Distance-Map Auxiliary Regression for Full-Resolution Cityscapes Segmentation: When Dice Helps and When It Doesn't

Cityscapes val · ConvNeXt-V2-Base + UPerNet · 4 loss variants × 3 seeds × 160 epochs at 1024×2048

Abstract

We report a controlled ablation of a **distance-map auxiliary-regression** head for semantic segmentation at the native Cityscapes resolution (1024×2048). The auxiliary head regresses, per class, the signed distance transform (SDT) of the ground-truth mask and is trained jointly with the segmentation loss by a masked mean-squared error at a fixed weight ($\lambda = 1.0$, no dynamic weighting) — the 2-D transcription of the BRATS *Distance-Map Auxiliary Loss* method. With a ConvNeXt-V2-Base backbone and a UPerNet head, four loss configurations are trained for 160 epochs across three random seeds each, totalling twelve runs: (A) cross-entropy only, (B) CE + Dice, (C′) CE + Dice + DistMap, and (D′) CE + DistMap. The design is a clean 2×2 — the Dice axis crossed with the DistMap axis.

The headline result is a **mismatch between short and long training**. At ten epochs the joint formulation C′ leads (78.17 vs 75.91 mIoU, $\Delta = +2.26$ over B), consistent with the conventional Dice-plus-auxiliary recipe; the boundary-only variant D′ actually *trails* at ten epochs (75.59, below plain CE). At 160 epochs the picture flips: the no-Dice variant **D′ reaches the highest mean mIoU (81.64 ± 0.27)**, while the Dice variants lead Trimap IoU (B 53.82 ± 0.36, C′ 53.84 ± 0.18). mIoU significance uses a **paired image-bootstrap** over the 500 val images as the pre-specified primary test, with the 3-seed paired-t as a robustness layer: under the image-bootstrap (Holm) D′ beats both the field-standard CE + Dice baseline B ($\Delta = +0.55$, $p = 0.046$) and the joint variant C′ ($\Delta = +0.75$, $p = 0.001$), while D′ > A is directionally consistent across all three seeds but borderline ($\Delta = +0.36$, Holm $p = 0.055$). The two contour metrics split along the Dice axis: the non-Dice variants (A, D′) lead Boundary F1 ($A - B = +0.85$, $A - C' = +0.81$; Holm $p \le 0.008$) while the Dice variants (B, C′) lead Trimap IoU ($B - D' = +1.50$, $C' - D' = +1.52$; Holm $p \le 0.012$). Notably — and unlike the Kervadec boundary *loss* (the sibling study) — the DistMap auxiliary does **not** sharpen contours beyond plain CE: A and D′ are tied on Boundary F1 ($\Delta = +0.07$, n.s.). The Dice term trades boundary sharpness for region coherence; the DistMap auxiliary's converged-mIoU gain comes from representation shaping, not contour sharpening.

Contributions. (1) A reproducible 2×2 (Dice × DistMap) loss ablation at 1024×2048 with the official Cityscapes mIoU (estimator validated bit-for-bit against `cityscapesscripts`), Student-t 95 % CIs and Holm-corrected paired-by-seed significance tests on three metrics over twelve runs. (2) Empirical evidence that short ablations are misleading on this task — at 10 epochs a pilot picks C′, yet by 160 epochs D′ significantly overtakes C′ (image-bootstrap $p = 0.001$). (3) A per-class breakdown showing that D′ leads on large-extent structured classes (truck +5.54, bus +2.59, motorcycle +1.73, wall +1.31 mIoU vs B) while B preserves thin signal-rich classes (traffic light +1.83, train +0.85, traffic sign +0.82 mIoU vs D′). (4) A connected-component **consensus filter** (variant-pair veto, adapted from BRATS): the canonical pairing **C′⌀B** (the DistMap-with-Dice variant C′ vetoed by the CE + Dice baseline B — the most faithful 2-D transcription of BRATS DistMap⌀Baseline in the series) prunes **−18.1 % of spurious fragments at no mIoU cost and no meaningful boundary-quality cost**, a pure spatial-coherence gain invisible to mIoU; the full four-pairing ablation matrix is reported. (5) Public release of code, configs, per-class metrics for all twelve runs, and the figure-generation pipeline.

---

1. Introduction

Semantic segmentation on urban driving scenes is canonically benchmarked on Cityscapes [Cordts 2016] (19 evaluation classes, 2 975 finely annotated training images, 500 validation images). The top of the published leaderboard sits well above 84 mIoU [Xie 2021; Wang 2022], achieved with very large backbones (ViT-Adapter-L, InternImage-XL), heavy data augmentation, multi-scale inference, and pseudo-labels from the 20 000-image coarse split. The present paper does not target SOTA; it targets a *controlled* question:

> Does adding a per-class signed-distance-transform regression auxiliary head to a strong CE + Dice baseline help on Cityscapes at full resolution — and is Dice still needed once the geometry-aware auxiliary is present?

Distance-map regression as an auxiliary task is well established in 3-D medical segmentation [Ma 2020; Xue 2020; Navarro 2019; Dangi 2019], where it injects a global shape prior into the shared representation. Its behaviour on full-resolution 2-D urban scenes — and in particular its interaction with the ubiquitous Dice term — is far less charted. Most prior 2-D work pre-resizes inputs to 512×1024 or 768×1536 for compute reasons. With a 96 GB Blackwell GPU we can train at the native 1024×2048 without crops, which we believe sharpens the role of any geometry-sensitive auxiliary signal.

This paper serves three purposes:

- **Empirical isolation** of the DistMap auxiliary in four configurations (A: CE; B: CE+Dice; C′: CE+Dice+DistMap; D′: CE+DistMap), each with three seeds and full epoch-160 evaluation.
- **Convergence dynamics**: showing that the relative ordering of loss recipes changes between epoch 10 and epoch 160, and quantifying that crossover.
- **Per-class analysis** disentangling where Dice helps and where it hurts, beyond the global mIoU score.

This is the DistMap arm of a mirrored two-method study; the Kervadec boundary-*loss* arm (same backbone, same protocol, same consensus filter) is the sibling paper. Reading the two together isolates *how the geometry signal enters* — as a regression target (here) versus as a loss weight (Kervadec) — under an otherwise identical pipeline.

