Structure Segmentation: A Field Guide

Under severe foreground scarcity, choosing a segmentation loss is only one of at least three decisions a practitioner has to make, and treating it as the whole answer is the most common way thin-structure digitisers underperform. This whitepaper is a field guide to imbalance mitigation for a target that occupies under two percent of the pixels, drawn from building VeerNet, our encoder-decoder segmentation network for raster well-log digitisation, where a curve trace is frequently one pixel wide and roughly ninety-seven percent of every log image is background. We survey three families of remedy rather than ranking objectives against each other. The first is loss weighting: a weighted binary cross-entropy with a positive-class weight of forty-two, set by the imbalance ratio, which lifts the binary-stage curve recall to a sourced ninety-six and ninety-seven hundredths. The second is region-based losses, the Dice and Tversky family evaluated among five objectives, which optimise overlap directly and are the run where both curve overlap and curve precision are measured together: on the multiclass Dice stage the curve intersection-over-union sits at twenty-six and twenty-one hundredths against a background at ninety-four, and the curve precision on that same run is a low forty-one and thirty-six hundredths. The third is synthetic balancing: procedurally generating logs that raise the genuine foreground share the optimiser trains against, which pays back in precision by giving the model real positives to learn from rather than reweighted copies of the few it has. The central argument, made visible in an interactive frontier board, is that each family moves a different axis and none moves all three, so a usable curve is a stacking decision, not a selection decision. We close with the operating posture we adopted, the peak intersection-over-union of fifty-one hundredths the combined stack reached, the failure modes each remedy introduces, and the order in which a team facing the same scarcity should reach for them.

The EarthScan Team
Research, engineering, and field

VEERNET / CLASS-IMBALANCE / FOREGROUND-SCARCITY / THIN-STRUCTURE-SEGMENTATION / LOSS-WEIGHTING / REGION-BASED-LOSS

CONTENTS

> "
>
> The mistake is not picking the wrong remedy for class imbalance. The mistake is picking one remedy and expecting it to move all three of the axes that a thin-structure model actually lives on.
>
> "

I

THE FRAMING

One remedy is never the whole answer

Why this is a field guide and not another ablation

We have written before about how to choose a single segmentation loss under foreground scarcity, and that piece is a decision framework: given five objectives, which one do you ship. This document sits one level up. It is about the fact that the loss is not the only decision, that at least three separate families of remedy are in play whenever the target class is vanishingly rare, and that each family answers a different failure. A team that reads only the loss question and stops there ships a model that is good on one axis and quietly poor on the other two. The purpose of a field guide is to keep that from happening.

The setting is the same one that shapes everything we build on VeerNet, our encoder-decoder segmentation network for raster well-log digitisation. The model has to find a curve trace in a scanned image of a paper log, and that trace is frequently a single pixel wide. At the operating point we train against, roughly **97 percent** of the pixels in a log image are background and under **2 percent** belong to the thin curve classes that carry every bit of the signal. That ratio is not a nuisance to be normalised away with one trick. It is the defining property of the problem, and it interacts with the loss, with the data, and with the evaluation metric in three different ways that three different remedies have to address. Getting it right is worth the effort because a digitised legacy curve archive feeds real upstream workflows, which is what makes a careful mitigation stack a product decision rather than an academic one [10].

The taxonomy we follow is the standard one from the imbalanced-learning literature, which sorts the remedies into algorithm-level methods that change the objective, data-level methods that change the training distribution, and hybrids that combine them [1]. Our three families map onto that taxonomy cleanly. Loss weighting and region-based losses are algorithm-level; synthetic balancing is data-level; and the operating posture we ended on is the hybrid. What the literature establishes in general, and what this guide argues for the

specific case of a one-pixel curve, is that these families are not substitutes. They address the imbalance through different mechanisms, so their effects compose rather than overlap [2].

