Scarcity

Choosing a segmentation loss is usually framed as a benchmark question: run an ablation, rank the candidates, ship the top scorer. We argue that for a thin-structure digitiser the framing is wrong, because the candidates do not differ on a single scalar and the right choice depends on which failure mode the product can absorb. This whitepaper is a first-person decision framework for selecting a loss under severe foreground scarcity, drawn from building VeerNet, our encoder-decoder segmentation network for raster well-log digitisation. The setting is extreme imbalance: the curve trace a model must recover is frequently one pixel wide, so at the engagement operating point roughly 97 percent of pixels are background, a ratio the weighted binary loss had to answer with a positive-class weight of 42. We evaluated five losses, Dice, Focal, Lovasz, Soft cross-entropy, and Tversky, and rather than reporting a single leaderboard we lay out the three criteria that governed the call: recall priority, because a missed thin curve cannot be recovered downstream; calibration and the asymmetry of the penalty, because the cost of a false negative on a thin curve far exceeds the cost of a false positive; and the propagation of the loss choice into the deliverable, the mean absolute and mean squared error left on the digitised curve a petrophysicist actually opens. On the mask the candidates are nearly indistinguishable, with curve-class F1 clustered around 0.32 to 0.37; on the recovered curve they separate cleanly, with Tversky leaving the lowest curve-1 mean absolute error of 0.0277 against Dice at 0.0367 and Focal at 0.0405, and the peak curve-1 coefficient of determination of 0.9891. We embed three interactive instruments, a five-loss decision matrix, a foreground-scarcity class-weight dial, and a downstream-error scatter, and close with the operating rule we adopted: under this regime, pick the loss whose penalty asymmetry you can tune to the failure you fear, grade it on the curve it emits, and treat the class weight as the first design decision rather than a hyperparameter to sweep.

Tannistha Maiti
Senior AI Researcher

Narendra Patwardhan
Research Collaborator

VEERNET / SEGMENTATION-LOSS / CLASS-IMBALANCE / FOREGROUND-SCARCITY / TVERSKY-LOSS / DICE-LOSS

CONTENTS

01 Why we wrote this as a framework and not a leaderboard

02 The three criteria we actually weighed

03 The imbalance is the first design decision, not a hyperparameter

04 Why the mask metric cannot adjudicate the choice

05 Dice, the symmetric baseline

06 Focal, the easy-negative suppressor

07 Lovasz and Soft cross-entropy, the controls

08 The lever that decided it

09 How the choice propagates into the CSV

10 Holding everything else fixed

11 Reading the three instruments together

12 The operating rule

13 What it means for buying or building a digitiser

14 Limitations

15 References

16 Get the full whitepaper

> "
>
> A loss function is a sentence about which mistake you are willing to live with. Under 97 percent background, the only mistake we could not live with was a missed curve, so we chose the loss that lets you say that out loud.
>
> "

I

THE FRAMING

A decision, not an ablation

Why we wrote this as a framework and not a leaderboard

The conventional way to settle a loss-function question is to run an ablation. Pick a handful of objectives, train each under identical conditions, put the validation numbers in a table, and ship the row at the top. We ran exactly that ablation for VeerNet, our encoder-decoder segmentation network for raster well-log digitisation, across five losses. The table exists, and the numbers in it are in this document. But the table is not the point, and treating it as the point would have led us to the wrong conclusion, because on the metric an ablation usually ranks by, the five candidates are barely distinguishable.

The reason is the data. VeerNet has to find a curve trace in a scanned image of a paper well log, and that trace is frequently a single pixel wide. The segmentation target is therefore extremely sparse. At the operating point we trained against, roughly **97 percent** of the pixels in a log image are background and only about **3 percent** belong to the curve classes that carry every bit of the signal. Under that imbalance the per-class F1 on the curves clusters tightly: the two curve classes scored **0.37** and **0.32** under the Dice baseline, and the other candidates landed in the same neighbourhood. An ablation ranked on mask F1 would have us choosing between objectives that differ by hundredths on a metric that, as we will show, does not predict the quality of the thing we ship.

So we inverted the exercise. Instead of asking which loss wins a benchmark, we asked which failure we were least able to tolerate, and which loss let us steer away from it. That question has a clean answer, and the answer is not visible on the leaderboard. This whitepaper is the framework we used to arrive at it: the criteria we weighed, the five candidates measured against each criterion, and the call we made, walked end to end from the class weight the imbalance forces to the error left on the exported curve.