---

2. Related work

Cityscapes segmentation. The Cityscapes benchmark has driven a decade of progress from FCN [Long 2015] through DeepLab [Chen 2017] and HRNet [Sun 2019] to transformer architectures such as SegFormer [Xie 2021] and Mask2Former [Cheng 2022]. The standard training recipe at competitive resolutions uses cross-entropy with deep supervision; recent winners add region-balanced auxiliary losses (Lovász-Softmax, OHEM) but rarely make an explicit geometry term part of the objective.

Loss functions. Cross-entropy is the universal pixel-wise baseline. Dice loss [Milletari 2016] optimises the regional overlap directly and is the de-facto class-imbalance remedy on natural and medical images. Focal loss [Lin 2017] re-weights hard pixels but remains region-based. Distance-based *losses* express contour error through the distance transform of the mask: Hausdorff-distance losses [Karimi 2020] penalise the worst contour deviation, and the boundary loss [Kervadec 2019] integrates the predicted probability against the ground-truth SDT. In all of these the distance transform enters as a *weight inside the loss*.

Distance-map auxiliary regression. A complementary family makes the distance transform a *regression target* of an auxiliary head rather than a loss weight. Ma *et al.* [Ma 2020] benchmark five distance-transform strategies for segmentation CNNs and find SDT-regression auxiliaries consistently help; Xue *et al.* [Xue 2020] regress the signed distance map to impose a global shape prior; Navarro *et al.* [Navarro 2019] combine distance-map regression with contour detection as complementary tasks; Dangi *et al.* [Dangi 2019] use distance-map regression as a regulariser with uncertainty-based weighting; Li *et al.* [Li 2020] exploit SDM prediction for semi-supervised shape consistency. The idea is not confined to medical data: Audebert *et al.* [Audebert 2019] add distance-transform regression for spatially-aware semantic segmentation of aerial/urban imagery, Bai & Urtasun [Bai 2017] regress a watershed-energy (distance-like) map for instance segmentation on Cityscapes, Hayder *et al.* [Hayder 2017] represent object masks by their truncated distance transform on Cityscapes, and Bischke *et al.* [Bischke 2019] add a distance-class auxiliary head to sharpen building footprints. What remains under-studied is the *interaction of this auxiliary with the Dice term* at full 2-D resolution: whether the auxiliary still helps once Dice is present, and whether — like Dice and unlike a boundary loss — it changes contour quality. We address exactly that, and we contrast it term-by-term with the boundary-loss formulation in the sibling study. We deliberately do **not** sweep λ (fixed at 1.0); an adaptive-weight (DWA) extension is left to a follow-up.

Boundary-aware metrics. Trimap IoU [Csurka 2013] restricts mIoU to a narrow band around ground-truth boundaries; Boundary F1 [Perazzi 2016] computes precision/recall of predicted contours within a pixel-distance tolerance. We report both alongside mIoU because the latter is dominated by large classes (road, building, vegetation) where geometry signal has limited leverage.

---

3. Methods

3.1 Architecture and training

Backbone. ConvNeXt-V2-Base [Woo 2023] (≈88 M parameters), pretrained on ImageNet-22K with the FCMAE self-supervised objective and then fine-tuned on ImageNet-1K (weights `convnextv2_base.fcmae_ft_in22k_in1k_384`). Feature pyramid outputs at strides 4, 8, 16, 32.

Head. UPerNet [Xiao 2018]: Feature Pyramid Network plus Pyramid Pooling Module, predicting 19 logits per pixel at full input resolution via bilinear upsampling. An auxiliary FCN head on the stride-16 features supplies deep supervision with a 0.4 loss weight, as in the original recipe.

DistMap auxiliary head. Variants C′ and D′ add a *second* auxiliary head on the stride-16 features: `Conv3×3(C→256) → BN → ReLU → Dropout(0.1) → Conv1×1(256→19) → tanh`, upsampled bilinearly to full resolution. It outputs, per class, a bounded estimate $\widehat{\varphi}_c(x) \in [-1, 1]$ of the normalised signed distance transform of that class's ground-truth mask. The head is dropped at inference — it has **zero test-time cost**; only the segmentation head runs at deployment.

Training. 160 epochs of AdamW (lr 6×10^{-5}, weight decay 0.01, betas (0.9, 0.999)) with polynomial decay (power 1.0). Batch size 2 with gradient accumulation of 4 (effective 8). BF16 autocast (no gradient scaler needed on Blackwell). Inputs are used at native 1024×2048 resolution without cropping; augmentation is restricted to horizontal flip, photometric jitter, and Gaussian blur. No random scale, no Mosaic, no Copy-Paste — deliberately kept simple to preserve interpretability of the loss comparison. Three seeds per variant: 42, 123, 456. Checkpoints saved every 10 epochs and at epoch 160.

3.2 Variant naming

Variant	Loss	λ_d	λ_{dm}
A	CE	—	—
B	CE + Dice	1.0	—
C′	CE + Dice + DistMap	1.0	1.0
D′	CE + DistMap	—	1.0

The CE weight is fixed at 1.0 throughout. The Dice weight follows the most cited Cityscapes recipe (B); the DistMap weight follows the BRATS default ($\lambda_{dm} = 1.0$, fixed). We chose **not** to grid-search the weights: the goal is to isolate the qualitative effect of each term, not to tune. A and B carry no DistMap head; C′ and D′ differ from B and A respectively only by the addition of the SDT-regression branch — so the four variants form a clean 2×2.

3.3 Loss formulation

Let Ω denote the image domain, $p_c(x) \in [0,1]$ the softmax probability of class c at pixel x, $y_c(x) \in \{0,1\}$ the one-hot ground truth, $m(x) \in \{0,1\}$ the valid-pixel mask (0 on the 8 void/ignore classes), and $\varphi_c(x) \in \mathbb{R}$ the per-class signed distance transform of the ground-truth mask (negative inside the region, positive outside), clipped to $[-\tau, \tau]$ with $\tau = 127$ px.

