Evaluating Diffusion-based Super-Resolution for Trustworthy Quantitative Metallography

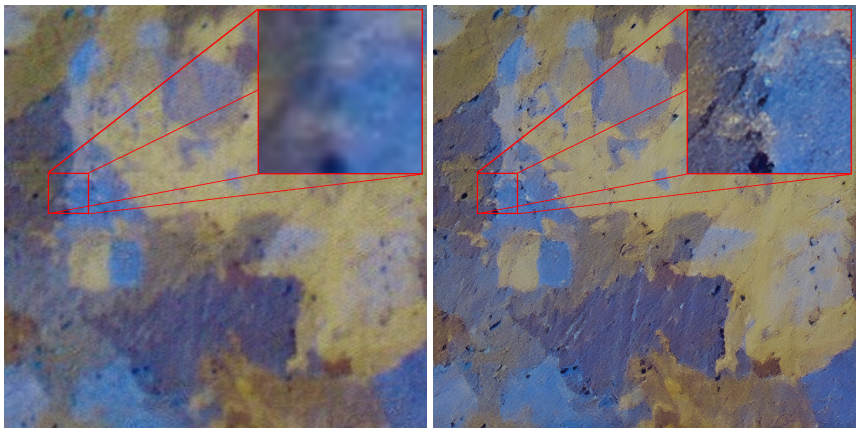
Boaz Meivar
Tel Aviv University
Israel
boazmeivar@mail.tau.ac.il

Inbal Cohen
Tel Aviv University
Israel
inbalc2@mail.tau.ac.il

Ofer Beeri
IAEC
Israel
oferb@iaec.gov.il

Shai Avidan
Tel Aviv University
Israel
avidan@eng.tau.ac.il

Gal Oren
Stanford University, Technion
United States
galoren@stanford.edu



(a) Original (b) Super-Resolution

Figure 1: **Original vs. Super-Resolution**. The left panel shows a TBM metallography image (original resolution), and the right panel shows its super-resolved version generated using our diffusion pipeline (OSEDiff with a metallography-specific prompt). The SR image exhibits enhanced clarity, making grain boundaries more continuous and easier to discern for analysis.

Abstract

Super-resolution (SR) holds promise for improving metallographic analysis, but diffusion-based methods raise concerns about hallucinated structures that could bias quantitative results. We present the first systematic study of diffusion SR in quantitative metallography. Using OSEDiff with a fixed domain prompt, we generate a fourfold super-resolved version of the Texture Boundary in Metallography (TBM) dataset (SR–TBM) and train uncertainty-aware edge detectors on both original and SR images. Expert audit confirmed that SR–TBM introduces no spurious grain boundaries, establishing that diffusion SR can be trusted under domain-guided prompting. At the same time, models trained on SR–TBM achieve a 47% reduction in grain-size error (Heyn intercept metric) compared to models trained on original TBM, surpassing prior baselines including MLOgraphy and AutoSAM. These results demonstrate that diffusion SR, when guided appropriately,

AI for Accelerated Materials Discovery (AI4Mat) Workshop, NeurIPS 2025.

both preserves scientific validity and enhances performance in grain-size estimation. We release SR–TBM and code to encourage reproducible, physics-aware evaluation of generative enhancement methods in materials science. [1]

1 Introduction

Quantitative metallography (QM) depends on extracting reliable grain-size statistics from microstructural images, yet this task is hindered by blurred, noisy, and low-resolution boundaries. Fine grains often appear fragmented or ambiguous, making automated detection difficult. Conventional vision approaches struggle: general-purpose models such as the Segment-Anything Model fail on purely textural images [5], while specialized pipelines like MLOgraphy apply heavy post-processing that can fracture true boundaries or erase uncertain segments [8]. Even state-of-the-art edge detectors degrade when resolution is reduced or context is limited [7, 4, 6, 2].

The Texture Boundary in Metallography (TBM) dataset was introduced to confront this challenge [9]. TBM consists of high-resolution images with intricate textures and incomplete boundaries, deliberately designed to test detection under adverse conditions. Standard pixelwise metrics such as IoU or Dice prove inadequate here, since boundaries are rarely perfectly closed. Instead, the physics-informed Heyn intercept method – long used in materials science (ASTM E112 [1]) – offers a more faithful evaluation by estimating grain size directly, thus aligning model performance with scientific goals [9].

