WHITEPAPER 18 PAGES / MARCH 2024

Evaluating Curve-Extraction Quality Beyond IoU and F1

Semantic-segmentation models are conventionally graded on pixel-overlap metrics, Intersection-over-Union and F1, and a raster-log digitiser built as a segmentation network inherits that habit. We argue the habit is wrong for this product class. The thing a digitiser delivers is not a pixel mask; it is a one-dimensional curve, a value at every depth, exported as a CSV that a petrophysics package consumes. When VeerNet, our encoder-decoder segmentation network for raster well-log digitisation, is graded on the mask it scores a peak IoU of 0.51 and an F1 of 0.55, with per-class IoU of 0.94 on background but only 0.26 and 0.21 on the two curve classes that actually carry the signal. Read in isolation those numbers look like a model that is not ready. Yet the same checkpoints, evaluated on the reconstructed curves they produce, reach a peak coefficient of determination of 0.9891 against the native LAS data, a lowest mean absolute error of 0.0132, and a lowest mean squared error of 0.0004. The gap between those two readings is not noise; it is structural, and it follows directly from what a one-pixel-wide trace does to an overlap metric. This whitepaper explains why pixel-space metrics systematically undersell thin-structure segmentation, defines a deliverable-space evaluation protocol built on per-curve R-squared, MAE, and MSE measured against ground-truth LAS at fixed depth sampling, walks the per-loss numbers from our five-loss ablation under both readings, and states the operating rule we now apply: a digitiser is graded on the curve it emits and the human-verification step that follows, not on the mask it passes through on the way there.

Tarry Singh
Founder & CEO

Tannistha Maiti
Senior AI Researcher

VEERNET / RASTER-LOG-DIGITISATION / EVALUATION-METRICS / IOU / F1-SCORE / R-SQUARED

CONTENTS

> "The petrophysicist never opens the segmentation mask. They open the CSV. If the curve in that CSV tracks the ground truth, the digitiser did its job, whatever the pixel-overlap score says."

I

THE MISMATCH

A digitiser graded on the wrong artefact

The metric a digitiser inherits is not the metric it should report

We built VeerNet as a semantic-segmentation network, an encoder-decoder with a transformer attention bottleneck, because per-pixel segmentation is the natural shape for finding a curve trace in a scanned well-log image. That architectural choice carries a default evaluation habit with it. Segmentation models are graded on Intersection-over-Union and on F1, the harmonic mean of precision and recall, and a leaderboard for the task we are nominally solving would rank entries by exactly those numbers.

On those numbers, VeerNet reads as a model that is not ready. The peak IoU across our runs is **0.51**. The peak F1 is **0.55**. Break F1 down by class and the picture looks worse where it matters: F1 of **0.97** on the background class, but only **0.37** and **0.32** on the two curve classes. Break IoU down the same way and it is **0.94** on background against **0.26** and **0.21** on the curves. A reviewer who saw only these figures would reasonably conclude the model finds the empty space between curves very well and the curves themselves poorly.

That conclusion is wrong, and the reason it is wrong is the entire subject of this whitepaper. A raster-log digitiser does not deliver a pixel mask to anyone. It delivers a curve: a value at every sampled depth, written to a CSV that a petrophysics package reads as if it had come off a modern logging tool. When we grade that deliverable instead of the mask, the same checkpoints that scored IoU 0.51 reach a peak coefficient of determination of **0.9891** against the native LAS data, a lowest mean absolute error of **0.0132**, and a lowest mean squared error of **0.0004**.

THE SAME MODEL, GRADED TWO WAYS

0.51

Peak IoU on the pixel mask

0.55

Peak F1 on the pixel mask

0.9891

deliverable

Peak R-squared on the reconstructed curve

0.0132

vs LAS

Lowest MAE on the reconstructed curve

The two readings describe one model. The disagreement between them is not measurement error. It is the predictable consequence of grading a one-dimensional product on a two-dimensional overlap metric, and once the mechanism is clear the gap stops being surprising.

The same VeerNet checkpoint, graded two ways. Toggle 'grade the mask' and the pixel-overlap metrics read as a broken model (IoU 0.51, F1 0.55, per-class IoU 0.94 on background but 0.26 / 0.21 on the curves). Toggle 'grade the curve' and the deliverable the petrophysicist actually consumes, the reconstructed 1D trace, lands at peak R-squared 0.9891, lowest MAE 0.0132, lowest MSE 0.0004. The verdict flips and the connecting line diverges. Every metric is the whitepaper's own; the dial fills are normalised for display and the divergence line is schematic geometry.