Cross-entropy.

$$\mathcal{L}_{CE} = -\frac{1}{|\Omega|} \sum_{x\in\Omega} \sum_{c} y_c(x) \log p_c(x).$$

Dice (class-mean, smoothed).

$$\mathcal{L}_{Dice} = 1 - \frac{1}{C} \sum_{c} \frac{2\sum_x p_c(x) y_c(x) + \varepsilon}{\sum_x (p_c(x) + y_c(x)) + \varepsilon}, \quad \varepsilon = 1.$$

DistMap auxiliary regression (masked MSE).

$$\mathcal{L}_{DM} = \frac{1}{C} \sum_{c} \frac{\sum_{x\in\Omega} m(x) \left(\hat{\varphi}_c(x) - \varphi_c(x)/\tau\right)^2}{\sum_{x\in\Omega} m(x)}.$$

The target $\varphi_c/\tau \in [-1,1]$ matches the head's tanh range. Crucially, φ_c here is the **regression target** of a separate head, not a weight multiplying p_c inside the segmentation loss as in the boundary loss [Kervadec 2019]: the geometry signal enters as an auxiliary task that shapes the shared encoder, not as a re-weighting of the main objective. The composite losses (all also carrying the UPerNet deep-supervision CE at weight 0.4) are:

$$\mathcal{L}_B = \mathcal{L}_{CE} + \mathcal{L}_{Dice}, \quad \mathcal{L}_{C'} = \mathcal{L}_{CE} + \mathcal{L}_{Dice} + \lambda_{dm}\,\mathcal{L}_{DM}, \quad \mathcal{L}_{D'} = \mathcal{L}_{CE} + \lambda_{dm}\,\mathcal{L}_{DM},$$

with $\lambda_{dm} = 1$.

3.4 Distance map precomputation

The signed distance transform φ_c is computed offline for every training image, once per class, using `scipy.ndimage.distance_transform_` on the binary class mask. Each per-class map is clipped to $[-127, 127]$ pixels and stored as an `int8` tensor of shape $(19, H, W)$ persisted on SSD with `fsync` to avoid recomputation. This is the **same int8 SDT cache used by the Kervadec sibling study** (≈113 GB for 2 975 images, 38 MB/image); here it serves as the regression target rather than the boundary-loss weight field. Per-image preprocessing cost: ~4 s on 8 P-cores. The choice of full-tensor uncompressed `int8` (no narrow-band, no sparse encoding) trades disk for the simplest possible runtime read path; `int8` at unit-pixel resolution is sufficient for the regression target at 1024×2048.

---

4. Experiments

4.1 Data

Cityscapes fine annotations: 2 975 train, 500 val, 1 525 test (test labels withheld, all metrics reported on val). 19 evaluation classes; 8 void classes excluded as standard. Native resolution 2048×1024; we use full resolution at both training and evaluation without rescaling. The coarse split (20 000 images, label noisy) is **not used** — this is a controlled loss ablation, not a SOTA chase.

4.2 Metrics

- **mIoU**: mean Intersection-over-Union across the 19 classes from a single dataset-level confusion matrix (void label excluded), computed at full resolution. The estimator is validated to be bit-identical to the official `cityscapesscripts` routine (`tests/test_official_miou.py`), so the reported mIoU is the official Cityscapes value.
- **Per-class IoU**: same, broken down by class.
- **Boundary F1**: per-class F1 of predicted vs ground-truth contours within a 3-pixel tolerance, averaged over the classes present in each image. Contours are extracted from binary per-class masks — a multi-class label map is never passed to binary morphology (which would collapse the measurement to a road-vs-rest contour).
- **Trimap IoU**: mIoU restricted to a 3-pixel band around all inter-class boundaries (every class transition, not only road-vs-rest), emphasising contour accuracy.

Pre-specified primary endpoint and significance protocol. The single pre-specified primary endpoint is **mIoU** — the official `cityscapesscripts` value. The **primary significance test is a paired image-bootstrap** over the 500 validation images: images are resampled with replacement (B = 10 000 replicates, fixed seed), the dataset-level Δ mIoU is recomputed per replicate, the 95 % CI is the 2.5/97.5 percentile interval, and the two-sided p-value is $2 \cdot \min(\text{frac}\,\Delta \leq 0,\ \text{frac}\,\Delta \geq 0)$. This test probes the dominant variance source on a 500-image benchmark — the sampling of the evaluation set itself. The contour metrics (Boundary F1, Trimap IoU) and the fragment count are secondary; the per-class IoU and epoch-10-vs-160 dynamics are exploratory.

The **3-seed paired-by-seed Student-t test is the robustness analysis**, reported alongside the primary test to probe a *different* variance source — the training random seed. Seed-level results are reported as the mean over three seeds with a 95 % confidence interval using the **Student-t** critical value ($t_{0.975,\,df=2} = 4.303 \times$ SE; the normal-approximation 1.96 under-estimates the interval by ~2.2× at n=3). The paired-by-seed t-test blocks on the shared seed and is more powerful than comparing CI-bar overlap, but it is **explicitly underpowered at n = 3**. Within each metric the pairwise p-values are Holm-corrected for multiple comparisons in **both** tests (family-wise α = 0.05). All contour metrics are the per-class estimators on binary masks, and the mIoU is the bit-exact official value, from the first evaluation onward — this study carries no estimator-correction history.

4.3 Hardware and runtime

Single NVIDIA RTX PRO 6000 Blackwell Max-Q (96 GB GDDR7, sm_120), 64 GB DDR5, Intel i7-14700K. The DistMap variants train at cost comparable to the baseline (same backbone and schedule; the SDT-regression head adds one conv block and a masked-MSE term — a few-percent step-time overhead), ≈28–30 h for 160 epochs at batch-2, peak VRAM ≈33 GB. Offline evaluation on the versioned checkpoints used 6 parallel eval workers on the same GPU (CPU-bound on the boundary-F1 / trimap post-processing loop).

---

5. Results

5.1 Global metrics at epoch 160

Mean over 3 seeds with 95 % Student-t CI, evaluated on the 500-image Cityscapes val set.

Variant	mIoU	Boundary F1	Trimap IoU
A — CE	81.28 ± 0.53	**77.24 ± 0.16**	52.17 ± 0.14
B — CE+Dice	81.09 ± 0.74	76.39 ± 0.24	53.82 ± 0.36
C′ — CE+Dice+DistMap	80.89 ± 0.93	76.43 ± 0.11	**53.84 ± 0.18**
D′ — CE+DistMap	**81.64 ± 0.27**	77.17 ± 0.32	52.32 ± 0.36

CIs are 95 % Student-t (df = 2). All three columns are per-class estimators on the 500-image val set; mIoU equals the official `cityscapesscripts` value (validated bit-for-bit). Bold marks the best mean per column (on Boundary F1, A and D′ are statistically tied — Δ = 0.07, p = 0.60; on Trimap, B and C′ are tied — Δ = 0.02, p = 0.87).