One possible route forward is generative super-resolution (SR), which can enhance boundary continuity and visibility. But diffusion-based SR introduces a crucial concern – hallucinated structures could compromise downstream measurements. The open question is whether SR can genuinely improve metallographic analysis without introducing misleading artifacts.

In this paper, we address this question through the first systematic study of diffusion-based SR for metallography. Using OSEDiff [11] – a one-step diffusion network for real-world image super-resolution – with a fixed domain prompt, we super-resolve TBM images fourfold to create SR–TBM, then evaluate identical uncertainty-aware edge detectors trained on original and super-resolved data. The SR-based model achieves markedly more accurate grain-size estimates – 47% lower error – without generating false boundaries, as confirmed by expert audit. It also surpasses existing baselines such as MLOgraphy and AutoSAM, establishing a new state-of-the-art [8, 10]. By releasing SR–TBM and code, we provide a reproducible framework for testing trustworthy enhancement at the intersection of AI and materials science.

2 Related Work

Grain-Boundary Detection in Metallography. Delineating grain boundaries in microstructural images is challenging because boundaries are often faint, incomplete, or rely on surrounding texture context [7, 4, 6, 9, 3]. The TBM benchmark [9] encapsulates these difficulties, providing a testbed for algorithms under realistic conditions of fragmented boundaries and high texture complexity. Traditional image segmentation or edge detection approaches struggle on TBM. MLOgraphy [8] addresses the task with a U-Net–based predictor followed by aggressive post-processing to produce closed contours. While effective on clean images, this method suffers on TBM due to its use of small image patches (losing global context) and the removal of uncertain edges which might actually be true boundaries. Subsequent work has moved towards using larger context windows and accepting partial (open) boundaries during training to reduce dependence on post-processing, thereby improving performance on datasets like TBM [3]. The Segment Anything Model (SAM), a foundation model for general segmentation, was also tested on metallography via an automated variant called AutoSAM [10, 3]. AutoSAM modifies SAM's prompt encoder to function without user input, producing full-image segmentation masks. However, without semantic cues, SAM-based approaches underperform on textures [5], although AutoSAM did achieve a reasonable baseline on TBM.

Physics-Aware Evaluation. Rather than judging predictions purely by pixel overlap with ground truth, Cohen et al. [3] advocated evaluating grain-size fidelity using the Heyn intercept method

[1] Source code and dataset: https://github.com/Scientific-Computing-Lab/SR-TBM.

(ASTM E112 standard [1]). This entails measuring the average grain size from predicted boundaries and comparing it to the true average grain size, thus focusing on the correctness of quantitative metallurgical outcomes. This metric tolerates small gaps or slightly misplaced boundaries as long as the overall grain size distribution is preserved. Prior TBM studies found that even when IoU/Dice are low, a model might still yield acceptable grain-size estimates. We adopt this Heyn intercept evaluation as our primary metric for assessing the effect of super-resolution on analysis.

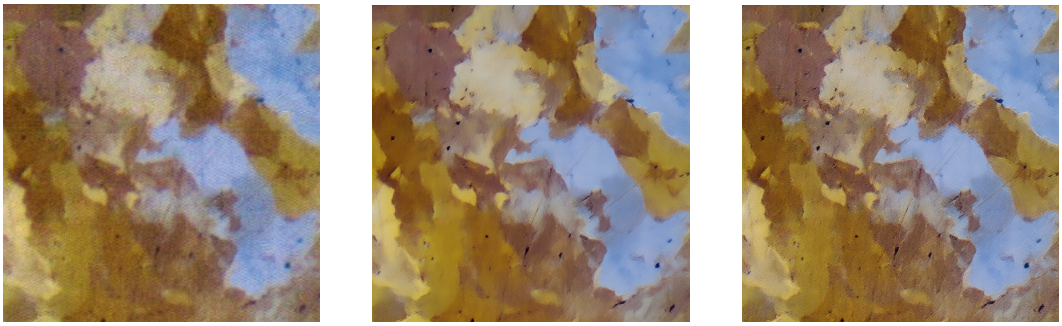
Uncertainty-Aware Edge Detection. Annotating grain boundaries is subjective – experts may disagree on faint edges or where a boundary begins/ends. Zhou et al. [13] introduced the UAED model, which learns a distribution over edge labels to account for annotation uncertainty. UAED produces both an edge probability map and an uncertainty map, and is trained with a specialized loss that emphasizes uncertain regions (to encourage robustness). In our work, we fine-tune the pretrained UAED, using a single ground-truth label per pixel, as we do not have multiple annotators. We use this pretrained model as a starting point for the optimization due to UAED's ability to model soft boundaries and ambiguity being well-suited to metallography, where some boundaries are weak and not all true edges are sharply defined.

