Iterating on the S23DR 2026 Baseline

Anatolii Shokha
Lviv Polytechnic National University
anatolii.shokha.mnitm.2025@lpnu.ua

Abstract

We describe our 5th-place entry to the S23DR 2026 challenge for 3D building wireframe reconstruction. We iterate on the organizers' baselines and report what each step contributes. We push a hand-crafted geometric pipeline to 0.4095 HSS, hybridize it with the learned baseline through vertex snapping and classifier-gated augmentation to reach 0.4997, and extend the learned model's resolution curriculum from 4k to 8k input points to reach 0.5009 public and 0.5388 private HSS with a single 8.85M-parameter model. Our main finding is that the value of the hand-crafted priors falls as the learned model improves: the snapping machinery worth $+0.045$ *HSS at 4k adds only* $+0.0005$ *at 8k. We also report negative results on scaling beyond 8k and on a stronger but slower point-transformer encoder.*

1. Introduction

The S23DR 2026 challenge [7] asks for a metric 3D building wireframe—vertices and edges—given only Structure-from-Motion (SfM) points [4], scale-fitted depth, and projected semantic segmentations, with no RGB images. Predictions are scored by the Hybrid Structure Score (HSS) [7], the harmonic mean of a vertex F1 term and an edge IoU term.

We iterate on the provided baselines and document what each step produces, working through three regimes in sequence. We first push a hand-crafted geometric pipeline, building on the 2025 second-place solution of Jurić [2], to 0.4095 HSS on the public split (Sec. 2). We then bridge this pipeline to the organizers' learned segment model through a geometric "seam" and two classifier-gated augmentation steps, reaching 0.4997 (Sec. 3). Finally, we extend the learned model's resolution curriculum from 4k to 8k input points, reaching 0.5009 public and 0.5388 private HSS (5th place) with a single 8.85M-parameter model (Sec. 4).

Our main finding is that the value of the hand-crafted priors falls as the learned model improves: the augmentation machinery worth $+0.045$ HSS at 4k adds only $+0.0005$ at 8k. We also report negative results from pushing past 8k (Sec. 5): longer context regressed, a latent-capacity experiment ruled out the obvious explanation, a per-scene router could not beat its own oracle, and a Point Transformer V3 encoder [8] was too slow for the two-hour T4 evaluation budget.

2. Maxing Out the Hand-Crafted Pipeline

The organizers provide a minimal hand-crafted baseline that, per view, finds vertices and line-like features as connected components in the Gestalt segmentation, connects nearby vertices into edges, lifts them to 3D using the fitted depth, and merges across views by proximity. On the public split it scores 0.3559 HSS (Tab. 1).

Base pipeline. We build on the 2025 second-place solution of Jurić [2], which adds a multi-view consistency filter that prunes 3D vertices by re-projecting them into all overlapping views and checking semantic agreement, an edge-recovery step that reconnects the surviving vertices, support for additional Gestalt vertex and edge classes, and robust handling of variable image orientations. Ported to the 2026 data, it scores 0.3771.

Our refinements. We add geometric proposal and validation steps, tuning each against the public split. Straight roof structures are detected with a probabilistic Hough transform (`HoughLinesP`); for each detected segment we order the nearby candidate vertices along it and connect consecutive ones, avoiding the spurious cross-connections of naive proximity linking. Together with a connectivity rulebook this reaches 0.3879, and tuning the 3D merge threshold a further 0.4019. A line-of-sight edge check (`verify_edge_mask`) rasterizes each candidate edge, intersects it with the semantic mask, and discards edges that cross empty space ($+0.006$). Overlap tuning brings the pipeline to its peak of 0.4095. We also tried and discarded snapping near-90° corners to exact right angles and a COLMAP-point DBSCAN vertex detector; neither helped on the public split.

At 0.4095 the hand-crafted pipeline exceeds the organizer baseline by $+0.054$. This is the ceiling we reach with

Configuration	HSS	Δ
Organizer hand-crafted baseline	0.3559	—
+ Jurić 2025 pipeline [2]	0.3771	+0.021
+ Hough detection & connectivity	0.3879	+0.011
+ 3D merge-threshold tuning	0.4019	+0.014
+ line-of-sight edge check	0.4079	+0.006
+ overlap tuning	**0.4095**	+0.002

Table 1. Hand-crafted pipeline on the public split, with Δ relative to the row above. Adopting Jurić's 2025 pipeline [2] is the base; our refinements add a further +0.032, peaking at 0.4095 before any learning.

geometry and rules alone; the rest of the report adds the learned model.

