What’s Under the Skin? Estimating Swine Body Condition

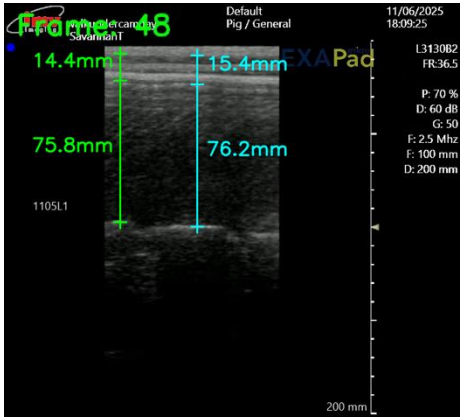
Automatic Pig Body Condition Scoring — Paper Presentation

Mk Bashar

May 28, 2026

Body Condition Score Method

- Input:
  - Raw single depth frames
    - ✓ pig walks under the ceiling-mounted camera
    - ✓ train on single random frames; average all frames at inference
- Output:
  - Backfat depth at last rib
  - Loin depth at last rib
  - Total tissue thickness (fat + loin)
    - ✓ Ground truth: Ultrasound Scan



Problem & Motivation

Body condition is critical for sow management

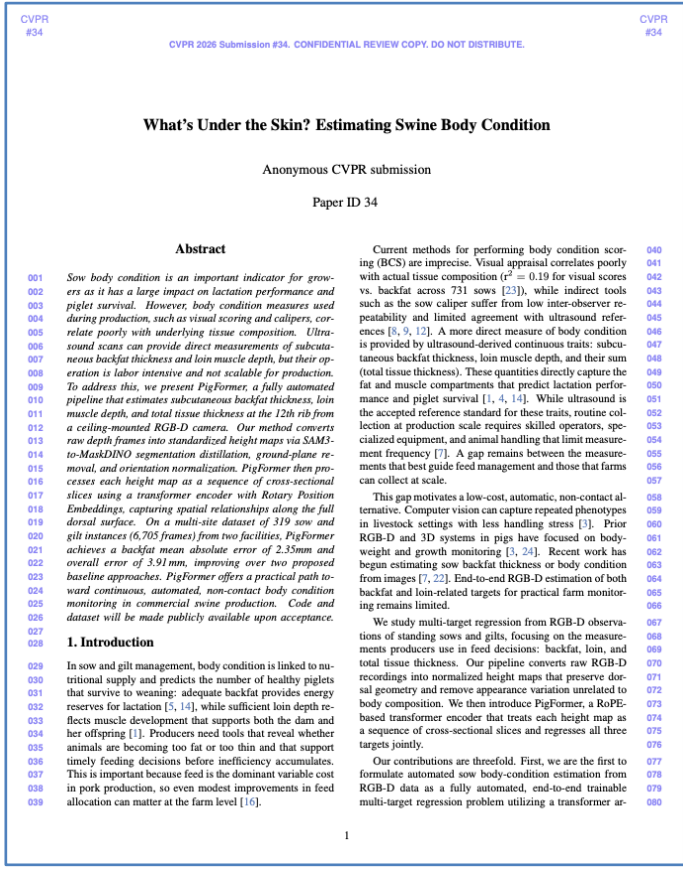
- Backfat → energy reserves for lactation & piglet survival
- Loin depth → muscle development supporting dam & offspring
- Feed is the dominant variable cost in pork production

Current methods are imprecise

- Visual scoring: $r^2 = 0.19$ vs. actual backfat (731 sows)[*]
- Sow caliper: low inter-observer repeatability
- Caliper explains only 56% of backfat, 44% of loin variance

Ultrasound is accurate but not scalable

- Requires skilled operators & specialized equipment
- Labor-intensive animal handling limits measurement frequency



CVPR #34 CVPR 2026 Submission #34. CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE. CVPR #34