Diffusion Models for Super-Resolution. Diffusion-based generative models have recently achieved impressive results in natural image super-resolution. OSEDiff (One-Step Effective Diffusion) by Wu et al. [11] is a state-of-the-art approach that formulates SR as a single diffusion step starting from a low-quality input image instead of pure noise. By fine-tuning a latent diffusion model (Stable Diffusion) with a specially designed loss, OSEDiff produces high-quality upscaled images in one step, offering $>100\times$ speedups over multi-step diffusion models, such as StableSR [11]. Importantly, OSEDiff allows text prompts to guide the super-resolved output, which can inject domain-specific prior knowledge. While diffusion SR has been extensively validated on natural images, its reliability on scientific images like micrographs remains uncertain. There is a legitimate worry that a powerful generative model might hallucinate structures that look plausible but are not real, thus corrupting measurements. To our knowledge, no prior work has systematically examined diffusion SR in metallography or its effect on physical metrics like grain size. Our study fills this gap, evaluating whether a diffusion model can be trusted to enhance metallographic images without biasing quantitative analyses.

3 Method: Diffusion-Based SR for Metallography

One-Step Diffusion (OSEDiff) with Domain Prompting. We leverage OSEDiff to perform $4\times$ super-resolution on metallography images. In the OSEDiff framework [11], a pre-trained Stable Diffusion U-Net (fine-tuned for SR) is used to transform a low-resolution image to a higher resolution in a single forward pass, guided by a text prompt. The original OSEDiff pipeline employs an automated prompt generator (DAPE from Wu et al. [12]) that uses a vision–language model to describe the input image. However, we found this approach unreliable for metallography: DAPE often produced irrelevant or misleading prompts (e.g., describing a micrograph as "snake skin texture" or "blue sky"), leading the model to generate artifacts or incorrect textures. Instead, we supply a fixed, domain-specific prompt — "metallographic image" — at inference time for all images. This simple prompt steers the diffusion model toward outputs that resemble real micrographs. As illustrated in Figure 2, this prompt choice is critical: using a generic VLM-generated prompt yields a washed-out image with under-defined boundaries (panel b), whereas our metallurgy-specific prompt produces a sharp, realistic microstructure with clearly defined grain boundaries (panel c). By guiding the generative model with domain knowledge, we ensure the SR output remains faithful to true metallographic patterns.

SR–TBM Dataset Creation. We applied OSEDiff (with the "metallographic image" prompt) to every image in the TBM dataset to create a super-resolved version. TBM contains 80 polarized light micrographs of metal microstructures; following the original benchmark protocol, we use 64 images for training, 8 for validation, and 8 for testing. Each original image (256×256 px after downsampling in prior work) is upscaled by a factor of $4\times$ to 1024×1024 px using OSEDiff's default settings. The super-resolved dataset, which we call SR–TBM, has a one-to-one correspondence with TBM: for every original image there is a corresponding SR image. To focus our evaluation on SR's contribution (rather than trivial scale differences), we also upscaled the original TBM images to 1024×1024 via simple bilinear interpolation (an "empty" magnification adding no new detail). This way, both Original and SR images have the same resolution and pixel density during training,



(a) Original TBM (b) OSEDiff + DAPE Prompt (c) Domain-Specific Prompt

Figure 2: **Effect of Prompt Choice on Super-Resolution Results.** (a) Original TBM micrograph patch. (b) Super-resolution using OSEDiff with an automatically generated prompt ("sky, yellow") results in an unrealistic, blurry output that fails to resolve grain boundaries. (c) Using the fixed domain prompt "metallographic image" yields a plausible high-resolution image with clearly defined grain boundaries and microstructural detail. This highlights the importance of appropriate prompting to obtain trustworthy SR results in scientific images.

and any performance differences can be attributed to the added high-frequency information from OSEDiff rather than an advantage of larger image size. Figure 1 (above) and Appendix B–Figure 5 show examples of the resulting SR images, which exhibit visibly enhanced detail while maintaining the integrity of grain boundaries.