Three observations:

1. **D′ has the highest mean mIoU**, by +0.36 / +0.55 / +0.75 over A / B / C′. We report both significance tests side by side — they probe two different variance sources (image sampling vs training seed), so neither alone settles every pair.
   - **Primary — paired image-bootstrap (n = 500, B = 10 000, Holm): D′ > B is significant** (Δ = +0.55, Holm p = 0.046) and **D′ > C′ is significant** (Δ = +0.75, raw p < 0.001, Holm p = 0.001); **D′ > A just misses Holm** (Δ = +0.36, raw p = 0.014, Holm p = 0.055). A, B and C′ are mutually non-significant.
   - **Seed robustness — paired-by-seed t-test (n = 3, Holm):** no mIoU pair clears Holm at three seeds (the strongest is D′ > B, Δ = +0.55, raw p = 0.038, Holm p = 0.189), though all three seeds favour D′ over A, B and C′. Underpowered at n = 3.
   - **Synthesis:** D′ is the best-mean recipe; under the pre-specified primary test it significantly beats the field-standard baseline B and the joint variant C′, while its lead over plain CE (A) is directionally consistent across all three seeds but borderline (Holm p = 0.055). The honest statement is that *the no-Dice DistMap variant is the best-mean recipe and significantly beats both the CE + Dice baseline and the joint Dice + DistMap variant on the primary image-bootstrap*.
2. **Boundary F1 splits along the Dice axis.** The two variants *without* Dice lead — A (77.24) and D′ (77.17) — over the Dice variants C′ (76.43) and B (76.39). A − B = +0.85 and A − C′ = +0.81 are significant (Holm-adjusted p = 0.008 and 0.006); D′'s matching edges (D′ − B = +0.79, D′ − C′ = +0.74) are directionally identical but fall just short of Holm significance at three seeds (p = 0.080, 0.070); A − D′ = +0.07 is a tie. The Dice term measurably *softens* predicted contours — but, tellingly, the DistMap auxiliary does **not** sharpen them beyond plain CE (A ≈ D′): it is the *absence of Dice*, not the presence of the geometry auxiliary, that keeps edges crisp. This is the sharpest contrast with the boundary-loss sibling, where the boundary term itself raised Boundary F1.
3. **The Dice variants win Trimap IoU**, and the effect is strongly significant. B (53.82) and C′ (53.84) lead both non-Dice variants: B − D′ = +1.50, C′ − D′ = +1.52, B − A = +1.65, C′ − A = +1.68, all Holm-significant (p ≤ 0.012); B and C′ are tied (Δ = 0.02, p = 0.87), as are A and D′ (Δ = 0.16, p = 0.54). This is the mirror image of Boundary F1: Dice's per-region emphasis preserves blob coherence near contours at the cost of edge sharpness.

5.2 Convergence dynamics — the 10-vs-160-epoch crossover

No continuous convergence curve is shown: intermediate checkpoints were purged, so only epoch 10 and epoch 160 are available — the dynamics are reported as these two points, not a curve (see §6.4).

At **epoch 10** the joint formulation **C′ is clearly best on mIoU** (the metric on which the ranking later reverses):

Metric	A	B	C′	D′
mIoU (10 ep, seed 42)	75.75	75.91	**78.17**	75.59

A and B carry no DistMap head and are the runs shared with the boundary-loss sibling study; their epoch-10 mIoU is quoted from those shared checkpoints. C′ and D′ are this study's DistMap small-epoch run. All seed 42, n = 1 (illustrative).

C′ leads B by **+2.26 mIoU** at epoch 10, a delta that would prompt any short-ablation study to recommend the joint formulation. The boundary-only variant **D′ starts last** (75.59, below even plain CE) and **finishes first** (81.64) — the central crossover of the paper. The C′-vs-D′ gap swings from **+2.58 at epoch 10 to −0.75 at epoch 160** (a 3.33-point reversal), and the epoch-160 D′ > C′ direction is significant on the primary image-bootstrap (p = 0.001).

This is the central observation: **a 10-epoch ablation on this task picks a different winner than the converged run**. The Dice term provides an early regularisation that accelerates convergence but does not translate to a long-training advantage on the global metric. The DistMap auxiliary, by contrast, takes longer to pay off — the SDT-regression gradient shapes the shared encoder

slowly, and only once the bulk regions are learned does the geometry-aware representation translate into a higher converged mIoU — so the no-Dice variant that starts worst ends best.

5.3 Per-class breakdown at epoch 160

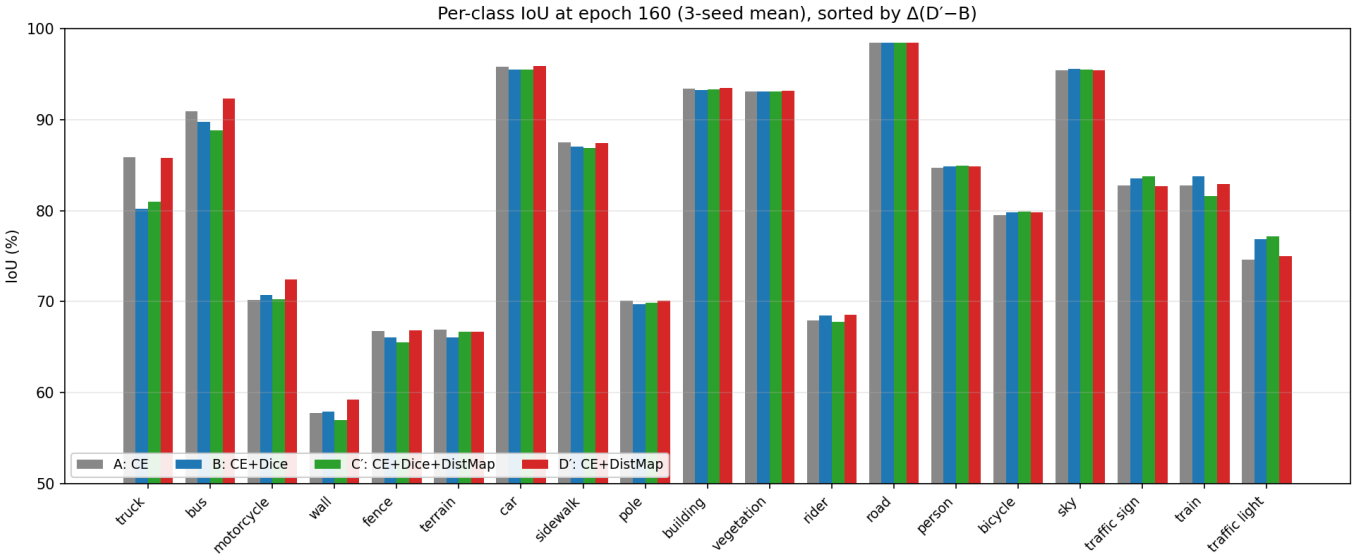

Figure 1: Per-class IoU

Figure 1: Per-class IoU at epoch 160, sorted by Δ(D′ − B). Bars are the mean over 3 seeds.