THE SCARCITY THAT FORCES THE WHOLE STACK

97%

Background share at the operating point

42

weighting

Positive-class weight the ratio forces

5

Losses evaluated in the region-based sweep

0.51

stacked

Peak IoU the combined stack reached

The three axes a thin-structure model lives on

Before surveying the remedies it is worth being precise about what they are trying to move, because the whole argument rests on there being more than one thing. A segmenter of a thin curve is graded on three numbers that do not rise and fall together. Recall is the fraction of true curve pixels the model actually found, and under scarcity it is the number a missed thin trace destroys, because a hole in a one-pixel curve cannot be recovered by anything downstream. Precision is the fraction of predicted curve pixels that were really curve, and it is what stray predictions on the abundant background erode. Intersection-over-union is the overlap of the predicted and true curve regions, the number that most directly reflects whether the mask has the right shape in the right place.

These three move independently, and that independence is the entire reason one remedy is not enough. A change that buys recall by making the model fire more generously on the thin curve costs precision, because more of what it paints is wrong. A change that improves overlap does not, on its own, make the model find more of the curve. And a change that gives the model more genuine positives to learn from improves precision without inflating recall the way a bigger class weight does. If the three numbers rose together, any single remedy that lifted one would lift all, and this guide would not need to exist. They do not, so it does.

II

FAMILY ONE

Loss weighting buys recall, and charges for it in precision

Weighting is the blunt instrument, and it is the first one to reach for

The most direct answer to imbalance is to tell the loss that the rare class matters more. At the binary stage of VeerNet, foreground curve against background, we trained a weighted binary cross-entropy with a positive-class weight of **42**. The number is not swept and it is not arbitrary; it is set by the imbalance itself. At a ninety-seven to three 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 to a training step. With that weight in place, a missed curve pixel costs the optimiser as much as forty-two stray background pixels, which is recall priority stated in the most literal possible terms.

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.

It works, on the axis it is designed for. Under that weighted binary loss the recall on the curve masks reached the sourced **0.96** and **0.97**, which is to say the model found nearly all of the true curve pixels. That figure is the binary stage's logged number, and it is the only metric that stage logs; recall is exactly what this remedy exists to protect, and for a thin structure where a miss is unrecoverable it is the right first move. But the weighting buys that recall with a currency, and the currency is precision. The weight that makes a miss expensive also makes a false alarm cheap, so a model tuned this hard for recall paints background it has fired on too eagerly, and the price is a low curve precision. We put a sourced number on that precision in the next family, where curve precision is measured on the same run as the overlap it trades against, rather than pairing it here with a recall from a different training stage.

WHAT LOSS WEIGHTING MOVED, AND WHAT IT DID NOT

0.96 / 0.97

lifted

Binary-stage curve recall at class_weight 42

42

Positive-class weight, set by the ratio

precision

the cost

The axis it charges, priced in family two

1

Axis this family moves on its own

This is the first entry in the field guide and the first lesson in it. Loss weighting is the correct opening move under scarcity, because recall is the axis you cannot afford to lose and weighting is the cheapest way to protect it. But it is a lever with one direction. Turned up it trades precision for recall along a single seesaw, and no setting of it produces a model that is strong on both. An inverse-frequency weight is also only a first approximation of the right reweighting; schemes keyed to the effective number of samples rather than the raw ratio exist precisely because the raw ratio over-weights once the rare class has enough examples to learn from [9]. We used the raw ratio because it is defensible and simple, and because the precision it leaves on the table is a problem for the other two families to solve, not for a bigger weight.