What's Under the Skin? Estimating Swine Body Condition

Anonymous CVPR submission

Paper ID 34

Abstract

Sow body condition is an important indicator for growers as it has a large impact on lactation performance and piglet survival. However, body condition measures used during production, such as visual scoring and calipers, correlate poorly with underlying tissue composition. Ultrasound scans can provide direct measurements of subcutaneous backfat thickness and loin muscle depth, but their operation is labor intensive and not scalable for production. To address this, we present PigFormer, a fully automated pipeline that estimates subcutaneous backfat thickness, loin muscle depth, and total tissue thickness at the 12th rib from a ceiling-mounted RGB-D camera. Our method converts raw depth frames into standardized height maps via SAM3-to-MaskDINO segmentation distillation, ground-plane removal, and orientation normalization. PigFormer then processes each height map as a sequence of cross-sectional slices using a transformer encoder with Rotary Position Embeddings, capturing spatial relationships along the full dorsal surface. On a multi-site dataset of 319 sow and gilt instances (6,705 frames) from two facilities, PigFormer achieves a backfat mean absolute error of 2.35mm and overall error of 3.91mm, improving over two proposed baseline approaches. PigFormer offers a practical path toward continuous, automated, non-contact body condition monitoring in commercial swine production. Code and dataset will be made publicly available upon acceptance.

1. Introduction

In sow and gilt management, body condition is linked to nutritional supply and predicts the number of healthy piglets that survive to weaning: adequate backfat provides energy reserves for lactation [5, 14], while sufficient loin depth reflects muscle development that supports both the dam and her offspring [1]. Producers need tools that reveal whether animals are becoming too fat or too thin and that support timely feeding decisions before inefficiency accumulates. This is important because feed is the dominant variable cost in pork production, so even modest improvements in feed allocation can matter at the farm level [16].

Current methods for performing body condition scoring (BCS) are imprecise. Visual appraisal correlates poorly with actual tissue composition ($r^2 = 0.19$ for visual scores vs. backfat across 731 sows [23]), while indirect tools such as the sow caliper suffer from low inter-observer repeatability and limited agreement with ultrasound references [8, 9, 12]. A more direct measure of body condition is provided by ultrasound-derived continuous traits: subcutaneous backfat thickness, loin muscle depth, and their sum (total tissue thickness). These quantities directly capture the fat and muscle compartments that predict lactation performance and piglet survival [1, 4, 14]. While ultrasound is the accepted reference standard for these traits, routine collection at production scale requires skilled operators, specialized equipment, and animal handling that limit measurement frequency [7]. A gap remains between the measurements that best guide feed management and those that farms can collect at scale.

This gap motivates a low-cost, automatic, non-contact alternative. Computer vision can capture repeated phenotypes in livestock settings with less handling stress [3]. Prior RGB-D and 3D systems in pigs have focused on bodyweight and growth monitoring [3, 24]. Recent work has begun estimating sow backfat thickness or body condition from images [7, 22]. End-to-end RGB-D estimation of both backfat and loin-related targets for practical farm monitoring remains limited.

We study multi-target regression from RGB-D observations of standing sows and gilts, focusing on the measurements producers use in feed decisions: backfat, loin, and total tissue thickness. Our pipeline converts raw RGB-D recordings into normalized height maps that preserve dorsal geometry and remove appearance variation unrelated to body composition. We then introduce PigFormer, a RoPE-based transformer encoder that treats each height map as a sequence of cross-sectional slices and regresses all three targets jointly.

Our contributions are threefold. First, we are the first to formulate automated sow body-condition estimation from RGB-D data as a fully automated, end-to-end trainable multi-target regression problem utilizing a transformer ar-

1

[*] Young, M.G., Tokach, M.D., Goodband, R.D., Nelssen, J.L., Dritz, S.S.: The relationship between body condition score and backfat in gestating sows. In: Kansas State University Swine Day Report (2001). https://newprairiepress.org/kaesrr/vol0/iss10/866/

Research Objective

Goal

- Fully automated, non-contact estimation of sow body condition from a ceiling-mounted RGB-D camera
- Three regression targets at the last rib: backfat, loin depth, total tissue thickness (mm)

Contributions: PigFormer

- **1.** First to formulate sow body condition as fully automated, end-to-end, multi-target regression on production-relevant measurements (backfat, loin, total)
- **2.** Two-stage PigFormer: Stage 1 geometric front-end (SAM3→MaskDINO distillation, ground-plane removal, orientation normalization) + Stage 2 Slice Attention Encoder (RoPE + dual pooling)
- **3.** 3.87 mm overall MAE on 319 instances — beats single-stage ResNet-18 by 22% and ViT-small by 39% (Stage 1 geometry adds info on top of pretrained backbones); UNet Stage-1 variant runs real-time (~7 ms/frame)