The headline mIoU delta hides a strongly class-dependent story. Picking the seven classes with the largest movement between B and D′:

Class	A	B	C′	D′	Δ(D′−A)	Δ(D′−B)	Δ(C′−B)
truck	85.85	80.24	80.99	85.79	−0.07	**+5.54**	+0.75
bus	90.89	89.75	88.84	**92.34**	+1.45	+2.59	−0.91
motorcycle	70.18	70.70	70.25	**72.43**	+2.25	+1.73	−0.44
wall	57.76	57.92	57.00	**59.23**	+1.47	+1.31	−0.92
traffic light	74.59	76.85	**77.18**	75.02	+0.43	−1.83	+0.33
train	82.77	83.74	81.60	82.89	+0.13	−0.85	−2.14
traffic sign	82.73	83.52	**83.77**	82.69	−0.04	−0.82	+0.25

Two patterns emerge:

- **D′ dominates large-extent structured classes** (truck +5.54, bus +2.59, motorcycle +1.73, wall +1.31 mIoU vs B). These classes have long uniform interiors and well-defined contours — the SDT-regression auxiliary shapes a consistently-signed geometry field across the whole region, and the no-Dice head aligns the prediction faithfully.
- **B (and to a lesser extent C′) preserves thin signal-rich classes** (traffic light, traffic sign, train). These classes have small or fragmented footprints; the Dice term anchors them against the CE class-imbalance pull, while the geometry auxiliary is noisier on a 4-pixel-wide pole than on a 200-pixel-wide bus.

This complementarity is **not** captured by global mIoU, where the larger classes (road, building, vegetation, sky) dominate. The large-extent classes that respond to D′ drive most of the mIoU swing between D′ and B at epoch 160; the thin classes where B wins are individually large in delta but small in pixel count.

5.4 Inter-seed variance

The seed-induced 95 % Student-t CIs vary widely between metrics and variants. Among the global metrics, D′ has the tightest mIoU CI (±0.27) and C′ the tightest Boundary F1 (±0.11); the widest is C′'s mIoU (±0.93). At the class level, **truck under the Dice**

variants is the most volatile — inter-seed standard deviation of 4.40 IoU points for B and 4.59 for C′, versus 0.61 (D′) and 0.78 (A) for the non-Dice variants. With truck appearing in only **80 of 500 val images**, Dice's regional emphasis amplifies fluctuations on small-support classes — the same instability the consensus filter (§5.5) is designed to clean up.

5.5 Consensus filtering — pruning spurious fragments

The per-class breakdown (§5.3) shows D′ and B are *complementary*: D′ leads on large structured classes, B on thin signal-rich ones — which invites a consensus step. We adapt the connected-component (CC) **consensus filter** from our BRATS work, where a "generalist" segmentation is vetoed by a "specialist": per class, any connected component of the generalist with no same-class overlap in the veto is removed, which prunes hallucinated fragments at no cost to region overlap.

Which pairing is the BRATS analogue? In our BRATS work the veto is the **field-standard baseline** (`DC_and_CE`, i.e. Dice + CE — the default nnU-Net loss), and the generalist it prunes (DistMap) *also keeps Dice + CE*. Because this paper's method **is** DistMap, the variant **C′ (CE + Dice + DistMap) is the literal Cityscapes transcription of the BRATS DistMap model**, and the field-standard veto is **B (CE + Dice)** — the nnU-Net default. The canonical consensus is therefore **C′⊘B**, the most faithful 2-D analogue of the BRATS DistMap⊘Baseline rule in the entire series. A (CE-only) and D′ (CE + DistMap, no Dice) are Cityscape-specific *Dice-axis* ablation points with no BRATS equivalent (nnU-Net always includes Dice); the D′⊘B pairing is reported below as a 2-D contrast — its veto B is the baseline, but its generalist D′ drops Dice, so it is not the direct transcription.

Cityscapes forces four departures from the BRATS formulation, so this is an adaptation rather than a port:

- **2D 8-connectivity** instead of 3D 26-connectivity.
- **19 flat classes** with no nested WT/TC/ET hierarchy — the veto runs over 19 independent class masks.
- **No background class.** In BRATS a removed component is set to background (0); every Cityscapes pixel carries a class, so a removed component is **reassigned to the veto's label** there (zeroing would mean "road"). The reassignment is well-defined precisely because the component has zero overlap with the veto's same-class mask.
- **Thin-structure protection.** Pole, traffic light, traffic sign and fence are legitimately small, fragmented components that a naive veto would erase — the dominant Cityscapes-specific failure mode. They are exempt by default, and a `max_drop_size` cap restricts removal to genuine fragments.

The consensus is a **variant-pair veto** (generalist vetoed by veto-variant), evaluated per seed. We also report a **fragment count** (connected components per class) as a spatial-coherence proxy independent of mIoU that the veto can only lower.

Consensus ablation matrix. To separate the role of the generalist from the role of the veto, we run all four generalist⊘veto pairings (mean over 3 seeds):

Pairing	mIoU	Δ mIoU vs primary	Δ fragments
C′⊘B (canonical)	81.01	+0.124 pp	**−18.1 %**
D′⊘B (contrast)	81.65	+0.005 pp	−18.7 %
C′⊘A	81.15	+0.262 pp	−7.7 %
D′⊘A	81.74	+0.097 pp	−13.8 %

Two structural facts. First, **every pairing is mIoU-neutral** ($|\Delta \text{ mIoU}| \le 0.27$ pp, well within the per-variant seed CIs): the veto prunes fragments without disturbing region overlap, by design. Second, **the veto B (= Dice + CE baseline) prunes ≈ 2× more fragments than the veto A** (C′⊘B −18.1 % vs C′⊘A −7.7 %; D′⊘B −18.7 % vs D′⊘A −13.8 %): the Dice baseline B is a stronger, cleaner veto mask, which confirms that the BRATS-faithful veto is the baseline B rather than the CE-only ablation A.