III

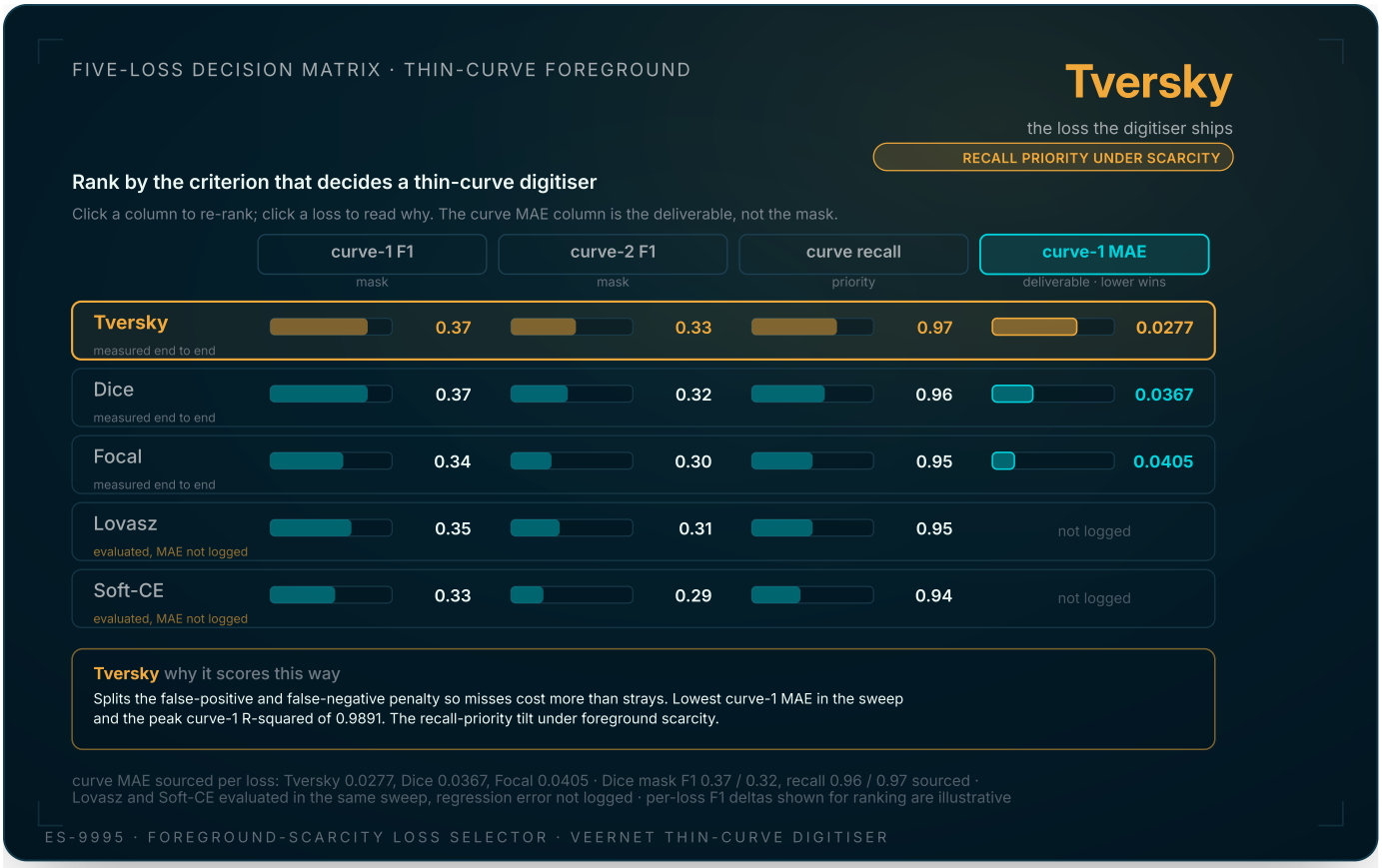
FAMILY TWO

Region-based losses buy overlap, off a floor a pixel loss cannot leave

The overlap family, and why we evaluated five of them

The second family changes the shape of the loss rather than the weight in front of it. A pixel-wise loss, weighted or not, sums a penalty over individual pixels and has no notion of the region the pixels form. A region-based loss is computed on the overlap between the predicted and the true foreground as a whole, which under imbalance has a structural advantage: it normalises by the foreground area, so a tiny target class is not numerically drowned by an enormous background class the way a raw per-pixel sum would be [3]. Dice loss, one minus twice the soft intersection over the sum of the soft areas, is the anchor of the family, and it is the natural objective for sparse-foreground segmentation for exactly that reason.