3. Hybridizing with the Learned Model

The organizers also release a *learned* baseline: a Perceiver-style encoder [1] over a fused, priority-sampled point cloud that predicts wireframe edges as oriented 3D segments. At its released 4k configuration it scores 0.4470 public HSS, above our hand-crafted pipeline (0.4095). The two make different errors—the learned model drifts on long edges, while the hand-crafted pipeline recovers some vertices the learned model misses—so we combine them, in two ways (Fig. 1).

The seam. Our first hybrid step (`snap_to_handcrafted`) moves each learned vertex toward the nearest hand-crafted vertex within a radius, blending positions with a fixed weight ($\alpha = 0.7$ toward the hand-crafted vertex). A median-distance guard disables the snap entirely when the hand-crafted vertices appear to be in a misaligned coordinate frame, preventing corruption on bad COLMAP reconstructions. Tuning the snap radius and blend weight, the seam lifted the learned baseline from 0.4470 to 0.4916 HSS (+0.045; Tab. 2).

Classifier-gated augmentation. We then adopted the central idea of the 2025 winning solution [5]: small PointNet [3] classifiers that score local 3D patches around candidate vertices and edges. Skvrna and Neumann used these as a *standalone* pipeline—generate 3D candidates, classify them, and emit the survivors as the wireframe. We use them as *augmentation gates*: each classifier scores the hand-crafted candidates, and only high-confidence ones are imported into the learned model's output, never replacing it. The candidates are the hand-crafted pipeline's own predicted edges and vertices (`predict_wireframe`, Sec. 2)—the same distribution scored at deployment—each drawn from roughly 13k training scenes. Edge candidates are labelled by their leave-one-out effect on HSS (whether removing the edge lowers the score); vertex candidates by proximity to a ground-truth vertex.

Configuration	HSS	Δ
Learned baseline (4k, no hybrid)	0.4470	—
+ seam (snap-to-HC, α=0.7)	0.4916	+0.045
+ edge augment (gated)	0.4984	+0.0068
+ vertex augment (gated)	**0.4997**	+0.0013

Table 2. Hybrid era on the public split, Δ relative to the row above. The classifier idea and architecture follow the 2025 winning solution [5]; we re-purpose them as augmentation gates.

We use a 6D (xyz+rgb) edge classifier on cylindrical patches and a dual-head vertex classifier on cubic patches, both architecturally close to the 2025 networks (val AUC 0.75 and 0.73). The edge augment lifted the seam from 0.4916 to 0.4984 (+0.0068); the vertex augment, which imports high-confidence hand-crafted vertices as orphan nodes, reached 0.4997 (+0.0013)—our best hybrid configuration.

Validation AUC versus deployment. A third classifier shows that validation AUC does not predict deployment value. We trained a gradient-boosted classifier to promote extra edge endpoints—a *different* candidate source, the endpoints of detected Hough segments, meant to fill the hand-crafted pipeline's recall gaps—labelling an endpoint positive if it lay within 0.5 m of any ground-truth vertex. It reached the highest validation AUC of the three (0.82) yet *lowered* the public score by 0.003. Because its confident positives sit near vertices the hand-crafted pipeline already proposes, they were dedup-rejected at import and added nothing, while its rarer errors hurt. A high score on the proxy objective (generic vertex-nearness) did not transfer because the deployment objective (adding vertices the pipeline is *missing*) was different. We also dropped an attempt to import edge-class "junction" candidates as orphans, which collapsed the bottom-quantile scores. Neither is part of the reported system.

4. Scaling the Learned Model

The hybrid of Sec. 3 reached 0.4997 around the *4k* learned baseline. We now scale the learned model itself.