Training UAED on Original vs. SR Data. To quantify the effect of super-resolution, we train two identical edge-detection models: one on the Original TBM images and one on SR–TBM images. We choose UAED [13] as the model for its robustness to uncertain edges. We fine-tune UAED from its publicly available pre-trained weights (trained on generic edge datasets) separately for each dataset. Both models use the same training hyperparameters, data splits, and augmentations, ensuring a fair comparison. In both cases, images are 1024×1024 (the original model sees the bilinearly upscaled TBM, the other sees the diffusion SR output). We emphasize that aside from the input data, everything is held constant: the train/val/test split is identical, augmentation (random flips/rotations) is identical, and training runs for the same number of epochs. This paired experiment isolates the impact of the SR content. At test time, each model produces a probability map of grain boundaries on the 8 held-out images. We convert the probability maps to binary edge maps by thresholding at the value that maximizes the F1 score on the validation set (ensuring a fair operating point for each). Additionally, we skeletonize the predicted edges to a single-pixel width before computing grain-size metrics, so that boundary thickness does not affect the Heyn intercept measurement (per ASTM standard practice).

4 UAED training

UAED Architecture. Our trained model is a dual-head UNet++ built on top of a flexible encoder API. The encoder is `efficientnet-b7` with ImageNet initialization and depth $= 5$. The decoder is a UNet++ variant with dense skip pathways and nearest-neighbor upsampling. Decoder channel widths are fixed to $(256, 128, 64, 32, 16)$ with BatchNorm and ReLU after each 3×3 convolution.

The network exposes *two capacity-matched decoder branches* sharing the same encoder features: (i) a *mean/logit* head that outputs a single-channel edge score map, and (ii) an *uncertainty* head that outputs a single-channel per-pixel dispersion map. Both heads apply a 3×3 `Conv2d` segmentation head at the final decoder stage. Spatial sizes are preserved by center-cropping the head outputs back to the input $H\times W$.

Learning Objective and Training Protocol. We instantiate a UAED that yields a mean (edge logit) and a per-pixel dispersion (uncertainty) map. Training follows the UAED codebase: a heteroscedastic, class-imbalanced edge loss with RCF-style balancing and a schedule that anneals the influence of

the uncertainty term over epochs.[1] We use AdamW with learning rate $1e{-4}$, weight decay $5e{-4}$, batch size 2, for 100 epochs, identical across Original TBM and SR-TBM. Mixed-precision training is enabled; no random crops are used. Augmentations are limited to the ones used in our training script (horizontal/vertical flips and 90° rotations unless otherwise specified); color-space transforms are disabled to avoid altering etch contrast. All other training hyperparameters (e.g., `itersize`, `std_weight`, `std_weight_final`) remain at their script defaults.

Inference and Post-processing. At inference, the mean head is converted to probabilities via sigmoid to produce a boundary probability map; the uncertainty head yields a non-negative dispersion map. We use a maximum F1-Score condition in threshold selection produce a binary boundary mask. For Heyn intercept calculation skeletons are extracted with standard thinning.