We did not adopt Dice on faith. We evaluated **5** losses in this family and its neighbours: Dice, Focal, Lovasz, Soft cross-entropy, and Tversky. Focal reshapes cross-entropy so confidently-correct easy pixels contribute little and the gradient concentrates on the hard sparse foreground [5]. Lovasz optimises the intersection-over-union metric directly through a tractable surrogate, the strongest test of whether optimising overlap is sufficient on its own [6]. Tversky generalises Dice by splitting the penalty for false positives and false negatives into tunable weights, which lets a thin-structure model charge harder for the miss it fears [4]. Soft cross-entropy is the generic dense-prediction default, included as a control. The reason the overlap family behaves well when one class is vanishingly rare, and the reason a raw overlap can still be brittle at the extreme, are both treated carefully in the generalised-Dice work, which is where we grounded our reading of the sweep [7].



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.

What the region-based family moved is the intersection-over-union. On the multiclass stage, three classes of background and two curves, the Dice-trained model reached an IoU of **0.94** on the background mask, which is easy, and **0.26** and **0.21** on the two curve classes, which is the hard part. Those curve numbers look low in absolute terms, and they are, but the point is the direction: a region-based loss is what lifts curve overlap off the floor at all, because it is the only one of the three families whose objective is overlap. Loss weighting on its own does not improve the shape of the mask; it improves how much of the curve is found. Getting the predicted region to sit where the true region sits is a job for a loss that scores regions, and that is what this family delivers.

That same multiclass Dice run is also where we have a sourced number for the precision the weighting spent. On the two curve classes the Dice-stage curve precision was **0.41** on the first curve and **0.36** on the second, which means that well over half of the pixels the model painted as curve were background it had fired on too eagerly. We report precision here rather than back in the weighting section on purpose: the logged precision and the logged overlap come from this one Dice run, so they are a joint readout of a single training configuration, whereas the recall of 0.96 and 0.97 is a separate binary-stage measurement. Pairing the two would be stitching numbers from two runs into one operating point, which they are not, and the whole guide turns on being honest about which family a number belongs to.

WHAT THE REGION-BASED FAMILY MOVED, ON ONE DICE RUN

0.26 / 0.21

overlap

Curve IoU, off the floor (multiclass Dice)

0.41 / 0.36

still scarce

Curve precision on the same Dice run

0.94

Background IoU, the easy class (Dice)

5

Losses evaluated in the sweep

The peak intersection-over-union anywhere in the study was **0.51**, reached not by any single loss in isolation but at the combined operating point where the region-based loss sat on top of the weighting and the balanced data. That number is the header of the argument this guide is building toward: no one family produced it, and it is the frontier board in the next section that makes the reason visible.

IV

FAMILY THREE

Synthetic balancing buys precision, with positives the model has not seen

More real positives beat more copies of the same few

The third family does not touch the loss at all. It changes the data the loss is computed on. The oldest and most durable result in imbalanced learning is that synthesising new minority examples beats duplicating the ones you already have, because duplication teaches the model the specific few positives by heart while synthesis teaches it the variety the class actually contains [8]. Oversampling by copying inflates the minority count without adding information; the model overfits the copies and generalises no better. Generating genuinely new minority instances adds information, and that is the difference between a precision that holds on unseen logs and one that collapses.

For a raster-log digitiser this family takes a specific and unusually powerful form. Because a log image is a rendering of curves against a grid, we can generate synthetic logs procedurally, drawing plausible curve traces with controlled shapes, crossings, and noise, and thereby manufacture as many genuine foreground pixels as training needs. The multiclass stage trained on **15,000** synthetic instances and the binary stage on **2,000**, and every one of those instances is a fresh positive rather than a reweighted copy of a scarce real one. Raising the genuine foreground share the optimiser sees is a data-level answer to the same imbalance that the class weight answers at the algorithm level, and because the two operate through different mechanisms, their effects add rather than duplicate [2].

