Fusing Forces: Deep-Human-Guided Refinement of Segmentation Masks

By Rafael Sterzinger, Christian Stippel, and Robert Sablatnig

KEY OVERVIEW

- **Previous Work:** photometric-stereo-scanning with deep learning for Etruscan mirror segmentation; high accuracy, some refinement needed
- **Contribution:** interactive, human-in-the-loop approach to **reduce manual refinement efforts by 75%** with an **intermediate quality difference of 26%**
- **Data Availability:** public access to code and dataset provided

METHODOLOGY

1. Obtain statistics of human annotation patterns to simulate realistic stroke widths
2. Simulate two operations:
   A. Adding missing segments
   B. Erasing superfluous segments
3. Refinement network improves segmentation using: original depth map, initial prediction, human input
4. Process repeats iteratively, focusing on worst patches first

OUTLOOK

- **Limitations:** Refinement network considers guidance only locally (near the annotation, within same patch), but could be used globally
- **Future Work:**
  - Add ability to learn online from guidance, e.g., via Gaussian processes for global refinements
  - Incorporate active learning to identify patches of uncertainty and maximize performance gain

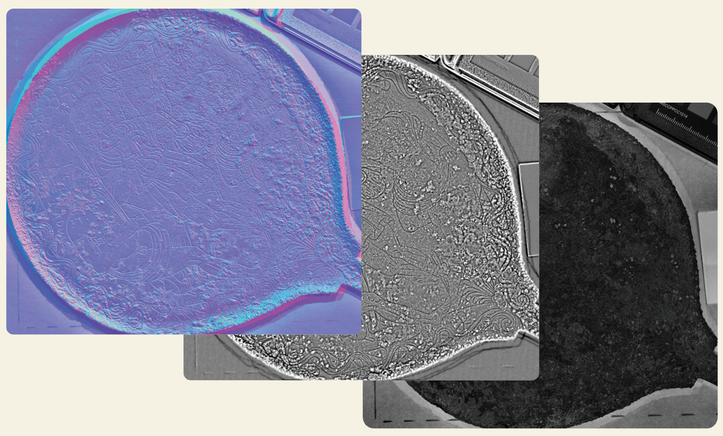
Etruscan Mirrors

- **Overview:** 3,000+ specimens; ~60 in AUT; Published in Corpus Speculorum Etruscorum; Backside features engravings depicting Greek mythology
- **Challenges:** Labor-intensive tracing; Damage leads to subjectivity; Limited available data

A typical Etruscan mirror: fine drawings of Greek mythology or Etruscan inscriptions adorn its backside

Dataset

- 59 Etruscan mirrors from Austria; 53 from Kunsthistorisches Museum Wien
- **Total: 29 annotations; 19 backs & 10 fronts**



Exemplary normal, depth, and albedo map obtained via photometric stereo

Initial Segmentation [1]

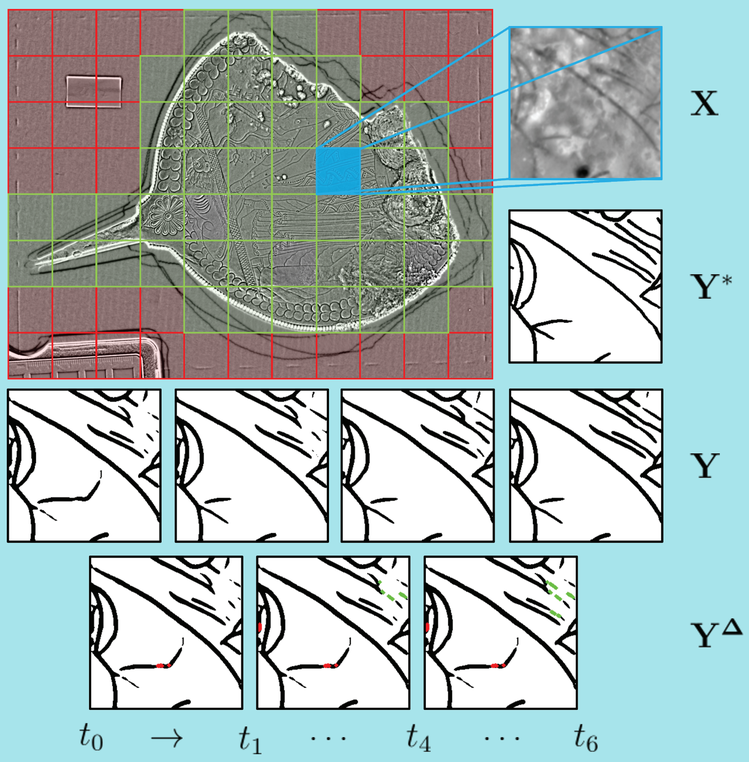
- **Architecture:** UNet with an EffcientNet encoder using solely depth maps; patch-level training with augmentations: rotations, flips, random crops; Generalized Dice Loss [2]
- **Performance Measure:** pseudo-F-Measure [3]