Data Collection

Two U.S. facilities

- MSU: 116 instances, 6 scan dates (Feb–Dec 2025); mean fat 17.93 mm, loin 54.52 mm
- USMARC / UNL: 203 instances, 11 scan dates (Jun–Dec 2025); mean fat 8.86 mm, loin 40.94 mm
  - *large per-site label gap → combined-site training matters*

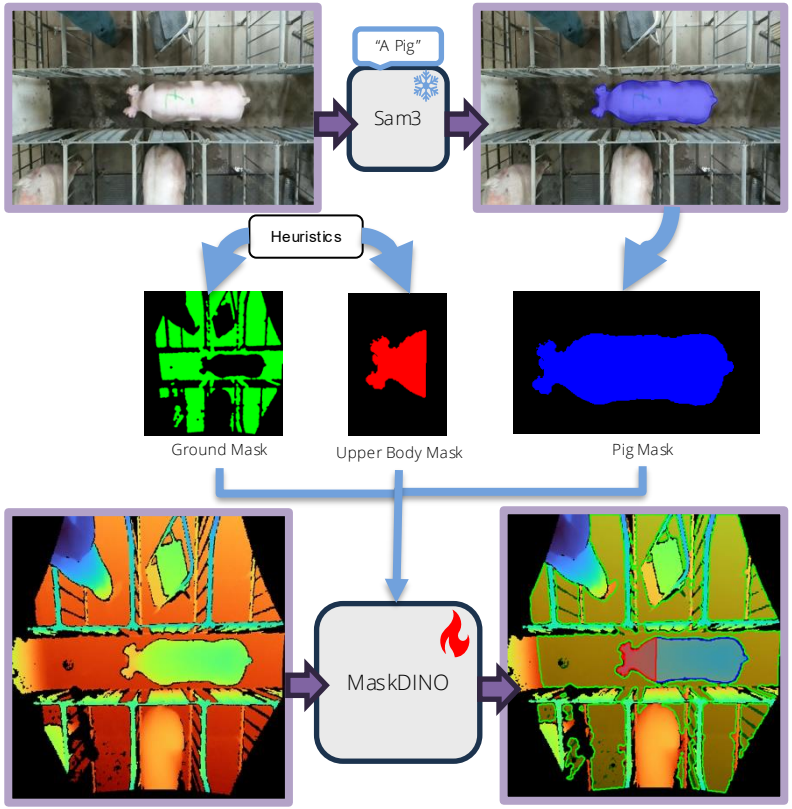
Dataset summary

- 319 total sow/gilt instances; 6,705 depth frames (ceiling-mounted Orbbec cameras)
- Each scan date treated as independent instance (body composition changes)

Ground truth

- B-mode ultrasound at last rib: backfat, loin depth, total tissue thickness
- Identity-level separation: test animals never seen during training