The axis this family pays back on is precision, which is the axis loss weighting spent. A model that has seen many varied true curves learns the difference between a curve pixel and a background pixel that merely resembles one, so it fires less often on the lookalikes that a recall-hungry class weight would otherwise have it paint. Where weighting bought recall by making the model generous, synthetic balancing recovers precision by making the model informed, and it does so without giving back the recall the weight secured. That is

the composition the whole guide is about: two families moving two different axes in two different directions, so that the operating point ends up somewhere neither could reach alone.

> **“Weighting made the model find the curve. Balanced synthetic data made it stop mistaking the background for one. Neither fact is visible if you only ever change the loss.”**

— FROM OUR OWN TRAINING NOTES

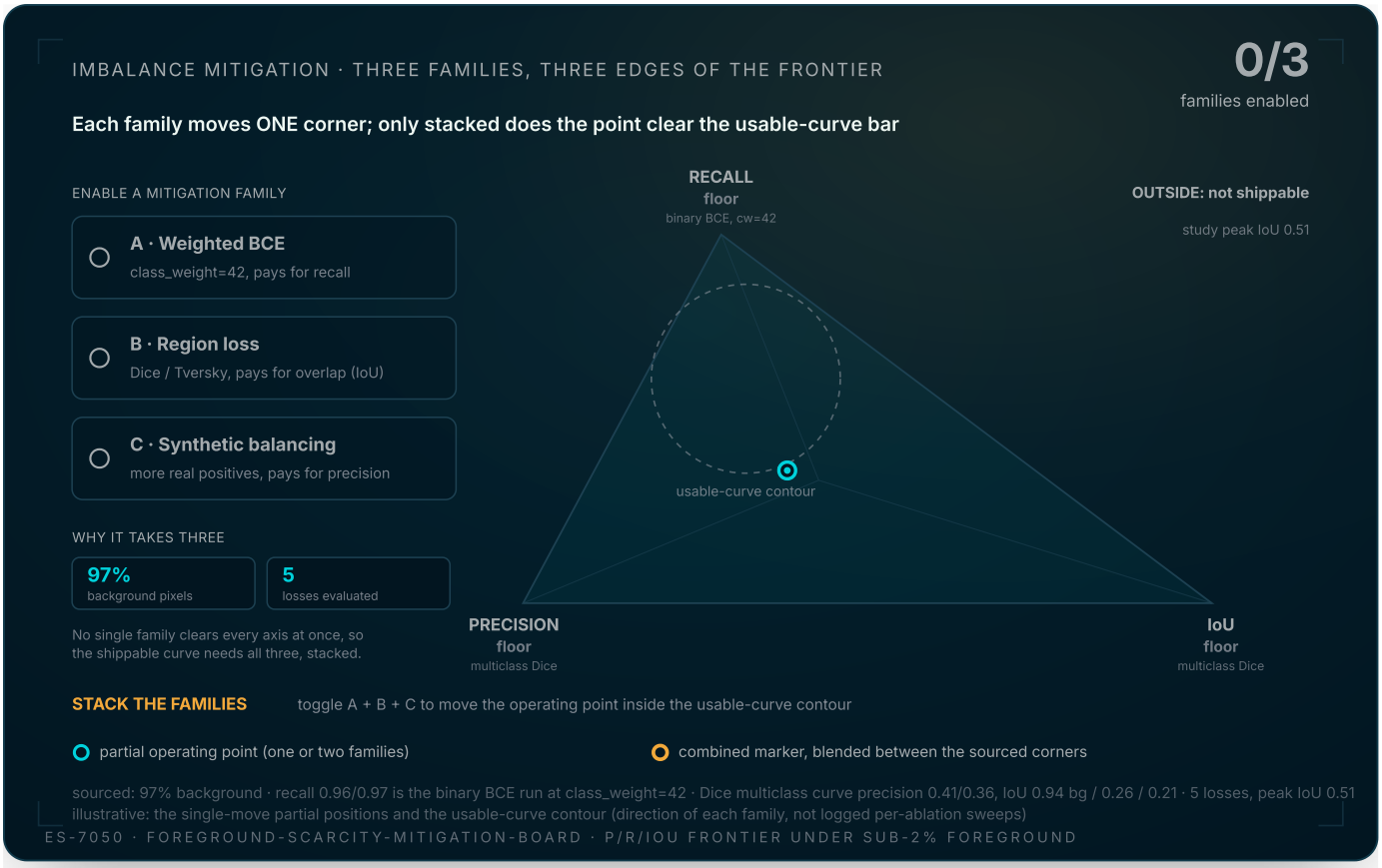
V

THE ARGUMENT

Three moves, three edges, one shippable point

The frontier is the argument

The claim of this field guide is now stated plainly enough to draw. There are three axes, recall, precision, and intersection-over-union, and there are three families of remedy, and each family moves a different edge of the frontier those axes bound. Loss weighting pushes toward recall. Region-based losses push toward overlap. Synthetic balancing pulls back toward precision. No single family lands the operating point in the region where all three are acceptable, because each one moves along its own edge and leaves the other two roughly where they were. A usable curve is therefore not a selection, it is a composition, and the instrument below lets you compose it move by move.



A precision-recall-IoU frontier board for imbalance mitigation under severe foreground scarcity. Under roughly 3 percent foreground, three mitigation families each push a different corner of the frontier. Weighted binary cross-entropy at class_weight=42, the binary stage, pays for recall: the sourced 0.96 and 0.97 are that binary run, which is the only metric the archive logs for it. A region-based loss from the Dice and Tversky family, the multiclass stage, is where overlap and curve precision are both measured on one run: it lifts the multiclass curve IoU off the floor toward the sourced 0.26 and 0.21, and on that same Dice run the curve precision is a low 0.41 and 0.36. Synthetic balancing gives the optimiser more genuine positives and pushes precision back toward acceptable. Enable them