Curriculum. The model is a Perceiver [1] with hidden width 256, 256 latent tokens, 7 latent layers and 64 segment queries—8.85M parameters, unchanged throughout. The organizers train it via a staged resolution curriculum—from scratch on 2048-point samples (125k steps), then fine-tuned at 4096 points—reaching 0.4470 at 4k (Tab. 3); training directly at high resolution overfits. We continue this curriculum to 8192 points, fine-tuning the released 4k checkpoint

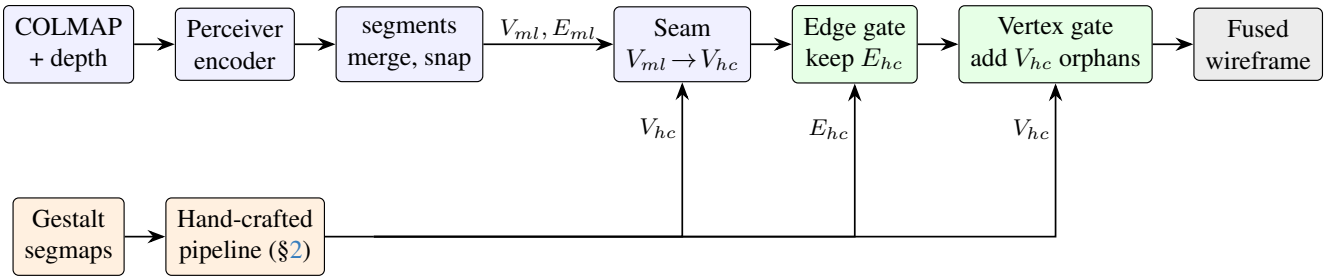

Figure 1. Hybrid pipeline. The learned branch (blue) encodes the fused COLMAP/depth cloud with a Perceiver and decodes segments into vertices and edges (V_{ml}, E_{ml}). The hand-crafted branch (orange, §2) produces (V_{hc}, E_{hc}) from the Gestalt segmentations. The seam snaps each learned vertex toward the nearest hand-crafted vertex; the edge and vertex gates (green) admit only high-confidence hand-crafted edges and vertices, the latter as orphans.

for a further 45k steps (the last 20k a linear cooldown) on the organizers' released 8k samples [6], at a gentle learning rate (3×10^{-5}) that preserves the lower-resolution representations. Each resolution step raises the input's structural coverage: roughly 66% at 2048, 74% at 4096, and higher at 8192.

Scaling to 8k. Doubling the input from 4096 to 8192 points, with 45k additional fine-tuning steps, lifted the *raw* learned model from 0.4470 to 0.5004 public HSS—a +0.053 jump, larger than the entire hand-crafted hybrid contributed at 4k. A further retuning of the confidence threshold ($0.5 \to 0.65$) added a small amount, reaching our public best of 0.5009. At this point the model alone, with no hand-crafted input, is already stronger than the best 4k hybrid.

Decay of the hand-crafted prior. At 4k, the seam was worth +0.045 and the classifier-gated augments a further +0.0081; the hand-crafted machinery accounted for most of the system's accuracy. At 8k, the *entire* hybrid—seam plus both augments—improved the raw model by only +0.0005 (Tab. 3, last two rows). The explanation is geometric: the seam and augments pull the model's vertices toward hand-crafted ones, which helps only while the model's vertices are less accurate than the hand-crafted candidates. At 8k the model's predictions are typically the more accurate of the two, so snapping toward the hand-crafted vertices moves them the wrong way as often as the right way. The mechanism that added +0.045 at 4k is, at 8k, indistinguishable from noise.

Our central result is that the value of a hand-crafted prior is not fixed but decreases with the learned model's quality: a hybrid that looks essential at one scale can be redundant, or mildly harmful (Sec. 3), at the next. This motivates our final submission (Sec. 6): at 8k we prefer the raw model, which does not depend on components that no longer help.

Stage	Context	Cum. steps	Public HSS
From scratch	2048	125k	0.4273
Fine-tune + cooldown	4096	170k	0.4470
Fine-tune + cooldown	8192	215k	**0.5004**
Hand-crafted hybrid contribution, by model scale:			
seam + augments @ 4k	4096	—	+0.053
full hybrid @ 8k	8192	—	+0.0005

Table 3. Resolution curriculum (top) and the decay of the hand-crafted prior with model scale (bottom). Each resolution step adds 45k fine-tuning steps. The $4k \to 8k$ gain (+0.053) exceeds the full hand-crafted hybrid at 4k, which itself adds only +0.0005 at 8k.