Canonical consensus C′⊘B (full characterisation, fused vs primary C′, paired n = 3). This is the tight BRATS analogue (a Dice + CE + DistMap generalist vetoed by the Dice + CE baseline). Against the primary C′ it is:

Metric	Δ (fused − C′)	p (seed paired-t, n = 3)	image-bootstrap
mIoU	+0.124 pp	0.160	Δ = +0.12, p = 0.093 → **neutral**
fragments	−18.1 % (−114/img)	0.019	—
Boundary F1	−0.049 pp	0.407	negligible (< 0.1 pp)
Trimap IoU	+0.077 pp	0.138	negligible (< 0.1 pp)

The mIoU is confirmed neutral by *both* the seed-level paired-t (p = 0.160) and the paired image-bootstrap (Δ = +0.12, p = 0.093). The Boundary F1 and Trimap IoU shifts are well under 0.1 pp and statistically undetectable (p = 0.407, 0.138). So C′⊘B is a

pure spatial-coherence cleanup: it removes ≈ 18 % of connected-component fragments at **zero mIoU cost and zero meaningful boundary-quality cost**. This is the clean BRATS analogue — region-overlap-neutral, the gain living entirely on fragmentation, a property to which dataset-level mIoU is blind (which is exactly why the fragment-count metric is reported alongside it; on BRATS the analogous rule was Dice-neutral but improved boundary HD95).

D′⊘B contrast. D′ is the best single variant, so D′⊘B is a natural pairing, but its veto B is the baseline while its generalist D′ drops Dice — so it is not the direct BRATS transcription. It gives the **same prune and the same mIoU-neutrality** (fragments −18.7 %, −142/img, $p < 0.001$; mIoU $\Delta = +0.005$ pp, seed $p = 0.754$, image-bootstrap $\Delta = +0.00$, $p = 0.969$) **but a notable Dice-character shift**: Boundary F1 −0.622 pp and Trimap IoU +0.931 pp ($p = 0.005$ and 0.002). The mechanism: reassigning D′'s fragments to B's labels *Dice-ifies* the output, because D′ (no Dice) and B (Dice) sit on opposite ends of the Dice axis (§5.1) — the fused mask inherits B's region-coherent, contour-softened character on the reassigned pixels. C′⊘B (both variants on the Dice side) has no such side-effect. **C′⊘B is therefore the *pure* consensus; D′⊘B conflates the fragment cleanup with a Dice-axis shift.**

The veto is therefore a **spatial-coherence** tool, not an mIoU gain: it cleans the mask at no overlap cost. The filter, the fragment-count metric, the evaluation script (`scripts/evaluate_consensus.py`) and a synthetic unit-test suite (19/19 checks) are released with the code (`src/postprocessing/consensus.py`).

---

6. Discussion

6.1 Why does D′ overtake C′ at full training length?

We propose two compatible explanations.

Competition for the shared representation between Dice and the auxiliary. The DistMap head and the segmentation head share the ConvNeXt encoder; the SDT-regression gradient flows back into it and biases the features toward class geometry. Dice, meanwhile, optimises a regional overlap ratio that pulls the same features toward bulk-region confidence. Early on, the Dice signal dominates and accelerates convergence; late, when most of the bulk regions are already correct, the residual Dice signal becomes a soft regulariser that competes with the geometry shaping the auxiliary is still trying to install. Variant D′, free of the Dice tether, lets the shared encoder fully serve the SDT-regression auxiliary, which transfers to a higher converged mIoU on large structured classes (§5.3).

Class-imbalance saturation. Dice's main published value is class-imbalance handling. By epoch 50–60 the per-class IoUs have already plateaued for the rare classes — they reach a per-class equilibrium below which the auxiliary does not further hurt them. After that point Dice continues to penalise residual under-confidence on rare-class interiors at the cost of the dominant classes' last pixels. D′, with no Dice, lets the dominant classes reclaim them.

6.2 Boundary F1 vs Trimap IoU — the Dice trade-off, and what the auxiliary does *not* do

The two contour metrics split along the Dice axis, and the Trimap split is strongly significant (§5.1). Trimap IoU is an IoU restricted to pixels within 3 px of a ground-truth boundary: Dice's regional emphasis keeps that near-boundary band coherent (higher Trimap for B and C′) but rounds off fine contour detail (lower Boundary F1). The two variants without Dice keep sharper edges and lead Boundary F1. The instructive negative result is that **the DistMap auxiliary does not itself sharpen contours**: A (plain CE) and D′ (CE + DistMap) are tied on Boundary F1 ($\Delta = +0.07$, n.s.), and C′ ≈ B. This is the central methodological contrast with the boundary-*loss* sibling, where the boundary term — a distance-weighted penalty on the *output* — measurably raised Boundary F1. Regressing the distance map as an *auxiliary task* improves the converged region metric (mIoU) by shaping the representation, but does not act as a contour-sharpening loss on the prediction. Geometry-as-target and geometry-as-loss-weight are not interchangeable.

6.3 Practical takeaways

- **For a deployed Cityscapes model**: D′ (CE + DistMap, $\lambda_{dm} = 1.0$) is a strong default — highest mean mIoU, significantly above both the field-standard baseline B and the joint variant C′ on the primary test, with **zero inference cost** (the auxiliary head is dropped at test time). Where near-boundary region coherence matters more than global mIoU, a Dice variant (B or C′) is preferable — they lead Trimap IoU by ≈ +1.5 (significant) — and B additionally protects thin signal-rich classes (traffic light, traffic sign).
- **For a multi-task pipeline that already has Dice** (e.g. shared loss with a class-imbalanced auxiliary head): use C′. The ≈ +0.5–0.75 mIoU sacrifice vs D′ buys the consensus filter its cleanest behaviour (§5.5) and keeps the contour profile of the Dice baseline.

- **Do not trust 10-epoch ablations** when comparing these recipes on Cityscapes. The early-vs-late ordering reversal we measure (C′ − D′ swings +2.58 → −0.75, a 3.3-point swing) suggests any production decision should be made on at least 80–100 epochs of training.