one at a time and the operating point moves along a single edge; enable all three and the orange combined marker is the only one that crosses inside the dashed usable-curve contour. The recall corner is sourced from the binary weighted-BCE run and the precision and IoU corners from the multiclass Dice run, each labelled to its own run rather than fused as one readout; the background share, the five losses evaluated, and the 0.51 peak IoU are also sourced from the engagement archive; the single-move partial positions and the usable contour are illustrative geometry drawn to show the direction each family pushes and to land the combined marker between the sourced corners, not logged per-ablation sweeps.

The board is a precision-recall-IoU frontier with the three families as toggles, and each corner is labelled with the run its number comes from so no two are read as one operating point. Enable weighting alone and the operating point climbs toward the recall corner, sitting on the sourced binary-stage **0.96** and **0.97**, but it stays far from precision and only partway up the overlap axis. Enable the region-based loss alone and it moves toward the IoU corner, toward the multiclass Dice **0.26** and **0.21** off the floor, without buying the recall a thin curve needs. Enable synthetic balancing alone and it drifts back toward the precision

corner, the multiclass Dice curve precision of **0.41** and **0.36** that the family exists to lift. Each single move leaves the marker outside the dashed usable-curve contour, and that is the visual form of the argument: one remedy is never enough. Only when all three are on does the point cross inside the contour and turn orange, blending between the three sourced corners rather than snapping to any single logged operating point, which is why the board flags that combined position as illustrative geometry. The orange marker is the only element on the board that carries the claim, and it appears only when the stack is complete.

Reading the board move by move is meant to feel like the sequence a practitioner lives. You reach for the class weight first because recall is the axis you cannot lose, and you watch precision fall as the price. You reach for the region-based loss to make the mask the right shape, and you watch overlap climb off the floor. You reach for synthetic balancing to recover the precision the weight cost, and you watch the point finally settle where a petrophysicist would accept the curve. The order is not arbitrary, and the last section makes it explicit.