$$\mathrm{pFM} = \frac{2 \times \text{p-Recall} \times \text{Precision}}{\text{p-Recall} + \text{Precision}}$$

- **Result:** outperforms classical binarization algorithms & sometimes en-par with human annotator

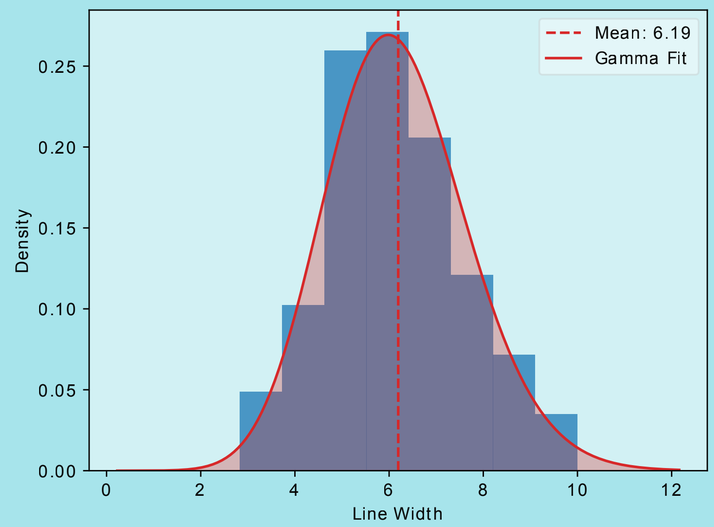
[1] Sterzinger R., Brenner S., Sablatnig R.: "Drawing the Line: Deep Segmentation for Extracting Art from Ancient Etruscan Mirrors" (2024, ICDAR)
[2] Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M.: "Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations" (2017, DLMIA)
[3] Pratikakis, I., Gatos, B., Ntirogiannis, K.: "ICFHR 2012 Competition on Handwritten Document Image Binarization" (2012, ICFHR)

Initial segmentation result, illustrating need for refinement



Simulation of Human Interaction

- **Statistics:** (1) obtain Euclidean distance in pixels to closest non-mask pixel, (2) skeletonize mask to obtain center, (3) fit distribution to sample realistic widths
- **Operations:** (1) find worst patch, (2) obtain superfluous and missing skeleton, (3) obtain longest lines segments for add & erase and pick the longer one



Distribution of stroke widths: after removing outliers with the two-sigma rule, a Gamma distribution fits the data best

Ablation Study

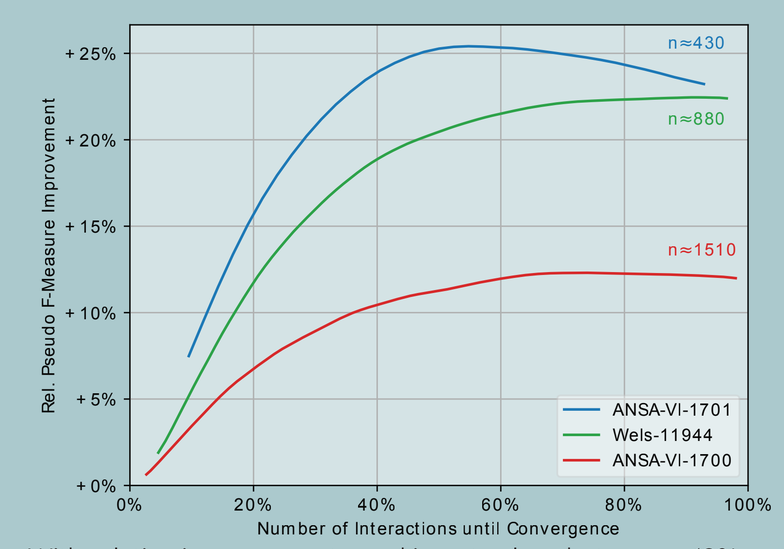
Evaluating the impact of different inputs: iterating over the initial prediction again does not cause improvement whereas providing human guidance does, however, providing both is best.

