1–1.7 points on houses, which remain most of the production traffic. The runtime should keep the house bundle and route commercial plans to a commercial bundle; the result cache is keyed on the wall-model file, so the two cannot share results.

Environment parity proved material. An earlier run of this comparison in a development environment with an older PDF library than the production worker showed exterior/interior accuracy falling from 83.3% to 67.1% on three sheets. With the worker's exact library version the same three sheets move from 87.1%, 85.3% and 71.4% to 83.7%, 81.7% and 78.2%. The class features read page text and drawings, and both change with the library version. Every result in this paper uses the matched environment.

## 7. Limitations and threats to validity

1. **Small commercial test set.** Five held-out sheets from five projects detect effects of several points but cannot rank close variants.
2. **Selection on the test set.** Line-picker checkpoints were chosen on the held-out tokens because five sheets leave no separate validation split; token-level numbers are optimistic.
3. **Bypassed gate.** The held-out runs use a lenient box. Today's production path would refuse two of the five sheets outright.
4. **One annotator.** Most pieces were traced by one person, so his conventions (centrelines through doors, what counts as a wall) are the model's conventions.
5. **Seed variance.** Seed-to-seed spread of about half a point on commercial F1 limits every RoomNet comparison below that size.


***Figure 10:*** *Held-out traces through the full pipeline. Top: H5, a cardiology clinic at 3/16", where production misses most of the exterior and many corridor walls. Bottom: H4, an angled plan, where found length rises from 708 to 896 of 1,113 LF. Orange marks invented walls, mostly casework and fixture outlines.*


***Figure 11:*** *The wall field itself on H5 (strongest direction per pixel). The production field fades on the thin exterior and the long corridor walls; those fades are the red spans in Figure 10.*

## 8. Reproducibility

The commercial gold, sheet list, RoomNet trainer and runners, full-pipeline harness, line-picker data builder and trainers, exterior/interior data builders, and the scripts that regenerate every figure and number in this paper are retained by Foreman AI Research, together with checkpoint hashes. The selected commercial RoomNet is round-4 run `R4_ftx10_s2`, step 1,500. Test runtimes are copies of

production runtime v16.2.4 with one model swapped and hash pins rewritten in the copy only. Nothing reported here has shipped; production still runs the house models.

## 9. Conclusion

Twenty traced commercial pieces were enough to adapt TRACE's wall model: on held-out commercial sheets the full pipeline now finds 93.0% of wall length instead of 77.4% and misses 455 LF instead of 1,466. The adaptation costs houses 1–2 pixel-F1 points, so the commercial model should be deployed alongside the house model rather than in place of it. The work also shows where effort should go next. More traced sheets no longer move the wall model, and a retrained line picker does not reduce invented length. The next gains will come from finding which stage admits invented walls, retraining the exterior/interior heads on commercial predictions, and fixing the sheet gate so commercial plans reach the model at all.

## Data availability

The plan sheets come from Foreman AI's proprietary corpus and are not redistributed; sheets are identified by neutral IDs. Aggregate results, per-sheet scores and checkpoint hashes are retained by Foreman AI Research.

## Competing interests

Kyle Rossignol is the founder of Foreman AI LLC, which develops commercial construction-estimating software incorporating the wall-recovery technology described here. This manuscript reports internal research and should be interpreted with that relationship disclosed.

## Funding

The work was conducted as internal research and development by Foreman AI LLC. No external funding was received.

## Author contributions

K.R. conceived the project, traced the commercial gold, and directed the experiments. Model training, evaluation and drafting of this manuscript were carried out with Claude (Anthropic), an AI model, under K.R.'s direction.

## References

1. K. Rossignol, “TRACE Residential: Candidate-Then-Verify Wall Recovery from Vector Construction Drawings,” Foreman AI Research, TRACE Series Report 1, Colorado, USA, Sep. 2026.
2. K. Rossignol, “Measurement Practice for Automated Wall Recovery from Architectural Permit Drawings,” Foreman AI Research, Colorado, USA, technical report, Aug. 2026.
3. O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, 2015.
4. I. Loshchilov and F. Hutter, “Decoupled Weight Decay Regularization,” in *International Conference on Learning Representations (ICLR)*, 2019.
5. M. F. Goodchild and G. J. Hunter, “A simple positional accuracy measure for linear features,” *International Journal of Geographical Information Science*, vol. 11, no. 3, pp. 299–306, 1997.
6. C. Liu, J. Wu, P. Kohli, and Y. Furukawa, “Raster-To-Vector: Revisiting Floorplan Transformation,” in *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 2195–2203, 2017.
7. A. Kalervo, J. Ylioinas, M. Häikiö, A. Karhu, and J. Kannala, “CubiCasa5K: A Dataset and an Improved Multi-Task Model for Floorplan Image Analysis,” arXiv:1904.01920, 2019.
8. W. Song, M. M. Abyaneh, M. A. A. Shabani, and Y. Furukawa, “Vectorizing Building Blueprints,” in *Proc. Asian Conference on Computer Vision (ACCV)*, pp. 1044–1059, 2022.
9. J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?,” in *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3320–3328, 2014.
10. J. Kirkpatrick et al., “Overcoming catastrophic forgetting in neural networks,” *Proceedings of the National Academy of Sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.