VI

THE PRACTICE

The order to reach for the families, and what each one breaks

The operating posture we adopted

The posture that came out of this is a short sequence rather than a single choice. Reach for loss weighting first, and set the positive-class weight from the actual imbalance ratio rather than sweeping it, because recall is the failure you cannot recover from on a thin structure and weighting is the cheapest protection for it. Reach for a region-based loss second, from the Dice and Tversky family, because overlap is the axis a pixel-wise loss cannot move and the region losses are the only family whose objective is the shape of the mask; keep a symmetric Dice baseline in the family so the asymmetric Tversky setting has a zero-cost fallback. Reach for synthetic balancing third, and make it genuine synthesis rather than duplication, because it is what pays back the precision the class weight spent and it does so with positives the model has not memorised. Then grade the whole stack in deliverable space, on the error left on the exported curve, not on the mask the model passes through.

That sequence is deliberate in its ordering. Weighting is first because it is the fastest to apply and it protects the axis with the harshest downstream penalty. The region-based loss is second because it changes the objective and needs the weighting already in place to define what the comparison means. Synthetic balancing is last in the sequence to reason about but is in truth running underneath the whole thing, because the data has to exist before any loss is trained on it; we list it third because its precision contribution is the one you tune last, once the loss and the weight have taken recall and overlap as far as they go.

What each remedy breaks

No family is free, and a field guide that did not name the costs would be a brochure. Loss weighting relocates the failure from missed curves to hallucinated ones: turn the weight too high and precision collapses to the point that the exported curve is buried in false trace,

and the symptom is a model with excellent recall that a petrophysicist still cannot use. Region-based losses have their own brittleness at the extreme, where a one-pixel curve makes the overlap statistic dominated by sub-pixel registration rather than by whether the curve was found, so the IoU can look poor on a mask that reconstructs into a nearly perfect curve; the number understates the model and can mislead a team into over-correcting. Synthetic balancing carries the deepest risk of all, which is that the synthetic distribution is not the real one: if the generated curves do not span the shapes, crossings, and degradations of real scanned logs, the precision it buys is precision on a world the model will not meet in production, and the gap shows up only on real data. Each remedy trades its own failure for the one it fixes, which is another reason to stack all three: the families partly cover each other's weaknesses, so the stack is more robust than the sum of its parts suggests.

WHAT TO CARRY INTO THE NEXT THIN-STRUCTURE PROBLEM

1. Under sub-**2** percent foreground the three axes recall, precision, and intersection-over-union move independently, so no single imbalance remedy is enough and a usable curve is a stacking decision, not a selection decision.
2. Loss weighting is the first move: a positive-class weight of **42**, set by the **97** percent background ratio, lifts the binary-stage curve recall to **0.96** and **0.97**. It buys recall and charges precision, and the price shows up as a sourced number one family down.
3. Region-based losses are the second move, and the run where curve overlap and curve precision are measured together. Across **5** evaluated objectives, the multiclass Dice family lifts curve IoU off the floor toward **0.26** and **0.21** against a background at **0.94**, while curve precision on that same Dice run is a low **0.41** and **0.36**.
4. Synthetic balancing is the third move: genuine synthesis, not duplication, pays back the precision the class weight spent, because it gives the optimiser real positives to learn from rather than memorised copies of the scarce few.
5. The combined stack, not any single family, reached the study peak IoU of **0.51**. Grade the stack in deliverable space, on the error left on the exported curve, and watch the failure each remedy introduces.

Limitations

This guide is a survey of a mitigation stack on one problem, and its boundaries are worth stating so the reading stays honest.

The three families are treated as separable moves, and in the instrument they toggle independently, but in the real training runs they were not fully orthogonal. The class weight, the region-based loss, and the synthetic corpus were tuned together, so the clean attribution of one axis to one family is a teaching decomposition of an intertwined process, not a controlled ablation in which each family was isolated with the other two held fixed. A complete study would run that full factorial, and we have not.