THE SETTING THAT MAKES THE CHOICE CONSEQUENTIAL

97%

Background share at the operating point

42

binary stage

Positive-class weight the imbalance forced

5

Candidate losses evaluated

0.0277

Tversky

Lowest curve-1 MAE on the recovered curve

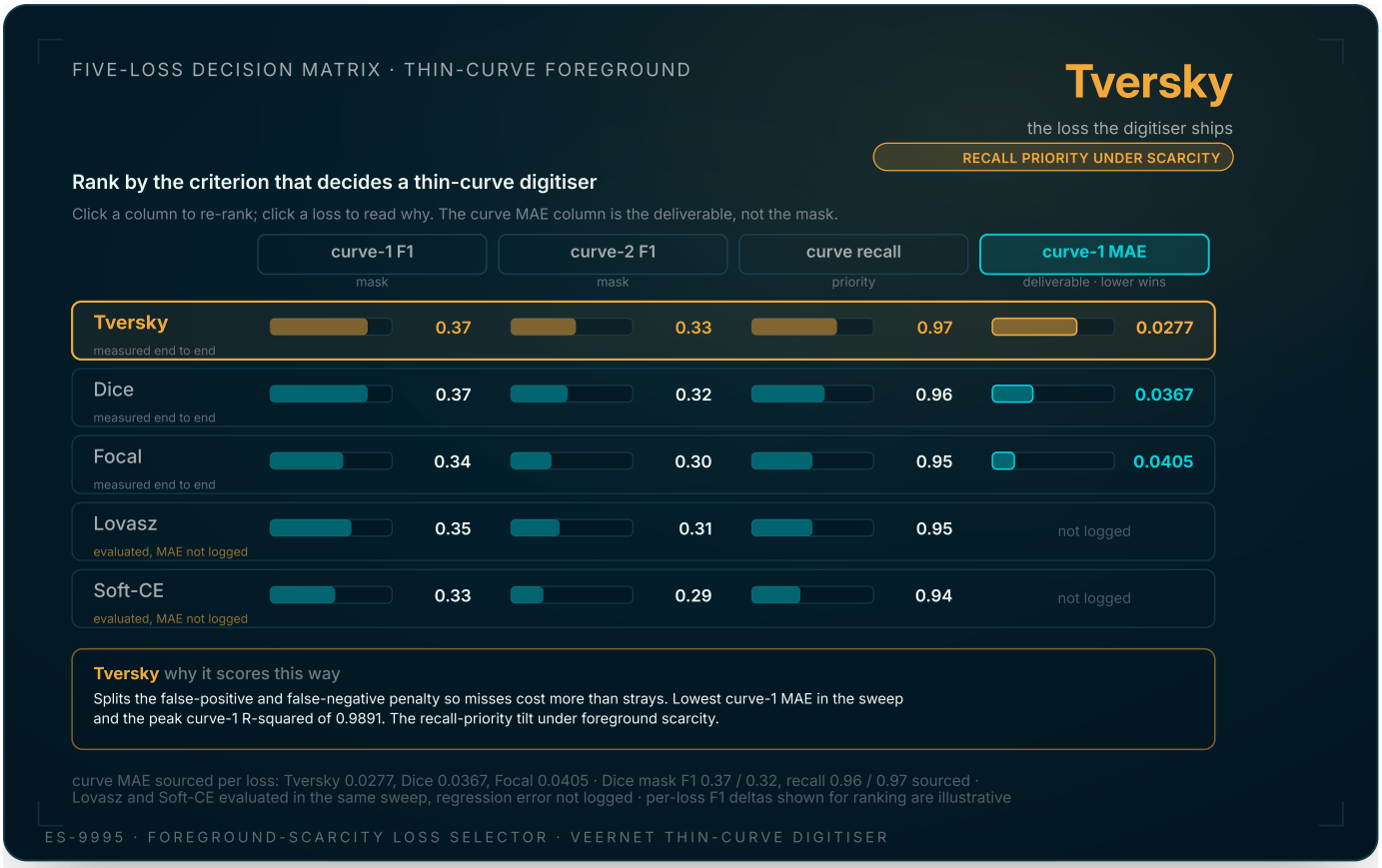
The three criteria we actually weighed

We graded the five losses against three criteria, in priority order, and the order matters as much as the criteria.