Input Modality		IoU	pFM	pFM_Δ
Init. Prediction [1]	-	32.86	49.28	–
Prediction	$\mathbf{Y}$	32.72	49.27	–
Interaction	$\boldsymbol{\Delta}$	$35.83 \pm .1$	$53.60 \pm .2$	$+5.8 \pm .23\%$
Both	$\mathbf{Y}, \boldsymbol{\Delta}$	$36.04 \pm .2$	$53.44 \pm .3$	$+5.5 \pm .55\%$

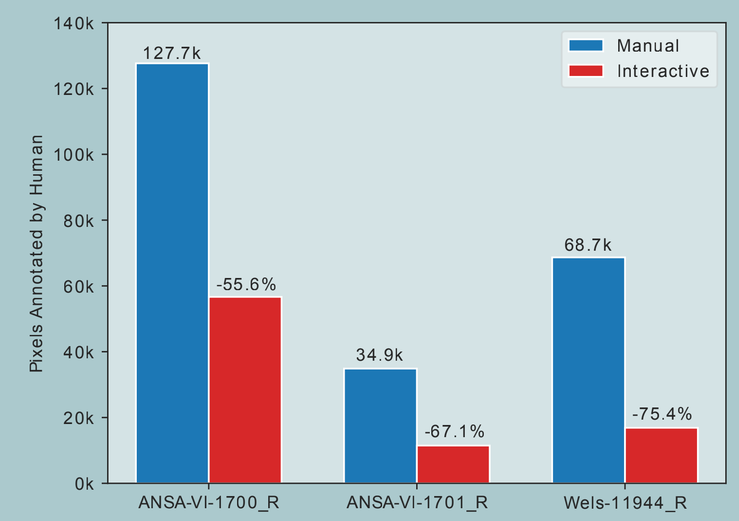
Evaluating the impact of adding and erasing: employing both operations will result in the highest delta of pseudo-F-Measure.

Interaction		IoU	pFM	pFM_Δ
Only Erasing	$\boldsymbol{\Delta}^-$	$38.25 \pm .06$	$58.92 \pm .07$	$+1.9 \pm .12\%$
Only Adding	$\boldsymbol{\Delta}^+$	$55.19 \pm .13$	$73.55 \pm .11$	$+8.4 \pm .16\%$
Both	$\boldsymbol{\Delta}$	$58.41 \pm .28$	$76.56 \pm .16$	$+12.3 \pm .17\%$

Results



With relative improvements peaking at values between +12% and +26%, our approach leads to better annotations earlier



At convergence, our approach requires fewer annotated pixels to reach equal performance, reducing workload from 56% to 75%

Conclusion

Our *human-in-the-loop* refinement drastically accelerates the annotation of complex engravings, e.g., found on Etruscan mirrors. By training a *deep neural network* to complete simulated human guidance, our method *reduces manual input* by up to 75%, with quality differences of up to 26%.

Paper | Code | Contact

This poster is presented at the International Conference on Pattern Recognition - ICPR 2024, Kolkata, India.
Corresponding Author: Rafael Sterzinger, rafael.sterzinger@tuwien.ac.at

This project has received funding from the Austrian Science Fund/Österreichischer Wissenschaftsfonds (FWF) under grant agreement No. P 33721.

CVL Computer Vision Lab
TU WIEN TECHNISCHE UNIVERSITÄT WIEN