Stage 1: Geometric Front-End — Segmentation



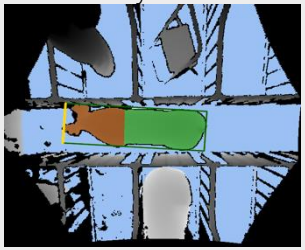
Pipeline

Raw Depth Map

640x576

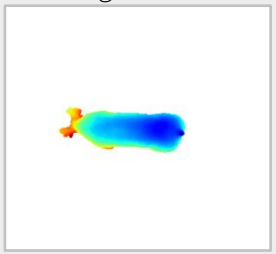
Mask DINO

Ground, Pig, Upper Body Mask



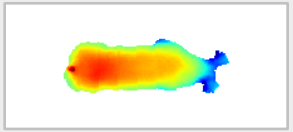
640x576

Depth Map of Pig Segmented



640x576

Height Map Rotated, Centered, Cropped

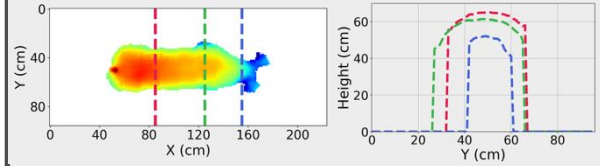


224x96

Stage 1: Geometric Front-End

Stage 2: Slice Attention Encoder

Tokens are **slices**, not patches

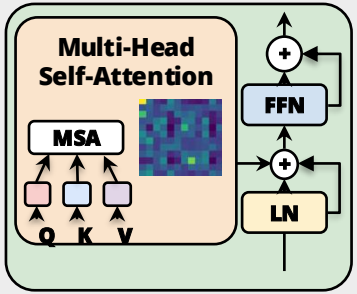


224x96

224 tokens

PigFormer Blocks

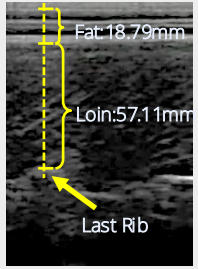
PigFormer Block



Regression Head

FC FC ... FC

Output: Fat & Loin Depth



Ultrasound Scan

Stage 2: Slice Attention Encoder

Sequence representation

- 96 × 224 height map → 224 column tokens (one per cross-sectional slice)
- Column-based tokenization outperforms patch-based approaches

Rotary Position Embedding (RoPE)

- Encodes relative distance between cross-sections in attention
- Critical for rib-site localization along the spine

Encoder + Dual Pooling

- Multi-head self-attention → FFN (GELU) with LayerNorm + dropout
- Mean pooling (overall condition) + Max pooling (anatomical features)

Training

- Huber loss on backfat, loin, and total tissue depth independently
- Single random frame per animal per epoch (natural augmentation)
- Inference: average predictions across all frames per animal

Baselines: Isolating Stage 1's Contribution

Main baselines are **single-stage** models fed **raw 576 × 640 depth** — they skip Stage 1 entirely, so the gap to PigFormer measures the value of the geometric front-end.

ResNet-18 (single-stage)

- ImageNet-pretrained CNN; conv1 surgery for 1-channel depth input
- z-score normalization only — no segmentation, ground-plane removal, or orientation normalization

ViT-small (single-stage)

- ImageNet-pretrained transformer; patch-embed surgery for 1-channel depth; same z-score-only input
- *CNN and MLP height-map encoders move to the supplement as Stage-2 architecture ablations.*

Main Results

Method	Fat ↓	Loin ↓	Total ↓	Overall ↓	Stage-1 ms
ViT-small (single-stage)	3.57	7.29	8.16	6.34	—
ResNet-18 (single-stage)	2.88	6.10	5.81	4.93	—
PigFormer w. MaskDINO	**2.43**	**5.01**	**4.19**	**3.87**	106.9
PigFormer w. Pruned MaskDINO	**2.34**	5.27	4.20	3.94	52.7
PigFormer w. UNet	2.40	5.20	4.26	3.95	**6.6**
Human Ultrasound Std	1.30	2.02	2.29	1.87	—

Key: PigFormer cuts error **22% vs. ResNet-18 and 39% vs. ViT-small** — Stage 1 geometry adds information on top of pretrained backbones. Fat (2.43 mm) is within ~2 × human ultrasound variability (1.30 mm). The UNet Stage-1 variant gives near-identical accuracy at ~6.6 ms (real-time).

Stage 1: Speed–Accuracy Trade-off

The segmentation front-end can be swapped without losing accuracy

Three front-ends, near-identical overall MAE:

- **Mask DINO** — best accuracy (3.87 mm) but slowest at 106.9 ms/frame
- **Pruned Mask DINO** — 3.94 mm at 52.7 ms/frame (2× faster)
- **UNet** — 3.95 mm at just 6.6 ms/frame — **16× faster than Mask DINO**

Efficiency of the UNet front-end:

- 2.2M parameters, 4.06 GFLOPs — lightweight enough for edge hardware

→ *End-to-end ~7 ms/frame on an A100: real-time body-condition estimation at production scale*

Ablation Studies

Stage-2 encoder architecture

- Slice-attention encoder wins (3.91 mm) vs. MLP (4.11) and CNN (6.02) on the same Stage-1 input
- Depth ablation: a single attention layer is best (4.00 mm); deeper stacks overfit