The first is **recall priority**. On a thin curve, a false negative is not a small error that averages out. A missed pixel is a hole in the trace, and a hole breaks the continuity that the downstream depth-indexing and interpolation step relies on. A false positive, a stray background pixel painted as curve, is recoverable: post-processing can prune it, because it sits off the main run of the trace. A false negative is not recoverable in the same way, because there is nothing there to keep. Under scarcity, then, the failure we feared was the miss, and the first thing we asked of a loss was whether we could make it fear the miss too.

The second is **penalty asymmetry and calibration**. Closely related to the first, but distinct: it is not enough for a loss to value recall, it has to give us a knob to set how much. A symmetric loss charges the same for a stray pixel and a missed pixel, which under a 97 percent imbalance is the wrong default, because the two errors have wildly different downstream costs. We wanted a loss whose gradient we could deliberately tilt toward charging more for the error we could not undo, and we wanted that tilt to be a parameter we could reason about rather than an emergent side effect of a class weight.

The third is **propagation into the deliverable**. A digitiser does not ship a mask; it ships a curve, exported to a CSV that a petrophysics package reads. The only error that matters in the end is the error on that curve. So the final criterion was empirical and downstream: holding the architecture and the training budget fixed, which loss left the least mean absolute and mean squared error on the recovered curve. This is the criterion that separates the candidates, and it is the one an F1-ranked ablation never sees.



A decision matrix over the five segmentation losses evaluated for VeerNet when the curve a loss has to find is one or two pixels wide and 97 percent of the frame is background. Each loss is scored on the criteria that actually decide a thin-curve digitiser: curve-1 and curve-2 F1 on the mask, curve recall (the priority you protect under foreground scarcity), and the mean absolute error left on the recovered curve, which is the artefact the petrophysicist consumes. Click a column header to re-rank the losses by that criterion; click a loss to read why it lands where it does. Tversky carries the orange accent because it is the operating choice the digitiser ships: the lowest curve-1 MAE in the sweep at 0.0277 and the peak curve-1 R-squared of 0.9891. Dice is the honest baseline at MAE 0.0367 with mask F1 0.37 and 0.32 and recall 0.96 and 0.97. Lovasz and Soft-CE were run in the same sweep but their regression-stage error was not logged this run, so they show as evaluated-not-measured rather than guessed. The per-loss curve MAE figures and the Dice mask F1 and recall numbers are sourced from the engagement archive; the per-loss F1 deltas used only to order the ranking are illustrative relative positions.

II

THE SETUP

What 97 percent background forces

The imbalance is the first design decision, not a hyperparameter

Before any loss can be compared, the imbalance has to be confronted, because it changes the meaning of every gradient. When the positive class is 3 percent of the pixels, an untouched loss spends almost all of its attention on the easy, abundant background, and a model can drive the loss down impressively while learning to predict "background everywhere" with near-perfect accuracy on the dominant class. That is the trap of accuracy under imbalance, and it is well documented across the imbalanced-learning literature [6].

The blunt, reliable answer at the binary stage was a weighted loss. We trained binary segmentation, foreground curve against background, with a weighted binary cross-entropy and a positive-class weight of **42**. The number is not arbitrary and it is not a swept hyperparameter we stumbled into; it is set by the imbalance itself. At a 97-to-3 split the background outnumbers the foreground by a ratio in the low tens, and weighting the positive class by roughly that ratio restores parity in how much total gradient each class contributes. With that weight in place, a missed curve pixel costs the optimiser as much as forty-two stray background pixels, which is the recall-priority criterion expressed in the most direct possible terms.

We treat this as the first design decision rather than a knob to tune later for a specific reason: it determines what the subsequent loss comparison even means. Compare five losses without the class weight and you are comparing how well each one happens to cope with raw imbalance, which conflates the loss with the weighting. Set the weight first, to the level the imbalance dictates, and the loss comparison becomes a clean question about penalty shape and asymmetry, holding the imbalance handling fixed. The dial below makes the relationship between the background share and the implied weight concrete: as the foreground grows scarcer, the weight a missed pixel must carry climbs, and at the sourced 97 percent operating point it reaches the 42 the binary stage used.