5. Probing the Ceiling: Negative Results

Scaling to 8k worked (Sec. 4); the attempts to push further did not. We report them because their failure modes are informative.

16k context. We doubled the input again, to 16384 points; as the organizers released samples only up to 8k, we generated the 16k samples ourselves from their cached full point-cloud release [6], using the same priority-sampling procedure. Fine-tuning with the same curriculum did not continue the trend: the raw 16k model scored 0.4485 public HSS, well below the raw 8k model's 0.5004 (Tab. 4). Our 170-scene local validation did not predict this—there the 16k model scored about the same as the 8k one (0.36 vs 0.357). The regression appeared only on the public split, a reminder that small validation sets can hide failures the test distribution exposes. (Consistent with Sec. 4, the hand-crafted hybrid recovered much of the loss on the weaker 16k model, lifting it back to 0.498—the prior helps again precisely because the model got worse.)

Latent capacity. Our first hypothesis was that the Perceiver's fixed bank of 256 latent tokens had become a bot-

tleneck: at 16k it must compress twice as many input points through the same latent budget. We tested this directly by widening the latent bank to 512 tokens—copying the trained latents and randomly initializing the new half—and fine-tuning at 16k. The wider model showed no improvement (0.355 local, versus 0.357 for the 256-token model at the same budget), ruling out latent capacity: whatever causes the 16k regression, it is not the number of latent tokens.

Point Transformer V3 encoder. To remove the latent bottleneck entirely we replaced the Perceiver encoder with a Point Transformer V3 [8], which maintains per-point features through serialized sparse attention rather than pooling into a fixed latent set. Trained from scratch at 8k for 200k steps, it plateaued at 0.323 local HSS—below the 0.357 of the curriculum-trained Perceiver. With far more training it might close this gap, but two factors made it impractical. First, from-scratch training did not reach Perceiver quality within our compute budget; the Perceiver benefited from its 2k→8k curriculum, which the sparse-convolution encoder does not directly inherit. Second, inference cost: the Point Transformer V3 pipeline ran at roughly 6 s per sample on an A5000, while the evaluation environment imposes a two-hour budget on a slower T4—which this pipeline would exceed. Deployment latency is a hard constraint here, and a more powerful per-point encoder is not automatically a better *submission*.

Per-scene router. A different way to push past the single-model ceiling is a mixture-of-experts view: for each scene, route to whichever source—the 4k pipeline, the 8k pipeline, or the hand-crafted prediction—would have scored best, and additionally gate whether to apply the seam. We built this as a two-stage gradient-boosted router (a 3-class source selector followed by a binary seam gate), trained on a per-scene dataset of cheap, zero-extra-cost features: hand-crafted and COLMAP geometry statistics available before any inference, plus the model's own confidence summaries available after the forward pass that runs anyway. The per-scene oracle—always picking the best of the three sources—is worth only $+0.0197$ HSS over the deployed hybrid, and even that margin is unreachable. The best single predictor we found (the 8k model's mean kept confidence) correlates with the hand-crafted source winning at only $r = -0.44$; under five-fold cross-validation the router recovered just $+0.0009$ of the $+0.0197$ ceiling (4.5%), and the seam gate sat at the majority-class accuracy of 67%. The per-scene winner is not predictable from features that are free at inference time, so we never submitted the router. We release the dataset and scripts for completeness.

These negatives mark the ceiling of our approach: the 8k Perceiver remained our system. The next section describes what we submitted.

Attempt	Context	Local	Public	Outcome
Perceiver (ours)	8192	0.357	0.5004	system
Perceiver	16384	0.360	0.4485	regressed
Perceiver, 512 latents	16384	0.355	—	no gain
Point Transf. V3	8192	0.323	—	slow / undertr.

Table 4. Attempts to push past the 8k Perceiver. Local is mean HSS on 170 validation scenes; public is the leaderboard score where a successful submission exists (the latent-512 and PT v3 runs were never submitted).