Multi-frame aggregation

- Averaging 3 frames is best (3.79 mm); all frames degrades to 5.52 mm. Pixel-level beats feature-level (4.12 vs. 4.67 mm)

Data augmentation

- Vertical flip helps (4.07 vs. 4.22 mm); horizontal flip hurts (4.86 mm) — head-to-tail orientation encodes spatial info

Training data composition

- Combined-site training nearly halves overall MAE vs. single-site; neither site alone generalizes to the mixed test set

Caliper vs. Ultrasound

Sow caliper scores are a poor proxy for tissue composition

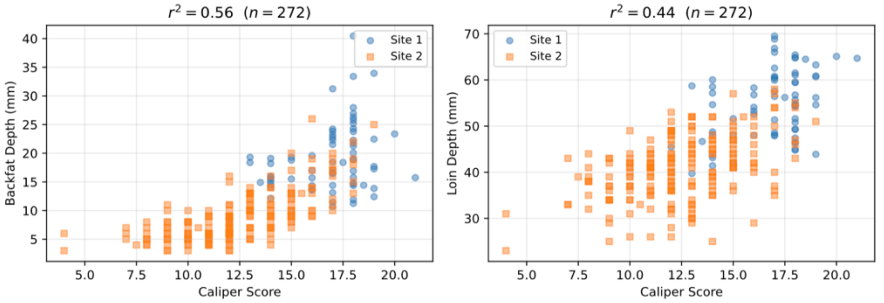
Caliper Score vs. Ultrasound (n = 272 instances with caliper data):

- Backfat depth: **$r^2 = 0.56$**
- Loin depth: **$r^2 = 0.44$**

Within individual sites, correlation is even weaker:

- MSU: $r^2 = 0.11$ (fat), 0.16 (loin)
- UNL: $r^2 = 0.48$ (fat), 0.21 (loin)

→ *Motivates continuous regression of ultrasound measurements over caliper-based ordinal scoring*



Why Not Just Use Global Depth Statistics?

Hand-crafted global features are a weak lower bound

Classical ML on global depth statistics:

- Regressors fed mean / std / percentiles of the depth map reach only **9.13 mm (fat)** and **6.16 mm (loin)**
- That is 2–4 × worse than PigFormer (2.43 mm fat / 5.01 mm loin)
- Aggregate statistics discard where tissue actually sits along the body

What the slice-attention encoder adds:

- Attention over body-axis slices learns which cross-sections (near the last rib) carry the tissue signal
- Learned spatial weighting is the gap between a global summary and a clinical measurement

→ *Spatial structure, not global shape statistics, drives accurate body-condition estimation*

Discussion & Conclusion

Key contributions

- First fully automated two-stage pipeline: raw depth → backfat, loin, total tissue (best overall MAE 3.87 mm)
- Non-contact, ceiling-mounted, real-time (~7 ms/frame, A100, UNet front-end)
- Fat MAE (2.43 mm) approaches human ultrasound variability (1.30 mm std)

Limitations

- No true cross-site test: both sites appear in training, so generalization to an unseen 3rd site is unverified
- No public benchmark exists — results are not directly comparable across studies
- 319 instances (ultrasound annotation cost); loin (5.01 mm) lags fat — more diffuse signal

Future work

- Collect a held-out 3rd site to measure true cross-site generalization
- Release code, dataset, and annotation tool to seed a public benchmark

Acknowledgements & References

Funding

- USDA awards 2022-67021-37858
- NSF Campus CyberInfrastructure grant #2200792 (MSU Data Machine)

Collaborators

- USMARC and MSU experts for data collection and annotation

Key references

- [1] Carion et al. SAM 3: Segment Anything with Concepts (2025)
- [2] Li et al. Mask DINO: Unified Transformer Framework (CVPR 2023)
- [3] Su et al. RoFormer: Enhanced Transformer with Rotary Position Embedding
- [4] Jian et al. Estimation of Sow Backfat Based on Machine Vision (2024)

Code: ***https://github.com/iambashar/Pigformer***

Annotation tool: ***https://www.egr.msu.edu/smarts/annotator/***

Dataset will be publicly available after publication.