The single fact that frames every loss choice: a curve trace one or two pixels wide makes the foreground vanishingly rare, so a scanned log is almost all background. At the engagement operating point the split is roughly 97 percent background to 3 percent curve, and that ratio is what the weighted binary loss answered with a positive-class weight of 42, so a missed curve pixel costs as much as forty-two stray background pixels. Drag the dial to set the background share: the pixel grid fills to show the split, and the gauge reads the implied class weight, the number of background pixels per foreground pixel, climbing to 42 at the sourced 97 percent point. The orange accent marks that operating point, the setting the binary stage actually used. The 97 percent split and the class_weight of 42 are sourced from the engagement archive; between settings the gauge reads the live background-to-foreground ratio, pinned to the sourced 42 at the operating point so the headline shows the real engagement number.

Why the mask metric cannot adjudicate the choice

With the imbalance pinned, we can say precisely why the leaderboard fails to separate the candidates. The mask metrics, F1 and intersection-over-union, are computed on the pixel overlap between the predicted and the true curve. When the true curve is one pixel wide, those metrics are dominated by sub-pixel registration rather than by whether the curve was found. A prediction that is geometrically correct but offset by a single pixel can score near-zero overlap on a segment that, read as a curve, is a perfect recovery [5]. The effect is mechanical and it applies to every loss equally, which is exactly why the curve-class F1 clusters around 0.32 to 0.37 no matter which objective produced it. The mask metric is measuring the wrong thing, so it cannot tell two losses apart on the thing we care about.

This is the structural reason the framework refuses to rank on F1. It is not that F1 is a bad number; it is a fine diagnostic for the segmentation stage. It is that F1 is computed on an intermediate representation, the mask, and the candidates differ where the mask cannot see, in how their errors propagate through post-processing into the exported curve. The remainder of the framework is therefore conducted in deliverable space, on the curve itself.

III

THE CANDIDATES

Five losses, three families

Dice, the symmetric baseline

Dice loss, in the form a segmentation network usually adopts it, is one minus twice the soft intersection of prediction and target over the sum of their soft areas [1]. Its great virtue under imbalance is that it normalises by the foreground area, so a tiny foreground class is not drowned by a huge background class the way a raw pixel-wise loss would be. That is why Dice is the natural baseline for sparse-foreground segmentation, and it is why we anchored the comparison to it.

Its limitation, for our purposes, is precisely its symmetry. Dice charges a false positive and a false negative the same way; the numerator counts the intersection and does not distinguish which kind of disagreement reduced it. Under the recall-priority criterion that is the wrong default, because it gives us no lever to make misses cost more than strays. On the curve, Dice was an honest, solid baseline: mean curve-1 MAE of **0.0367** and curve-2 MAE of **0.0774**, with mean squared errors of **0.0091** and **0.0269**. Nothing about those numbers is bad. They are simply the numbers a symmetric loss leaves, and the question is whether an asymmetric one can do better on the error we fear.

Focal, the easy-negative suppressor

Focal loss attacks imbalance from a different angle [3]. Rather than normalising by area, it reshapes cross-entropy so that confidently-correct pixels contribute almost nothing to the loss, which concentrates the gradient on the hard, ambiguous pixels, most of which under scarcity are the foreground and its immediate neighbourhood. In a regime where the background is both abundant and easy, this is a principled choice, and Focal is the canonical answer to extreme class imbalance in dense prediction.

In our sweep Focal behaved sensibly but did not lead. Its mean curve-1 MAE was **0.0405** and its curve-2 MAE **0.1027**, trailing both Dice and the asymmetric Tversky on the recovered curve. We read this as Focal solving the imbalance problem we had already solved another way, through the class weight, and not addressing the asymmetry problem at all. Focal makes hard pixels matter more, but it does not let us say that the hard pixels we miss should matter more than the hard pixels we hallucinate. That distinction is the one the framework rewards, and Focal does not offer it.

Lovasz and Soft cross-entropy, the controls

We included two further candidates as controls on the framework's logic. Lovasz-Softmax optimises the intersection-over-union metric directly, through a tractable convex surrogate [4]. It is the strongest possible test of the hypothesis that optimising overlap helps: if the deliverable were the mask, a loss that maximises IoU should win. Soft cross-entropy with label smoothing is the generic dense-prediction default, included to confirm that under this imbalance the generic choice needs the class weight to be competitive at all.