Why this is not special pleading

There is a lazy version of this argument that says: the headline metric is low, so report a different one until something looks good. That is not the claim. The claim is narrower and, we think, defensible: for this product class the pixel mask is an intermediate representation, not the output, and grading an intermediate representation tells you about a stage in the pipeline rather than about the thing you ship.

A useful contrast is the classical-vision lineage of raster-log digitisation, gridline elimination followed by morphological curve extraction [4]. That work was never evaluated on IoU either, because its authors understood the output to be an extracted curve. The deep-learning reframing imported segmentation metrics along with the segmentation

architecture, and the metrics came in without anyone asking whether they fit the deliverable. They do not. The rest of this document is the argument for grading the deliverable directly, and the protocol for doing it honestly.

II

THE MECHANISM

What a one-pixel trace does to an overlap metric

Sparse foreground breaks IoU before training even starts

The curve trace VeerNet is trained to find is, on the source raster, frequently a single pixel wide. The segmentation target is therefore extremely sparse: in a typical log image the curve classes occupy a tiny fraction of the pixels and the background occupies almost all of them. That sparsity is what poisons the overlap metric, for a reason that has nothing to do with model quality.

IoU is the count of correctly predicted curve pixels divided by the union of predicted-curve and actual-curve pixels. When the true curve is one pixel wide, a prediction that is geometrically correct but offset by a single pixel can have **zero** overlap with the ground truth on that segment, even though, read as a curve, the prediction is off by one pixel of depth, which is below the noise floor of the source scan. The overlap metric punishes a one-pixel lateral shift as if the model had missed the curve entirely. A human reading the two traces overlaid would call them identical.

> A prediction that is one pixel off the ground truth can score zero IoU on that segment and still be a perfect curve. The metric is measuring registration, not recovery.

This is also why the per-class breakdown looks the way it does. Background is a large, dense, easy region, so its IoU of 0.94 is genuine and uninformative. The curve classes are thin and sparse, so their IoU of 0.26 and 0.21 is dominated by sub-pixel registration error rather than by whether the curve was found. The widely-noted failure of overlap metrics on thin structures is exactly this effect; the loss-function literature that targets IoU directly exists precisely because the metric is so unforgiving on sparse foreground [1]. The aggregate IoU of 0.51 is then a weighted blend of one inflated easy class and two deflated hard ones, which is to say it is not a number that means anything about the product.

The deliverable lives one post-processing step downstream

The mask is not the end of the pipeline. After segmentation, a deterministic post-processing stage groups the predicted foreground pixels into a single trace per curve, resolves the depth axis, and samples the trace at fixed depth intervals to produce the exported curve. Our validation notebooks sample at **300** interpolated depth points per curve before comparing against ground truth. That post-processing step is where the sub-pixel registration error washes out: grouping and interpolation turn a slightly jagged pixel mask into a smooth one-dimensional function, and a one-pixel lateral wobble that destroyed the IoU contributes almost nothing to the value of the curve at a given depth.

So the metric that survives to the deliverable is not pixel overlap. It is the agreement between two one-dimensional functions, the predicted curve and the LAS ground truth, sampled at the same depths. That agreement is what R-squared, MAE, and MSE measure, and it is the only thing the downstream petrophysics workflow ever sees.

III

THE PROTOCOL

Grading the curve, not the mask

A deliverable-space evaluation protocol

If the deliverable is a curve, the evaluation has to be on the curve. We grade VeerNet output with three measurements taken per curve, per example, against the native LAS data resampled to the same depth axis.

R-squared against LAS

- Fraction of variance in the ground-truth curve explained by the predicted curve
- Scale-aware: rewards getting the curve shape right, which is what a petrophysicist reads
- Our peak: 0.9891 on the hardest curve-1 example under Tversky loss
- The right headline number for a digitiser

Mean Absolute Error (MAE)

- Average absolute difference between predicted and true value, in the curve's own units
- The number to quote when asked how far off a single sample typically is
- Our lowest: 0.0132; per-curve Dice baseline ran 0.0367 on curve 1

- Cannot be inflated by an easy background class, there is no background in a 1D curve

