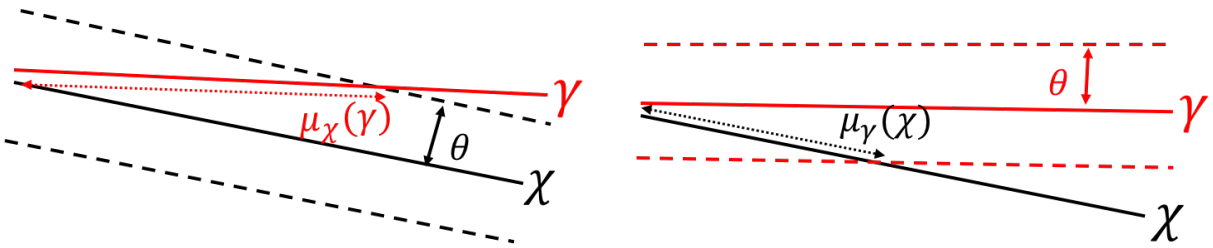
A. Supplementary Material

A.1. Metrics

In Section 4 we used the notions of *completeness*, *correctness* and *quality* [32] to compare the performance of different delineation methods. They are better suited for this task than traditional pixel-wise metrics such as precision and recall because a small shift in the prediction may lead to zero precision and recall in this region. Contrary to this, completeness, correctness and quality operate on skeletonized delineations and they relax the notion of precision and recall to include region around the skeleton's neighborhood.

Denote the set of ground truth skeleton pixels by $\mathcal{X}$ and the set of prediction skeleton pixels by γ. We define the subset of prediction skeleton pixels that *match* the ground truth as $\mu_{\mathcal{X}}(\gamma)$. Similarly, the subset of ground truth skeleton matching prediction is $\mu_{\gamma}(\mathcal{X})$. The matching is defined in terms of a threshold θ on a distance $d()$ to the nearest point of the other set $\mu_B(A) = \{a \in A | \exists b \in B,\ d(a,b) < \theta\}$. The correctness, a rough equivalent of precision, is then defined as $\frac{|\mu_{\mathcal{X}}(\gamma)|}{|\gamma|}$. Similarly, $completeness = \frac{|\mu_{\gamma}(\mathcal{X})|}{|\mathcal{X}|}$ and $quality = \frac{|\mu_{\mathcal{X}}(\gamma)|}{|\gamma| - |\mu_{\gamma}(\mathcal{X})| + |\mathcal{X}|}$.



(a) (b)

Figure 9: (a) Matching ground truth with prediction skeleton. (b) Matching prediction with ground truth skeleton.

A.2. Ablation Study

We performed an extended ablation study and report the results in Table 5 below. The left column shows that using any modern architecture to compute the topology loss yields an improvement. The right column shows that the results are relatively insensitive to the choice of the μ parameter of Eq. 3 that controls the relative influence of the topology loss and binary cross-entropy, as long as it remains in the range 0.001 to 10. The influence of number of layers used to compute the topology loss and number of refinement iterations is presented in Table 1 of the paper.

Feature extractor	F1 score	μ parameter	F1 score
None	0.7952	**0**	0.7952
AlexNet	0.8053	**0.001**	0.8059
VGG19 with random initialization	0.8037	**0.01**	0.8142
VGG19	0.8140	**0.1**	0.8140
VGG16	0.8106	**10**	0.8058
ResNet	0.8190	**inf**	0.7987

Table 5: Ablation study. F1 scores for **OURS-NoRef** method on the **EM** dataset when using different architectures to compute the topology loss of Eq.(2) (left) and when using different weighting parameter μ between the topology loss and binary cross-entropy (right).