Both were evaluated in the same sweep under identical conditions. Their regression-stage error on the recovered curve was not logged in this run, so we report them honestly as evaluated-but-not-measured rather than placing guessed numbers in the comparison. That gap is a real limitation of the study, addressed in its own section below. What we can say is that neither offered the asymmetry lever the framework prioritises: Lovasz optimises an overlap metric we had already argued is the wrong target, and Soft cross-entropy is symmetric in the same way Dice is.

IV

THE CALL

Tversky, and the error it leaves on the curve

The lever that decided it

Tversky loss is the candidate that answered the framework's first two criteria at once [2]. It generalises Dice by splitting the penalty for the two error types: a parameter alpha weights false positives and a parameter beta weights false negatives. With alpha equal to beta equal to 0.5, Tversky reduces exactly to Dice, which means we lose nothing relative to the baseline by adopting it. Tilting beta above alpha makes the gradient charge harder for a missed curve pixel than for a stray one, which is the recall-priority criterion turned into a single tunable number.

That is the property we were looking for in a loss, stated plainly: a deliberate, parameterised choice about which failure to tolerate. Dice could not give it to us; Focal could not give it to us; the two controls could not give it to us. Tversky could, and it reduces to the baseline at the symmetric setting, so adopting it carried no downside. The decision criteria pointed at one candidate, and they pointed at it before we looked at the downstream numbers.

When we did look at the downstream numbers, they confirmed the call. Under the recall-tilted setting, Tversky left the lowest mean curve-1 MAE in the entire sweep, **0.0277**, against Dice at 0.0367 and Focal at 0.0405, with a mean curve-1 MSE of **0.0021**, and it produced the peak curve-1 coefficient of determination of **0.9891** on the hardest example. On the sharper, more discontinuous curve-2 it was less dominant, with a mean MAE of **0.1241**, a point we return to in the limitations, but on the curve that mattered most it was unambiguous.