5 Results

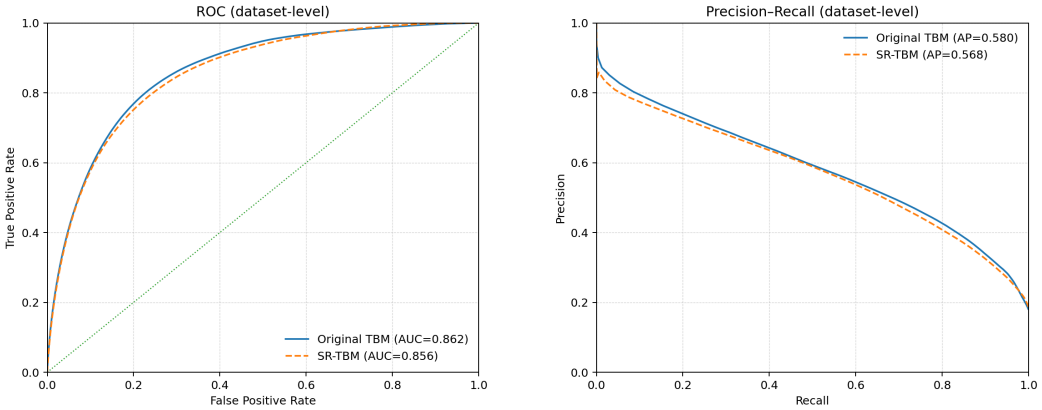
Qualitative Evaluation. Visually, the diffusion-based super-resolution greatly improves the clarity of micrographs. In Figure 1, the super-resolved image reveals grain boundaries that were barely discernible in the original. The fine details of the microstructure are reconstructed with higher contrast and continuity. Crucially, we observe no hallucinated boundaries or unrealistic textures in the SR images — grains appear physically plausible and consistent with the original structures. We had an experienced material scientist examine all SR test images, and no spurious grain boundaries were found, providing confidence that the SR process did not introduce false features. Appendix B–Figure 5 provides additional examples comparing original vs. SR images from TBM: in each case, the SR version presents a sharper image with preserved true boundaries (even faint ones become more continuous) and no obvious artifacts. These results indicate that, with our guided prompting, diffusion SR can enhance image quality without compromising scientific validity in metallography.

Table 1: **Heyn grain size comparison on TBM test set.** Each model's predicted average grain size (mean lineal intercept length, $\bar{\ell}$) is compared to the ground truth (GT) value, and the error is reported in pixels and as a percentage of GT. Incorporating diffusion super-resolution (UAED SR–TBM) dramatically reduces grain-size error relative to using original images, indicating improved fidelity to true grain statistics. (All values scaled to 1024 px image size for consistency.)

Dataset	$\bar{\ell}^{\text{pred}}$ [px]	$\bar{\ell}^{\text{GT}}$ [px]	$\Delta\lvert\text{pred} - \text{GT}\rvert$ [px]	% Error $\left(\frac{\Delta}{\bar{\ell}^{\text{GT}}} \times 100\right)$
MLOgraphy	115.30	105.44	9.86	9.35%
AutoSAM	113.7	105.44	8.26	7.83%
UAED TBM	95.46	105.44	9.99	9.47%
UAED SR-TBM	110.82	105.44	**5.38**	**5.10%**

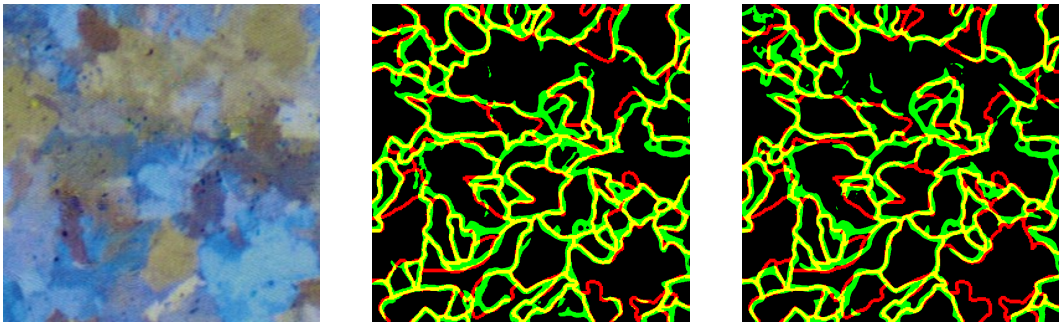
Quantitative Grain-Size Analysis. Table 1 reports the average grain size estimated from the predicted boundary maps, using the Heyn intercept method, for several models (detailed . We compare our UAED-based edge detector trained on original images (UAED TBM) versus on super-resolved images (UAED SR–TBM), and also include two reference baselines from prior work: MLOgraphy and AutoSAM (applied to the TBM test set). All methods are evaluated on the same 8 test micrographs, and all grain-size values are normalized to correspond to a 1024 px image scale (for fairness, baseline outputs originally at 256 px were scaled up by $4\times$ when measuring intercept lengths). As shown, the SR-trained UAED yields by far the most accurate grain-size estimates. Its predicted mean lineal intercept ($\bar{\ell}$) is 110.82 px, very close to the ground-truth 105.44 px, resulting in an absolute error of 5.38 px (5.10%). This is a 47% reduction in error compared to the same model trained on original TBM images (which had 9.99 px error, 9.47%). It is also a 35% improvement over the previous best method, AutoSAM (8.26 px error, 7.83%). Notably, the UAED model trained on original low-quality images underestimates the grain size (predicted 95.46 px vs. true 105.44 px), likely because it misses some boundaries or produces overly fragmented edges, whereas the SR-trained model gets much closer. MLOgraphy and UAED (original) have similar error rates around 9–9.5%, while AutoSAM was somewhat better at $\sim$7.8%; our SR approach substantially outperforms all of them on this physics-based metric. This confirms that diffusion SR can meaningfully boost quantitative grain-size accuracy in metallography.