6. Final Submission and Results

What we submitted. The challenge allows two submissions to count toward the private leaderboard. Our highest public score was the 8k hybrid at 0.5009, but Sec. 4 gives a concrete reason to distrust the hybrid on the held-out split: at 8k it contributes only $+0.0005$, and that contribution flows through two PointNet classifiers trained on roughly thirteen thousand scenes, which could be tuned to the public distribution rather than to the task. We therefore selected, as our primary entry, the *raw* 8k model with no hand-crafted post-processing (0.5004 public), trading a negligible 0.0005 of public score for independence from those components. As a complementary second entry we kept the earlier 4k hybrid, whose much stronger bottom-quantile behaviour ($q_5 = 0.086$ versus 0.015 for the raw 8k model) hedges against distribution shift on hard scenes. The two entries are intentionally different: one classifier-free and accurate on typical scenes, one classifier-assisted and robust in the tail.

Result. On the private leaderboard our raw 8k model scored 0.5388 HSS, placing 5th overall after team merging (Tab. 5). Every team gained 0.03–0.05 between the public and private splits, so the ranking was largely preserved; the raw-model choice neither helped nor hurt relative to the hybrid, consistent with the decayed hybrid contribution of Sec. 4. The result sits above the organizers' learned (0.474) and hand-crafted (0.391) baselines, with a single 8.85M-parameter model—one to two orders of magnitude smaller than the multi-stage, multi-model pipelines of the top entries.

7. Conclusion

We studied one baseline across three regimes. Beginning from the 2025 second-place hand-crafted pipeline [2], we pushed pure geometry to 0.4095 HSS; bridged it to the organizers' learned model with a snapping seam and classifier-gated augments adapted from the 2025 winning solution [5],

Rank	Team	Private HSS
1	VRG	0.6542
2	LundUni	0.6476
3	StarAtNyte1	0.6060
4	kkcc	0.5743
5	**Ours**	**0.5388**
–	Organizer learned baseline	0.4739
–	Organizer hand-crafted baseline	0.3907

Table 5. Final private leaderboard down to our entry (top teams, after team merging) and the organizer baselines. Our single 8.85M-parameter model places 5th.

reaching 0.4997; and scaled the learned model along a resolution curriculum to 0.5009 public (0.5388 private, 5th place).

Across the three regimes, the value of a hand-crafted prior decreases with the learned model's quality. The same snapping machinery that lifted the 4k model by $+0.045$ added only $+0.0005$ to the 8k model, and a per-scene oracle over all sources offered a ceiling of just $+0.0197$ that no cheaply-gated router could capture. Pushing past 8k stopped us: context beyond 8k regressed for reasons not attributable to latent capacity, and a per-point Point Transformer V3 encoder neither reached Perceiver quality in our training budget nor met the two-hour T4 inference limit. We hope the scaling curriculum and the negative results are useful to others building on these baselines.

References

[1] Andrew Jaegle, Felix Gimeno, Andy Brock, Andrew Zisserman, Oriol Vinyals, and Jo ao Carreira. Perceiver: General perception with iterative attention. 2021.

[2] Antonio Jurić. Exploring what can be done with a hand-crafted 3d reconstruction pipeline. Technical report, University of Zagreb, Faculty of Electrical Engineering and Computing, 2025. S23DR 2025 Challenge, 2nd place; CVPR 2025 Workshop on Urban Scene Modeling.

[3] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, pages 652–660, 2017.

[4] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, pages 4104–4113, 2016.

[5] Jan Skvrna and Lukas Neumann. Structured semantic 3d reconstruction (s23dr) challenge 2025 – winning solution. In *CVPR Workshop on Urban Scene Modeling*, 2025.

[6] USM3D. S23DR 2026 datasets: `hoho22k_2026_trainval` and the pre-sampled / cached point-cloud releases. https://huggingface.co/usm3d, 2026. Separate releases under the `usm3d` org: `sampled_{2048,4096,8192}` and `cached_full_pcd`, derived from `hoho22k_2026_trainval`.

[7] USM3D. Structured 3d reconstruction challenge 2026. https://huggingface.co/spaces/usm3d/S23DR2026, 2026.

[8] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer V3: Simpler, faster, stronger. In *CVPR*, 2024.