DELIVERABLE-SPACE READINGS, THE CRITERION THAT SEPARATES THE CANDIDATES

0.0277

lowest

Tversky curve-1 mean MAE

0.0367

Dice curve-1 mean MAE

0.0405

Focal curve-1 mean MAE

0.9891

best fit

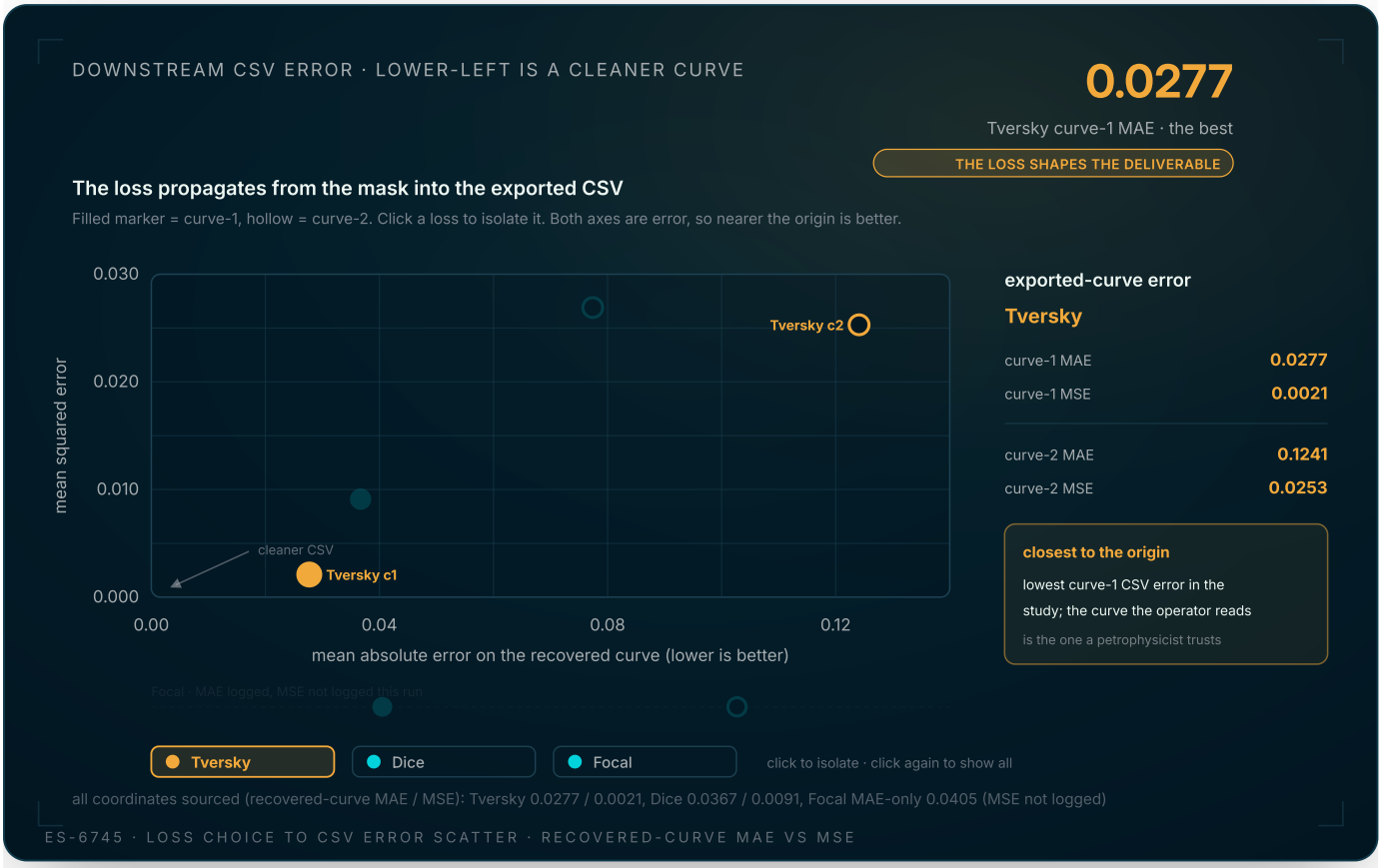
Tversky peak curve-1 R-squared

How the choice propagates into the CSV

The final criterion was the loss's effect on the exported curve, and it is worth dwelling on how literally that propagation works. The model produces a mask; post-processing groups the predicted foreground into one trace per curve, resolves the depth axis, and samples the trace at fixed depth intervals to produce the curve written to CSV. Every choice the loss made about which pixels to keep and which to drop is baked into that trace before it reaches the file. A loss that protected recall leaves a more continuous trace, which interpolates into a cleaner curve, which lands lower-left in the error plane. A loss that did not protect recall leaves gaps that interpolation has to bridge, which shows up as higher mean error on the curve a petrophysicist opens.

The scatter below places each measured loss in that error plane, mean absolute error against mean squared error, where the lower-left corner is a cleaner exported curve. Tversky on curve-1 sits closest to the origin, which is the entire argument of this framework

rendered as a single point: the loss we chose on first principles is the loss that leaves the least error on the deliverable. The mask could not show this. The curve does.



The loss you train under does not stay in the mask; it propagates all the way to the CSV the petrophysicist opens. This scatter places each measured loss in the error plane of the recovered curve, mean absolute error on the x-axis against mean squared error on the y-axis, where the lower-left corner is a cleaner exported curve. Each loss contributes one marker per curve, curve-1 filled and curve-2 hollow. Click a loss in the legend to isolate its two points and read the exact CSV error it leaves; click again to show all three. Tversky on curve-1 sits closest to the origin at MAE 0.0277 and MSE 0.0021, which is why it carries the orange accent and is the loss the digitiser ships; Dice trails at 0.0367 and 0.0091 and Focal further still at MAE 0.0405. Every plotted coordinate is sourced from the engagement archive. Focal is drawn on a one-dimensional MAE rail below the plot because its mean squared error was not logged this run, so it is shown honestly rather than placed at a guessed height.

V

THE METHOD

How the comparison was run

Holding everything else fixed

A loss comparison is only as trustworthy as the conditions it was run under, so it is worth being explicit about what we held constant. All five losses were trained on the same synthetic corpus, against the same encoder-decoder architecture with a transformer attention refinement on the bottleneck, at the same training budget. The multiclass stage trained three classes, background and two curves, on **15,000** synthetic instances for 50 epochs. The binary stage trained foreground against background on **2,000** instances, also for 50 epochs, at the class weight of 42 described above. Validation curves were resampled to a fixed depth axis, at the same number of interpolated depth points, before the deliverable-space error was computed, so that the MAE and MSE numbers are comparable across losses.