[1] Exact loss and annealing expressions are implemented in our training script and left unchanged; see the released code for the `cross_entropy_loss_RCF` function.



(a) ROC curve comparison (b) Precision–Recall curve comparison

Figure 3: **ROC and PR curve comparison.** Models trained on original TBM (blue) and super-resolved SR–TBM (orange dashed) achieve nearly identical ROC (AUC $= 0.862$ vs. 0.856) and PR (AP $= 0.580$ vs. 0.568) performance, confirming that super-resolution does not alter predictive capability.



(a) Input Image (b) TBM Prediction (c) SR-TBM Prediction

Figure 4: **Edge prediction on original vs. super-resolved images.** (a) Input image. (b) Prediction from TBM-trained model. (c) Prediction from SR–TBM-trained model. Yellow = True Positive (TP), Red = False Negative (FN), Green = False Positive (FP). Both models yield visually similar, high-quality edge maps, showing SR preserves boundary integrity.

Importantly, we verify that this improvement does not come at the cost of pixel-level performance. The UAED–SR model's traditional edge detection scores remain on par with the UAED model trained on original data. Specifically, the average precision (AP) and ROC–AUC for detecting boundary pixels are essentially unchanged (within 0.5% difference between SR-trained and original-trained models on the test set) (Figure 3,Figure 4). This further validates that the SR-TBM does not result in significant hallucinated edge boundaries that would skew pixel based metrics. This result, coupled with the improvement shown for the average grain size, suggests that diffusion SR provides genuinely better input data, enabling more accurate detection of true boundaries that translate into better grain measurements. In summary, by integrating a trustworthy SR step in the metallographic analysis pipeline, we achieve both qualitatively clearer images and quantitatively superior grain-size estimates.

Acknowledgments

This work was supported by the Pazy Foundation.

References

[1] E 112-13: Standard test methods for determining average grain size. *ASTM International*, 2013.

[2] Inbal Cohen, Boaz Meivar, Peihan Tu, Shai Avidan, and Gal Oren. Texturesam: Towards a texture aware foundation model for segmentation. *arXiv preprint arXiv:2505.16540*, 2025.

[3] Inbal Cohen, Julien Robitaille, Francis Quintal Lauzon, Ofer Beeri, Shai Avidan, and Gal Oren. Avoiding post-processing with context: Texture boundary detection in metallography. In *AI for Accelerated Materials Design-NeurIPS 2024*, 2024.

[4] Brian L DeCost, Matthew D Hecht, Toby Francis, Bryan A Webler, Yoosuf N Picard, and Elizabeth A Holm. Uhcsdb: Ultrahigh carbon steel micrograph database. *Integrating Materials and Manufacturing Innovation*, 6(2):197–205, 2017.

[5] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[6] Julian Luengo, Raul Moreno, Ivan Sevillano, David Charte, Adrian Pelaez-Vegas, Marta Fernandez-Moreno, Pablo Mesejo, and Francisco Herrera. A tutorial on the segmentation of metallographic images: Taxonomy, new metaldam dataset, deep learning-based ensemble model, experimental analysis and challenges. *Information Fusion*, 78:232–253, 2022.

[7] Mengyang Pu, Yaping Huang, Yuming Liu, Qingji Guan, and Haibin Ling. Edter: Edge detection with transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1402–1412, 2022.

[8] Matan Rusanovsky, Ofer Beeri, and Gal Oren. An end-to-end computer vision methodology for quantitative metallography. *Scientific Reports*, 12(1):4776, 2022.

[9] Matan Rusanovsky, Ofer Be'eri, Shai Avidan, and Gal Oren. Universal semantic-less texture boundary detection for microscopy (and metallography). In *NeurIPS 2023 Workshop on Machine Learning and the Physical Sciences*, 2023.