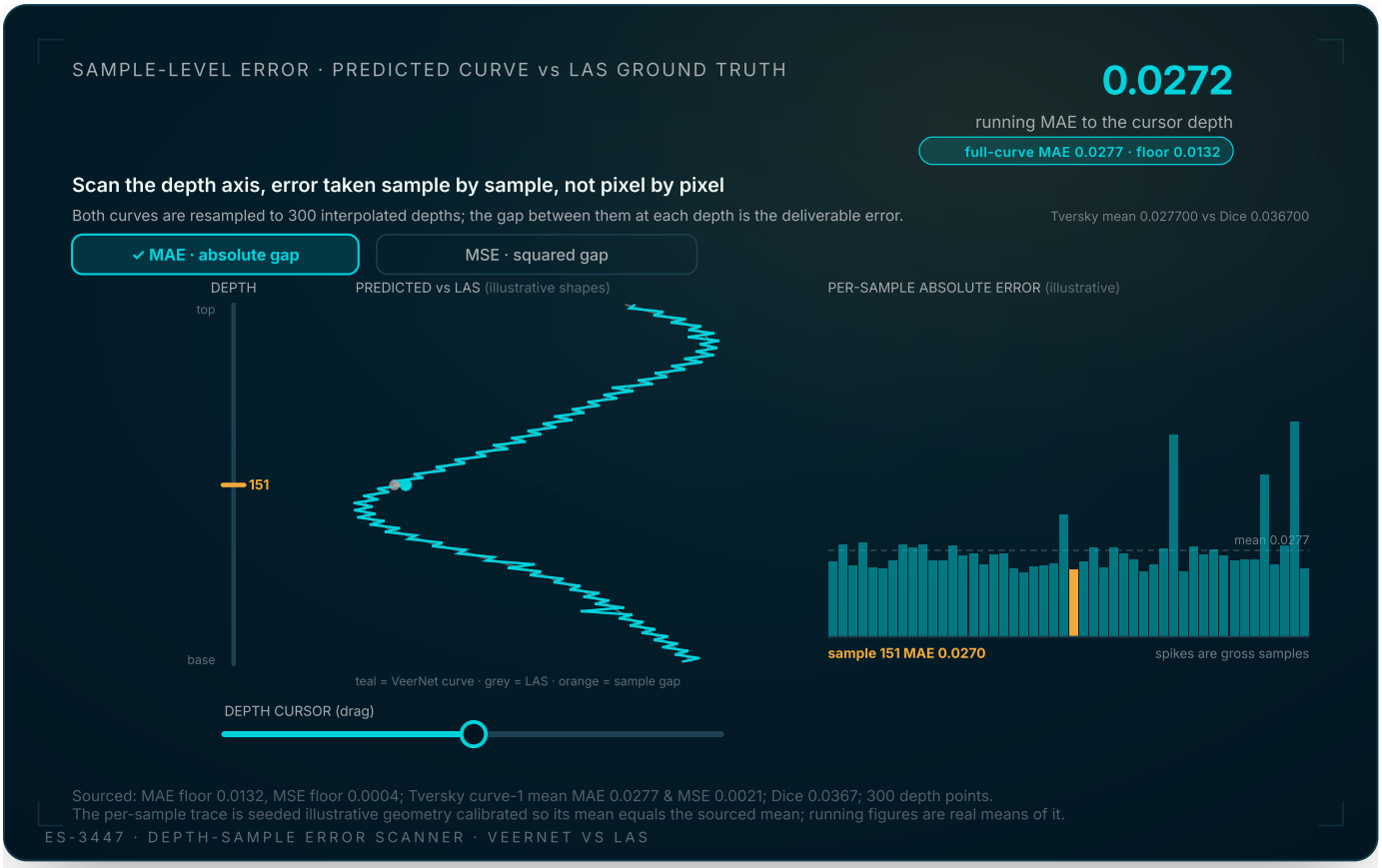
Mean Squared Error (MSE)

- Squares the error, so large mistakes count far more than small ones
- Surfaces rare gross errors that MAE would average away
- Our lowest: 0.0004; curve-1 mean 0.0021 under Tversky loss
- Reported alongside MAE because the two fail in different directions

Each measurement answers a different question, and we report all three rather than collapsing to one, because they fail in different directions and a digitiser can be good on one and weak on another.

- **R-squared** asks whether the predicted curve explains the variance in the true curve. It is the right headline for a digitiser because it is scale-aware and rewards getting the shape right, which is what a petrophysicist reads a curve for. Our peak is **0.9891**.
- **MAE** asks how far off the predicted value is at a typical depth, in the curve's own units. It is the number to quote when someone asks how much they can trust a single sample. Our lowest is **0.0132**.
- **MSE** asks the same question but punishes large errors more than small ones, so it surfaces rare gross mistakes that MAE would average away. Our lowest is **0.0004**.