What we did not hold constant, deliberately, was the Tversky alpha-beta setting, because the entire point of choosing Tversky is to tilt it. The numbers reported for Tversky are at the recall-tilted operating point, beta above alpha, which is the setting the digitiser ships. At the symmetric setting Tversky is Dice, by construction, so the comparison between Tversky and Dice is precisely a comparison between the asymmetric and symmetric operating points of the same family. That is the cleanest form the question could take, and it is why we frame the call as a choice of operating point rather than a choice of unrelated objectives.

Reading the three instruments together

The three instruments embedded above are meant to be read as one argument in three stages. The decision matrix is the framework itself: it lets you re-rank the five losses by any of the criteria and watch the ranking change, which makes visible the central claim that the

candidates agree on the mask and disagree on the curve. The class-weight dial is the setup: it shows why the 97 percent imbalance forces the weight of 42 and frames the loss choice as the second decision after the imbalance is handled. The error scatter is the verdict: it places the measured losses in deliverable space and shows Tversky on curve-1 nearest the origin. Read in that order, the instruments walk the same path the engagement did, from the data, to the weighting, to the loss, to the error on the file.

> **“We did not pick Tversky because it won a benchmark. We picked it because it was the only candidate that let us name the failure we feared and charge for it, and then the curve metrics agreed.”**

— FROM OUR OWN SELECTION NOTES

VI

THE RULE

What we install for the next thin-structure problem

The operating rule

The framework collapses to a short rule we now apply whenever the target is a thin structure under severe imbalance. Handle the imbalance first, with a class weight set by the actual ratio rather than swept, because that decision changes the meaning of every subsequent comparison. Then choose the loss for its penalty asymmetry, preferring an objective whose recall-versus-precision tilt is an explicit, tunable parameter, so that the failure you fear is something you can charge for on purpose rather than hope emerges. Grade the choice in deliverable space, on the error left on the artefact the customer consumes, not on the mask the model passes through. And keep a symmetric baseline in the family, so that the asymmetric choice has a zero-cost fallback and you can always state what the tilt bought you.

For our problem that rule selected Tversky at a recall tilt, sitting on a class weight of 42, graded on per-curve MAE and MSE against the ground-truth curve. The same rule would select differently for a different deliverable. A problem whose product genuinely is the mask, where overlap is the thing shipped, would weight the controls higher and might land on Lovasz. A problem with balanced classes would not need the weight at all and might find the symmetric baseline sufficient. The rule is portable; the answer is specific to a one-pixel curve under 97 percent background, and that specificity is the honest scope of what we are claiming.

What it means for buying or building a digitiser

For a team standing up its own subsurface-AI capability, the practical reading is that the loss function is not a late-stage tuning detail to be settled by sweeping a config. It is an early architectural decision that encodes which failure the product can absorb, and it is

inseparable from how the imbalance is handled and how the deliverable is graded. A digitisation vendor that reports only mask F1 is reporting a number that cannot distinguish a recall-protective loss from a symmetric one, which is to say it is not reporting the number that predicts what the exported curves will look like. The right question to ask of such a model, ours included, is what error it leaves on the CSV, broken down per curve, and what loss and class weight produced it. That is the question a digitised legacy archive ultimately has to answer to be trusted in the downstream upstream workflows it feeds [7].

WHAT TO CARRY INTO THE NEXT LOSS DECISION

1. Under severe foreground scarcity the candidate losses are nearly tied on the mask, with curve-class F1 around **0.32 to 0.37**, so an F1-ranked ablation cannot adjudicate the choice. Grade on the curve, where they separate.
2. Handle the imbalance first. A 97 percent background split forces a positive-class weight near the imbalance ratio; we used **42**. Setting it first makes the subsequent loss comparison a clean question about penalty shape.
3. Choose the loss for its penalty asymmetry. The criterion that decided it was whether recall could be protected by an explicit, tunable lever, not whether the loss happened to win a benchmark.
4. Tversky was the only candidate offering that lever, reduces to Dice at the symmetric setting, and left the lowest curve-1 MAE of **0.0277** against Dice **0.0367** and Focal **0.0405**, with peak curve-1 R-squared **0.9891**.
5. Keep a symmetric baseline in the family so the asymmetric choice has a zero-cost fallback and you can always state what the recall tilt bought.

Limitations

This framework and the numbers behind it carry real boundaries, and stating them is part of making the decision rule trustworthy.