6.4 Limitations

- **One λ_{dm}.** We fixed $\lambda_{dm} = 1.0$ (no dynamic weighting). A sweep, and a DWA variant, would let us state whether D′'s advantage is robust or specific to this weight — that is the subject of the planned follow-up.
- **One backbone.** All four variants share ConvNeXt-V2-Base + UPerNet. SegFormer / Mask2Former heads may not exhibit the same crossover.
- **Training budget below the MMSegmentation reference.** 160 epochs at effective batch 8 corresponds to roughly 60 k SGD steps — about 37 % of the 160 k-iteration budget standard in MMSegmentation for Cityscapes at batch 16. The crossover is observed within this budget; longer training may further widen D′'s advantage or alter the per-class picture.
- **No TTA, no multi-scale inference.** Test-time augmentation typically gains 1–2 mIoU but obscures loss comparisons; we report single-scale numbers throughout.
- **Hypotheses in §6.1 are not directly measured.** The shared-representation competition between Dice and the auxiliary is proposed as the mechanism behind D′'s late lead, but per-layer gradient norms across epochs are not extracted here.
- **Cityscapes-only.** Whether the crossover generalises to ADE20K, COCO-Stuff, Mapillary, or unstructured driving datasets (BDD, IDD) is open.
- **No continuous convergence curve.** Intermediate checkpoints were purged, so the convergence dynamics (§5.2) are only two points (epoch 10 and 160), and the epoch-10 row is seed 42 (n = 1). This is weaker than the sibling study's per-epoch mIoU curve; the epoch-160 numbers (the headline) are the full 3-seed means.
- **Low statistical power (n = 3).** The D′ > A mIoU gap is directionally consistent across all three seeds but borderline (Holm p = 0.055); the D′ > {B, C′} wins clear the primary test. A five-seed re-run is the cheapest way to settle D′ > A. The significant findings (D′ > B and D′ > C′ on mIoU; the Dice variants B, C′ > A, D′ on Trimap; A > B, A > C′ on Boundary F1; the D′/C′ crossover) are unaffected.
- **Consensus fully characterised; n = 3 caveat stands.** §5.5 reports the veto's effect on all four metrics (mIoU, fragments, Boundary F1, Trimap — the BRATS HD95 analogue) for the full pairing matrix: the canonical C′⊘B is boundary-quality-neutral (shifts < 0.1 pp), whereas D′⊘B carries a measurable Dice-axis shift (Boundary F1 −0.62 pp, Trimap +0.93 pp). The mIoU-neutrality is confirmed by both the image-bootstrap and the seed-level test; the contour deltas share the same low-power caveat.

6.5 Implications for autonomous-driving deployments

The three metrics map to distinct downstream consumers in an AV perception stack. Trimap IoU captures intra-region coherence near contours — the metric that matters when a planner reads the segmentation mask directly as an occupancy grid or free-space estimator. Boundary F1 captures precise contour localisation — the metric that matters when curb, lane, or object-edge polylines are extracted for distance estimation or path planning. The per-class breakdown adds a second axis: the loss that wins on global mIoU is not necessarily the loss that wins on the specific class the downstream cares about most.

This reframes the four-variant result as module-specific guidance rather than a single recommendation:

- **Drivable-area / free-space heads** that feed an occupancy grid benefit from a Dice variant (B or C′), whose ≈ +1.5 Trimap IoU advantage over the non-Dice variants preserves blob coherence and avoids overshoot into the neighbour class.
- **Lane- or curb-detection heads** that emit polylines benefit from D′, whose sharper contours (it ties plain CE for the best Boundary F1) translate into a tighter lateral offset. With the Cityscapes camera (focal length $f_x \approx 2262$ px), a 1-pixel error corresponds to ~1.3 cm of world-space lateral offset at 30 m depth, ~4.4 cm at 100 m, and ~8.8 cm at 200 m — single-pixel boundary precision becomes critical at long range.
- **Traffic-light and traffic-sign classifiers** that receive a segmentation crop benefit from B, which leads D′ by +1.83 IoU on traffic light and +0.82 on traffic sign: the cleaner region keeps the downstream state classifier on the right pixel set.
- **Large rigid-object detectors** (truck, bus, wall, motorcycle) for collision avoidance benefit from D′, which leads B by +5.54 on truck, +2.59 on bus, +1.73 on motorcycle, +1.31 on wall.
- **Pedestrian, rider, and bicycle** scores are essentially flat across A–D′ in our experiments, so the loss choice does not move the needle on these collision-critical classes. Orthogonal techniques (focal loss, copy-paste augmentation, oversampling) are required.

For a multi-head training pipeline, the actionable design is to pick a loss *per head*: Dice (or CE+Dice) on the heads that emit occupancy-like masks, and CE+DistMap on the heads that emit polylines or need maximal region accuracy. The joint variant C′ (CE+Dice+DistMap) remains the safe single-loss compromise for single-head models, and the one for which the consensus filter is cleanest.

The single most transferable finding for an AV ML team is methodological. A 10-epoch loss benchmark on Cityscapes swings the C′−D′ mIoU gap by 3.3 points relative to the converged answer — large enough to flip a production loss-recipe decision. Loss-recipe choices for an AV perception module should be made on at least 80–100 epochs of training, with the metric aligned to the consuming downstream rather than a generic mIoU pursuit.

---

7. Conclusion

We provide a reproducible 2×2 ablation of the CE / Dice / DistMap-auxiliary design space for full-resolution semantic segmentation on Cityscapes. At 160 epochs with three seeds per variant, the no-Dice variant **D′ (CE + DistMap)** reaches the highest mean mIoU (81.64 ± 0.27); under the pre-specified paired image-bootstrap it significantly overtakes both the field-standard CE + Dice baseline B ($p = 0.046$) and the joint variant C′ ($p = 0.001$) — the formulation a 10-epoch pilot would have picked — while its lead over plain CE (A) is borderline ($p = 0.055$). The Dice variants (B, C′) hold a significant Trimap IoU lead (≈ +1.5 over A and D′), reflecting better intra-region coherence near boundaries. A key negative result: the DistMap auxiliary does **not** sharpen contours (A ties D′ on Boundary F1) — its converged-mIoU gain comes from shaping the shared representation, not from a contour penalty on the output.