None of these can be inflated by a large easy background class, because there is no background in a one-dimensional curve. Every sampled depth is a real measurement that the model either got right or did not. That property alone makes deliverable-space metrics more honest than pixel overlap for this task.



The deliverable is graded sample by sample, not pixel by pixel. VeerNet's predicted curve and the LAS ground truth are resampled onto the same axis at 300 interpolated depths, and the gap between the two one-dimensional functions at each depth is the error that survives to the customer. Drag the depth cursor: the running mean accumulates from the top of the curve to the cursor, settling toward the sourced full-curve figures (lowest MAE 0.0132, lowest MSE 0.0004; Tversky curve-1 mean MAE 0.0277, mean MSE 0.0021). Toggle MAE versus MSE to see why both are reported: MSE squares the gap, so the rare gross sample that MAE averages away spikes the squared-error bar. The per-sample trace is seeded illustrative geometry calibrated so its mean equals the sourced mean; all printed figures are either sourced constants or true running means of the trace.

THE REPORTING RULE WE ADOPTED

Report per-curve R-squared, MAE, and MSE against LAS at fixed depth sampling as the headline. Report mask IoU and F1 as diagnostics for the segmentation stage, not as the product score. Never lead with an aggregate IoU that blends an inflated background class with deflated curve classes.

The protocol applied across the loss ablation

We trained five loss functions under otherwise identical conditions [1][2][3]. Read on the mask, the ranking is muddy and every candidate looks mediocre. Read on the curve, the ranking is clean and the strong candidates are clearly strong. The two readings do not even agree on the ordering, which is the sharpest possible evidence that the mask is the wrong scoreboard.

DELIVERABLE-SPACE READINGS FROM THE MULTICLASS ABLATION

0.9891
Peak R-squared, Tversky loss, hardest curve-1 example

0.0277
Mean MAE on curve 1, Tversky loss

0.0367
Mean MAE on curve 1, Dice loss

0.0021
Mean MSE on curve 1, Tversky loss

The Tversky objective, which lets us tilt the precision/recall trade-off explicitly [2], produced the best per-curve reconstruction on the hard examples and the peak R-squared of 0.9891 on a curve-1 example that the overlap metric scored unremarkably. Dice was a close, honest baseline on the curve metrics, with mean curve-1 MAE of **0.0367** against Tversky's **0.0277**. The point is not which loss wins; the companion VeerNet whitepaper settles that with a two-loss schedule. The point is that you cannot see the difference between these losses on IoU, and you can see it immediately on the curve.

> “The same five checkpoints rank one way on the mask and a different way on the curve. Only one of those rankings predicts what a petrophysicist will say when they open the file.”