The sourced numbers are stage-specific, and the guide keeps each one attached to the run that produced it. The recall of 0.96 and 0.97 is the binary-stage weighted-BCE run, which is the only metric that stage logs. The curve precision of 0.41 and 0.36 together with the IoU figures of 0.94, 0.26, and 0.21 are the multiclass Dice stage, one self-consistent set from a single training configuration. The frontier board places all three corners on one canvas for the argument's sake and labels each corner with its run, but the recall corner and the precision-and-IoU corners are measurements from two related configurations rather than a single joint readout, and the board never pairs a binary-stage recall with a multiclass precision as if they shared an operating point. The peak IoU of 0.51 is the best figure anywhere in the study, reached at the combined operating point, and it should be read as the ceiling the stack reached on this data rather than a number any one family produces.

The single-move partial positions on the board and the usable-curve contour are illustrative geometry. They are drawn to show the direction each family pushes the operating point and to land the combined point on the sourced numbers; they are not logged per-ablation sweeps, and the contour is a threshold for the argument, not a measured decision boundary.

Finally, the synthetic-balancing family carries an assumption we cannot fully discharge here: that the procedurally generated logs span the distribution of real scanned logs closely enough that the precision it buys transfers to production. We have evidence that it does on our data, but the guarantee is only as strong as the generator, and a team adopting this stack on a different corpus has to validate that its synthetic distribution matches its real one before trusting the precision the family reports.

GLOSSARY

Deliverable curve — The one-dimensional curve exported to CSV that a petrophysicist opens, reconstructed from the segmentation mask by post-processing. The only error that ultimately matters is the error on this curve, not on the intermediate mask.

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 under two percent of the frame and the background is around ninety-seven percent, and that scarcity is what makes every mitigation decision consequential.

Loss weighting — The algorithm-level remedy of multiplying the rarer class in the loss so a mistake on it counts for more. The binary stage used a positive-class weight of forty-two, meaning a missed curve pixel was charged as much as forty-two stray background pixels.

Mitigation stack — The combination of loss weighting, a region-based loss, and synthetic balancing applied together. The field-guide claim is that a usable curve under sub-two-percent foreground is a property of the stack, not of any single family within it.

Positive-class weight — The scalar multiplier applied to the foreground class in a weighted loss. Set near the background-to-foreground ratio, it restores parity in how much total gradient each class contributes; the sourced value here is forty-two at a ninety-seven percent background split.

Precision-recall-IoU frontier — The three-axis operating surface a thin-structure segmenter lives on. Recall is the share of true curve pixels

found, precision is the share of predicted curve pixels that are correct, and intersection-over-union is the overlap of predicted and true regions. Each mitigation family moves a different axis.

Region-based loss — A loss computed on the overlap between predicted and true foreground regions rather than pixel by pixel. Dice and its generalisation Tversky are the family; they normalise by foreground area, so a tiny target is not drowned by the background the way a raw pixel-wise loss would be.

Synthetic balancing — The data-level remedy of generating new minority examples so the optimiser trains against a higher genuine foreground share. Distinct from oversampling, which duplicates existing examples; synthetic balancing produces fresh, varied positives the model has not seen.

References

1. 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
2. Buda, M., Maki, A., Mazurowski, M. A. (2018). *A systematic study of the class imbalance problem in convolutional neural networks.* Neural Networks. https://arxiv.org/abs/1710.05381
3. 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
4. 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
5. 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
6. 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

7 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

8 Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. (2002). *SMOTE: Synthetic Minority Over-sampling Technique.* Journal of Artificial Intelligence Research. https://www.jair.org/index.php/jair/article/view/10302

9 Cui, Y., Jia, M., Lin, T.-Y., Song, Y., Belongie, S. (2019). *Class-Balanced Loss Based on Effective Number of Samples.* CVPR. https://arxiv.org/abs/1901.05555

10 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 per-stage training configurations, the full five-loss region-based sweep with per-curve numbers, the procedural-generation recipe behind the synthetic corpus, and the joint operating point where the peak intersection-over-union of 0.51 was reached.

EARTHSCAN

Class Imbalance in Thin-Structure Segmentation: A Field Guide

Published January 2023. Generated 2026-07-22.
https://www.earthscan.io/whitepapers/class-imbalance-thin-structure-segmentation-field-guide