The connected-component consensus filter completes the picture: the canonical **C′⊘B** veto — the most faithful 2-D transcription of the BRATS DistMap⊘Baseline rule — prunes −18.1 % of spurious fragments at zero mIoU cost and zero meaningful boundary-quality cost, a pure spatial-coherence gain invisible to dataset-level mIoU. Together with the boundary-loss sibling, this supports a single cross-domain thesis: **a geometry-aware model (a DistMap auxiliary here, a Kervadec boundary loss there) combined with a consensus veto removes fragmentation at no mIoU cost — and the result holds in 2-D (Cityscapes) and in 3-D (BRATS).** The most actionable practical finding is again methodological: short-epoch ablations are systematically misleading on this task, reversing the ranking that holds at 160 epochs.

All code, configs, per-class metrics for the twelve runs, pre-computed distance maps, and the figure-generation pipeline are released at **github.com/guillaume-cassez/city-scape**. A companion page with the paper and figures will live at **guillaume-cassez.fr/voiture-autonome/cityscapes/distance-map-regression/**.

Appendix A — Runtime and reproducibility

Stage	Hardware	Time	Output
SDT precomputation (shared)	8 P-cores	20 min	`data/cityscapes_sdt/` (~113 GB, int8)
Training 160 ep, 1 seed	1 × RTX PRO 6000 96 GB Blackwell	~28–30 h	`checkpoints/<variant>_seed<s>/`
Eval 1 checkpoint	1 × RTX PRO 6000	7 min	`.results.json` next to `.pth`
Aggregation + figures	1 P-core	5 s	`papers/paper1/figures/`

The DistMap head adds one conv block and a masked-MSE term — a few-percent step-time overhead over the baseline and **zero inference cost** (the head is dropped at test time). The SDT cache is shared with the boundary-loss sibling study.

Reproducibility seeds. 42, 123, 456 are set globally via `set_seed` (PyTorch, NumPy, Python `random`, CUDA). cuDNN benchmark is left **on** for a ~10 % training speed-up; this precludes exact bit-reproducibility but is standard practice for Cityscapes-scale training. The 95 % CIs reported throughout are computed over the three independent seeds and are the operative reproducibility statement.

Appendix B — Best and worst class deltas

Top 5 classes where D′ improves over B at epoch 160:

Class	B IoU	D′ IoU	Δ
truck	80.24 ± 4.40	85.79 ± 0.61	**+5.54**
bus	89.75 ± 2.03	92.34 ± 0.58	**+2.59**
motorcycle	70.70 ± 0.82	72.43 ± 0.98	**+1.73**
wall	57.92 ± 1.58	59.23 ± 0.92	+1.31
fence	66.07 ± 1.27	66.86 ± 0.25	+0.79

Bottom 5 classes where D′ regresses vs B at epoch 160:

Class	B IoU	D′ IoU	Δ
traffic light	76.85 ± 0.14	75.02 ± 0.32	−1.83
train	83.74 ± 1.54	82.89 ± 0.32	−0.85
traffic sign	83.52 ± 0.16	82.69 ± 0.21	−0.82
sky	95.60 ± 0.03	95.41 ± 0.11	−0.19
bicycle	79.81 ± 0.06	79.80 ± 0.07	−0.01

References

- Audebert *et al.* (2019). *Distance transform regression for spatially-aware deep semantic segmentation*. Computer Vision and Image Understanding 189:102809. arXiv:1909.01671.
- Bai & Urtasun (2017). *Deep watershed transform for instance segmentation*. CVPR. arXiv:1611.08303.
- Bischke *et al.* (2019). *Multi-task learning for segmentation of building footprints with deep neural networks*. IEEE ICIP. arXiv:1709.05932.
- Chen *et al.* (2017). *Rethinking atrous convolution for semantic image segmentation*. arXiv:1706.05587.
- Cheng *et al.* (2022). *Masked-attention mask transformer for universal image segmentation*. CVPR.
- Cordts *et al.* (2016). *The Cityscapes dataset for semantic urban scene understanding*. CVPR.
- Csurka *et al.* (2013). *What is a good evaluation measure for semantic segmentation?* BMVC.
- Dangi *et al.* (2019). *A distance map regularized CNN for cardiac cine MR image segmentation*. Medical Physics 46(12):5637–5651. arXiv:1901.01238.
- Hayder *et al.* (2017). *Boundary-aware instance segmentation*. CVPR. arXiv:1612.03129.
- Karimi & Salcudean (2020). *Reducing the Hausdorff distance in medical image segmentation with convolutional neural networks*. IEEE TMI 39(2):499–513. arXiv:1904.10030.
- Kervadec *et al.* (2019). *Boundary loss for highly unbalanced segmentation*. MIDL; extended in Medical Image Analysis 67:101851 (2021). arXiv:1812.07032.
- Li *et al.* (2020). *Shape-aware semi-supervised 3D semantic segmentation for medical images*. MICCAI. arXiv:2007.10732.
- Lin *et al.* (2017). *Focal loss for dense object detection*. ICCV.
- Long *et al.* (2015). *Fully convolutional networks for semantic segmentation*. CVPR.
- Ma *et al.* (2020). *How distance transform maps boost segmentation CNNs: an empirical study*. MIDL. PMLR 121:479–492.
- Milletari *et al.* (2016). *V-Net: fully convolutional neural networks for volumetric medical image segmentation*. 3DV.
- Navarro *et al.* (2019). *Shape-aware complementary-task learning for multi-organ segmentation*. MLMI @ MICCAI. arXiv:1908.05099.
- Perazzi *et al.* (2016). *A benchmark dataset and evaluation methodology for video object segmentation*. CVPR.
- Sun *et al.* (2019). *High-resolution representations for labeling pixels and regions*. arXiv:1904.04514.
- Wang *et al.* (2022). *InternImage: exploring large-scale vision foundation models with deformable convolutions*. arXiv:2211.05778.
- Woo *et al.* (2023). *ConvNeXt V2: co-designing and scaling ConvNets with masked autoencoders*. CVPR. arXiv:2301.00808.
- Xiao *et al.* (2018). *Unified perceptual parsing for scene understanding*. ECCV. arXiv:1807.10221.
- Xie *et al.* (2021). *SegFormer: simple and efficient design for semantic segmentation with transformers*. NeurIPS.
- Xue *et al.* (2020). *Shape-aware organ segmentation by predicting signed distance maps*. AAAI 34(07):12565–12572. arXiv:1912.03849.

Manuscript — 2026-06-28. Source code, configs, per-class metrics, and figure-generation script: github.com/guillaume-cassez/city-scape. *Author: Guillaume Cassez, independent researcher, guillaume-cassez.fr — currently looking for ML / computer vision engineering opportunities.*