— FROM OUR OWN ABLATION NOTEBOOKS

IV

THE CONSEQUENCE

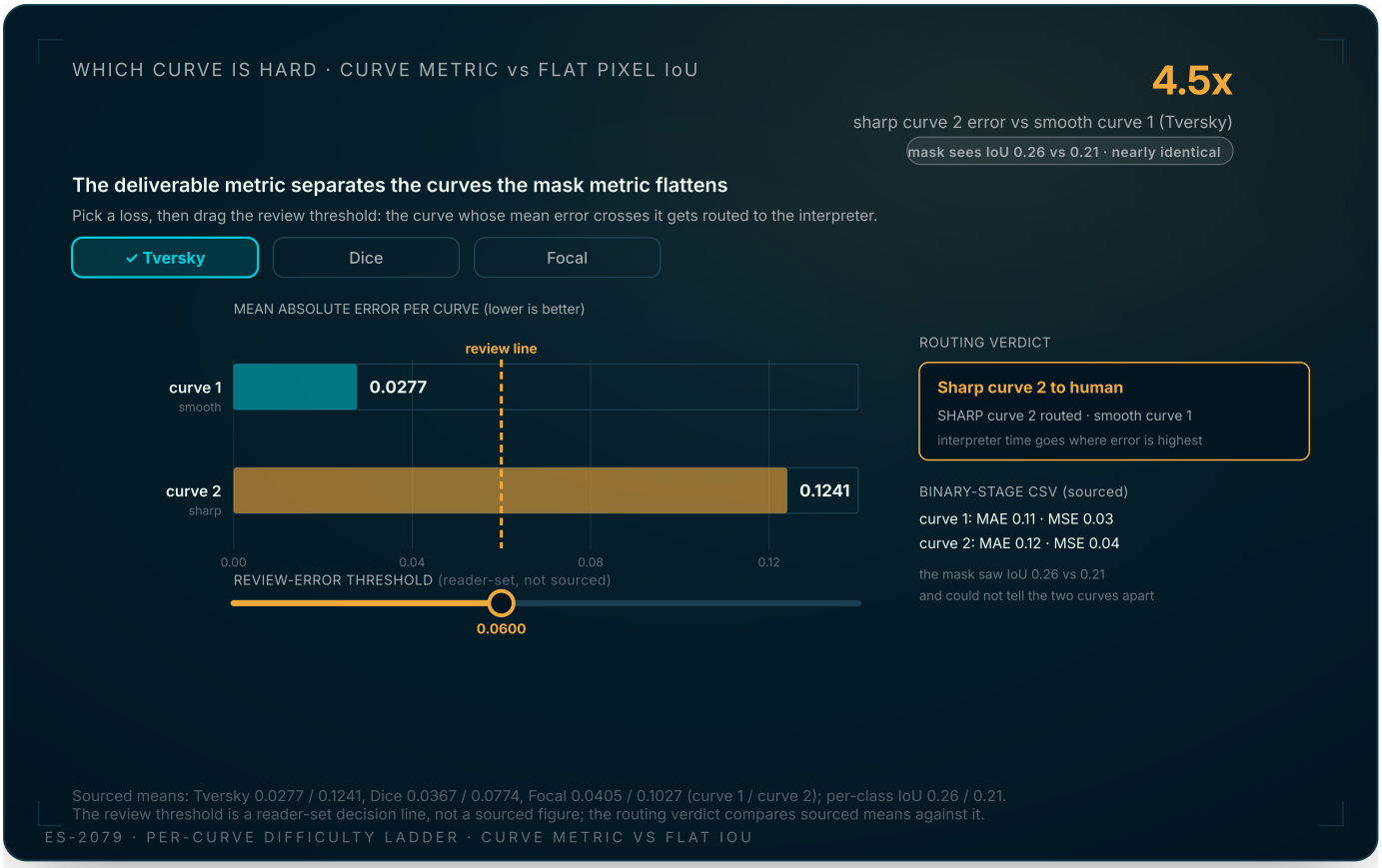
What changes when you grade the deliverable

Per-curve behaviour the mask hides entirely

Grading on the curve does not only make a good model look good. It also tells you where the model is genuinely weak, which the mask cannot. The two curve classes behave differently under reconstruction in a way that IoU 0.26 versus 0.21 completely flattens.

On the smoother, more continuous curve, reconstruction is strong: low MAE, high R-squared, the model recovers the shape and the values. On the sharper, more discontinuous curve, the Dice-loss reconstruction carried a higher mean MAE of around **0.0774** on curve 2 against **0.0367** on curve 1, and binary-stage per-curve CSV errors landed at MAE **0.11** and **0.12** with MSE **0.03** and **0.04**. That is a real, actionable signal: it tells us the sharper curve needs more training data or a recall-tilted loss, and it points at the fix. The pixel-overlap numbers gave no such guidance, because both curves looked equally bad for the same uninformative sub-pixel-registration reason.

This is the practical payoff of measuring the deliverable. The metric stops being a single discouraging headline and becomes a diagnostic that maps onto the work.



The deliverable metric does what the mask metric cannot: it tells you which curve is hard. Pixel IoU flattens both curve classes to a near-identical 0.26 versus 0.21, hiding that the smoother curve 1 reconstructs cleanly while the sharper, more discontinuous curve 2 carries roughly twice the deliverable-space error under every loss (Tversky mean MAE 0.0277 versus 0.1241, Dice 0.0367 versus 0.0774, Focal 0.0405 versus 0.1027). Pick a loss, then drag the review-error threshold: the curve whose mean error crosses the line is routed to the human interpreter, and curve 2 crosses first every time. That is the actionable signal the overlap metric never surfaces. The error means and the per-class IoU are the whitepaper's own; the threshold is a reader-set decision line, labelled as such, and the routing verdict is a direct comparison of the sourced means against it.

The human-verification step is part of the metric

A digitiser does not run unattended. The output is reviewed by an interpreter who corrects the small fraction of curves the model got wrong. That review step is part of the product, and it changes what "good enough" means. A model whose curves are right almost everywhere and confidently flagged where they are not is a model that makes the interpreter fast, regardless of its IoU.