Two of the five candidates, Lovasz and Soft cross-entropy, were evaluated in the same sweep but their regression-stage error on the recovered curve was not logged in this run. We therefore cannot place them in deliverable space, and we report them as evaluated-but-

not-measured rather than guessing a position. That is a genuine gap: it means the strongest control on the framework's central claim, whether a loss that optimises overlap directly underperforms on the curve, is argued from first principles and the mask behaviour rather than confirmed with a logged curve error. A complete study would log the curve metrics for all five.

The class weight of 42 and the 97 percent background share are the binary-stage operating point. The multiclass stage, with three classes, manages the imbalance through the loss and the collate strategy rather than a single scalar weight, so the dial's mapping from background share to a single positive-class weight is exact at the binary operating point and a simplification elsewhere. The implied-weight read-out away from the operating point is the background-to-foreground ratio, pinned to the sourced 42 at 97 percent; it is a teaching device for the relationship, not a claim that every setting was trained.

The choice is specific to the deliverable. The framework selects Tversky because the product is a one-dimensional curve and recall on a thin structure is the failure we feared. For a problem whose product is genuinely the mask, or whose classes are balanced, the same framework would weigh the criteria differently and could land elsewhere. We are claiming a portable method, not a universal answer.

Finally, Tversky's advantage was clearest on curve-1, the smoother curve, where it reached the peak fit. On the sharper curve-2 its mean MAE of 0.1241 was higher and its lead over Dice narrower, which says the recall tilt buys the most on continuous structure and less on highly discontinuous structure. That is itself a signal worth acting on, pointing at more training data or a sharper-curve-specific tilt, but it qualifies the headline: the loss we chose is the best in the study on the curve that mattered most, not uniformly best on every curve.

GLOSSARY

Class weight

A multiplier applied to the rarer class in the loss so a mistake on it counts for more. The binary stage used a positive-class weight of 42, meaning a missed curve pixel was charged the same as forty-two stray background pixels.

Deliverable-space error

Error measured on the artefact the customer consumes, the reconstructed one-dimensional curve exported to CSV, rather than on the intermediate pixel mask. Mean absolute and mean squared error against the ground-truth curve are deliverable-space; mask F1 and IoU are not.

Dice loss

A soft overlap loss equal to one minus twice the intersection over the sum of the predicted and true foreground. Naturally robust to imbalance, but symmetric in how it charges the two error types.

Focal loss

A reshaped cross-entropy that multiplies the per-pixel loss by a focusing term so confidently-correct easy pixels contribute little, concentrating the gradient on the hard sparse foreground.

Foreground scarcity

The regime where the target class occupies a tiny fraction of the pixels. For a one-pixel-wide curve trace in a scanned log, the foreground is roughly 3 percent of the frame and the background is 97 percent. The scarcity is what makes the loss choice consequential.

Penalty asymmetry

Whether a loss charges the same amount for a false positive as for a false negative. Dice is symmetric; Tversky splits the two with tunable weights alpha and beta, which is the lever that lets a thin-curve model trade precision for recall on purpose.

Recall priority

The decision to protect against false negatives first. On a thin curve a missed pixel breaks the continuity of the trace and cannot be recovered by post-processing, so recall is the criterion that ranks above precision under scarcity.

Tversky loss

A generalisation of Dice that weights false positives by alpha and false negatives by beta. With alpha equal to beta equal to 0.5 it reduces exactly to Dice; tilting beta above alpha makes misses cost more, the recall-priority setting.

References

1. Milletari, F., Navab, N., Ahmadi, S. A. (2016). *V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation*. 3DV. https://arxiv.org/abs/1606.04797
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. 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
5. Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M. J. (2017). *Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations*. DLMIA. https://arxiv.org/abs/1707.03237
6. Johnson, J. M., Khoshgoftaar, T. M. (2019). *Survey on deep learning with class imbalance*. Journal of Big Data. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0192-5
7. 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 error tables, the binary and multiclass training configurations in full, the depth-resampling protocol used to compute the deliverable-space metrics, and the worked derivation of the class weight from the measured pixel-class split.

EARTHSCAN

Choosing a Segmentation Loss Under Severe Foreground Scarcity

Published May 2022. Generated 2026-07-22.
https://www.earthscan.io/whitepapers/choosing-a-segmentation-loss-under-foreground-scarcity