So the metric that actually governs the economics is not even R-squared in isolation. It is the fraction of curve that clears a confidence threshold without human touch, multiplied by how well the reconstructed curve tracks ground truth where it does clear. Both of those are deliverable-space quantities. Neither is visible on the mask. Grading the curve is what

connects the model's numbers to the interpreter's time, and the interpreter's time is what the customer is paying to get back. This is the same reframing the broader upstream-AI literature keeps arriving at: the value of a model is set by the workflow it feeds, not by its offline score in isolation [5].

WHAT TO REMEMBER

1. A raster-log digitiser ships a **curve**, not a mask. Grade the deliverable. The same checkpoints that score IoU **0.51** and F1 **0.55** on the mask reach peak R-squared **0.9891**, lowest MAE **0.0132**, and lowest MSE **0.0004** on the curve.
2. Sparse one-pixel foreground breaks IoU mechanically: a one-pixel lateral shift can score zero overlap on a segment that, read as a curve, is a perfect recovery. The metric is measuring registration, not recovery.
3. Per-class IoU of **0.94 / 0.26 / 0.21** is one inflated easy class blended with two deflated hard ones. The aggregate is not a product score; it is an artefact of class sparsity.
4. Report per-curve R-squared, MAE, and MSE against LAS at fixed depth sampling as the headline; keep IoU and F1 as segmentation-stage diagnostics only.
5. Deliverable-space metrics also diagnose where the model is truly weak (the sharper curve), which the overlap metric flattens away. They turn a discouraging headline into an actionable map.

GLOSSARY

Deliverable-space metric	A metric computed on the artefact the customer actually consumes, the reconstructed 1D curve, rather than on an intermediate representation like the pixel mask. R-squared, MAE, and MSE against LAS are deliverable-space; IoU and F1 are not.
F1	Harmonic mean of precision and recall on the pixel mask. Like IoU, it is a segmentation-stage diagnostic, not a digitiser product score. Peak F1 here is 0.55 overall, 0.37 and 0.32 on the two curve classes.
IoU	Intersection-over-Union. Correctly predicted curve pixels divided by the union of predicted-curve and actual-curve pixels. Mechanically harsh on thin structures: a one-pixel lateral shift can score zero on a segment that is a perfect curve recovery.
LAS	Log ASCII Standard, the canonical text format for digital well-log curves. The ground truth a digitised curve is graded against, and the format the deliverable is ultimately exported to.
MAE	Mean Absolute Error between the predicted curve and the LAS ground truth, resampled to the same depth axis. In the curve's own units. Lowest here: 0.0132.
MSE	Mean Squared Error between predicted and ground-truth curve. Penalises large errors more than MAE. Lowest here: 0.0004.
R-squared	Coefficient of determination. The fraction of variance in the ground-truth curve explained by the prediction. The deliverable-space headline for a digitiser. Peak here: 0.9891.

Thin-structure segmentation	Segmentation where the target foreground is one or a few pixels wide. Overlap metrics behave pathologically here because sub-pixel registration error dominates the score regardless of whether the structure was found.

References

1. Berman, M., Triki, A. R., Blaschko, M. B. (2018). *The Lovasz-Softmax loss: a tractable surrogate for the optimization of the intersection-over-union measure in neural networks*. CVPR. https://arxiv.org/abs/1805.02396
2. Salehi, S. S. M., Erdogmus, D., Gholipour, A. (2017). *Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks*. MLMI. https://link.springer.com/chapter/10.1007/978-3-319-67389-9_44
3. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollar, P. (2017). *Focal Loss for Dense Object Detection*. ICCV. https://arxiv.org/abs/1708.02002
4. Yuan, B., Yang, Q. (2019). *Digitization of Well-Logging Parameter Graphs Based on Gridlines-Elimination Approach*. Journal of Petroleum Exploration and Production Technology. https://link.springer.com/article/10.1007/s13202-019-0656-3
5. Koroteev, D., Tekic, Z. (2021). *Artificial intelligence in oil and gas upstream: Trends, challenges, and scenarios for the future*. Energy and AI. https://www.sciencedirect.com/science/article/pii/S2666546820300033

Get the full whitepaper

This page is the long-form summary. The complete whitepaper adds the full per-loss, per-curve, per-example tables under both readings, the depth-resampling and confidence-thresholding details of the protocol, and the worked reconciliation between the mask metrics and the curve metrics on the same held-out examples.

EARTHSCAN

Evaluating Curve-Extraction Quality Beyond IoU and F1

Published March 2024. Generated 2026-07-22.
https://www.earthscan.io/whitepapers/evaluating-curve-extraction-beyond-iou-f1