[10] Tal Shaharabany, Aviad Dahan, Raja Giryes, and Lior Wolf. Autosam: Adapting sam to medical images by overloading the prompt encoder. *arXiv preprint arXiv:2306.06370*, 2023.

[11] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2406.08177.

[12] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. SeeSR: Towards semantics-aware real-world image super-resolution. *arXiv preprint arXiv:2311.16518*, 2023.

[13] Caixia Zhou, Xiaotao Sun, Hao Lin, Zhen Zhang, Yi Zhou, Xitong Shen, Yuesong Xu, Qian Yu, Jiwen Zhou, Xiangyu He, and Fei Wang. Uaed: Uncertainty-aware edge detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22645–22654, 2023.

A Appendix: Evaluation Metrics

We evaluate model performance using both a physics-informed grain-size metric and standard edge detection metrics. All metrics are computed on the held-out test split of 8 images.

Heyn Intercept (Grain-Size) Error

A skeletonization procedure is first applied to the edge map, thinning all edges to a single pixel width. This ensures that the metric is unaffected by edge width. Let B be a binary boundary map (GT $B^\star$ or prediction $\hat{B}$). Following the method described in the TBM benchmark and [9], we estimate the mean lineal intercept (Heyn) length by sampling $M = 50$ randomly oriented test lines $\{\ell_m\}_{m=1}^{M}$, with orientations $\theta_m \sim \mathcal{U}[0, \pi)$ and uniformly sampled positions. For each line, let L_m be its length within the field of view and $n_m(B)$ the number of intersections with B. The Heyn estimate is

$$\bar{\ell}(B) = \frac{\sum_{m=1}^{M} L_m}{\sum_{m=1}^{M} n_m(B)}.$$

We report the absolute Heyn error (in pixels)

$\Delta_{\mathrm{Heyn}} = \left|\bar{\ell}(\hat{B}) - \bar{\ell}(B^\star)\right|$ as well as the percent change from GT: $\Delta^{\%}_{\mathrm{Heyn}} = 100 \cdot \Delta_{\mathrm{Heyn}}/\bar{\ell}(B^\star)$.

This physics-informed metric captures whether edge maps preserve grain-size statistics, remaining robust to small gaps or soft boundaries. We evaluate Heyn only for our UAED variants (Original vs. SR) to isolate the effect of super-resolution; baselines are compared with AP/AUC.

A.1 Average Precision (AP) and ROC–AUC

To compare pixel-level boundary detection performance, we also calculate the Average Precision (precision–recall AUC) and the ROC–AUC for the edge probability maps produced by the models. Ground-truth edges are treated as positive class pixels. We note that baselines like MLOgraphy and AutoSAM were originally evaluated on TBM using these metrics. In our experiments, we found that the UAED model trained on SR–TBM achieves virtually the same AP and ROC–AUC as the model trained on original TBM, indicating no loss (and in fact a slight gain) in per-pixel accuracy despite the generative upscaling.

Average Precision (AP). Given per-pixel scores $\{s_i\}$ and binary labels $y_i \in \{0, 1\}$, thresholding at τ yields

$$\mathrm{TP}(\tau),\ \mathrm{FP}(\tau),\ \mathrm{FN}(\tau),\ \mathrm{TN}(\tau).$$

Precision–recall at τ are

$$\mathrm{Prec}(\tau) = \frac{\mathrm{TP}(\tau)}{\mathrm{TP}(\tau) + \mathrm{FP}(\tau)}, \qquad \mathrm{Rec}(\tau) = \frac{\mathrm{TP}(\tau)}{\mathrm{TP}(\tau) + \mathrm{FN}(\tau)}.$$

AP is the area under the PR curve:

$$\mathrm{AP} = \int_0^1 P(R)\, dR \ \approx\ \sum_{k=1}^{K} \big(R_k - R_{k-1}\big)\, \tilde{P}(R_k),$$

where $\{(R_k, P_k)\}_{k=1}^{K}$ are PR points obtained by sweeping τ and $\tilde{P}$ denotes the (monotone) interpolated precision. AP is threshold-independent and is well-suited to sparse, partially labeled boundaries.

ROC-AUC. Define true/false positive rates at τ:

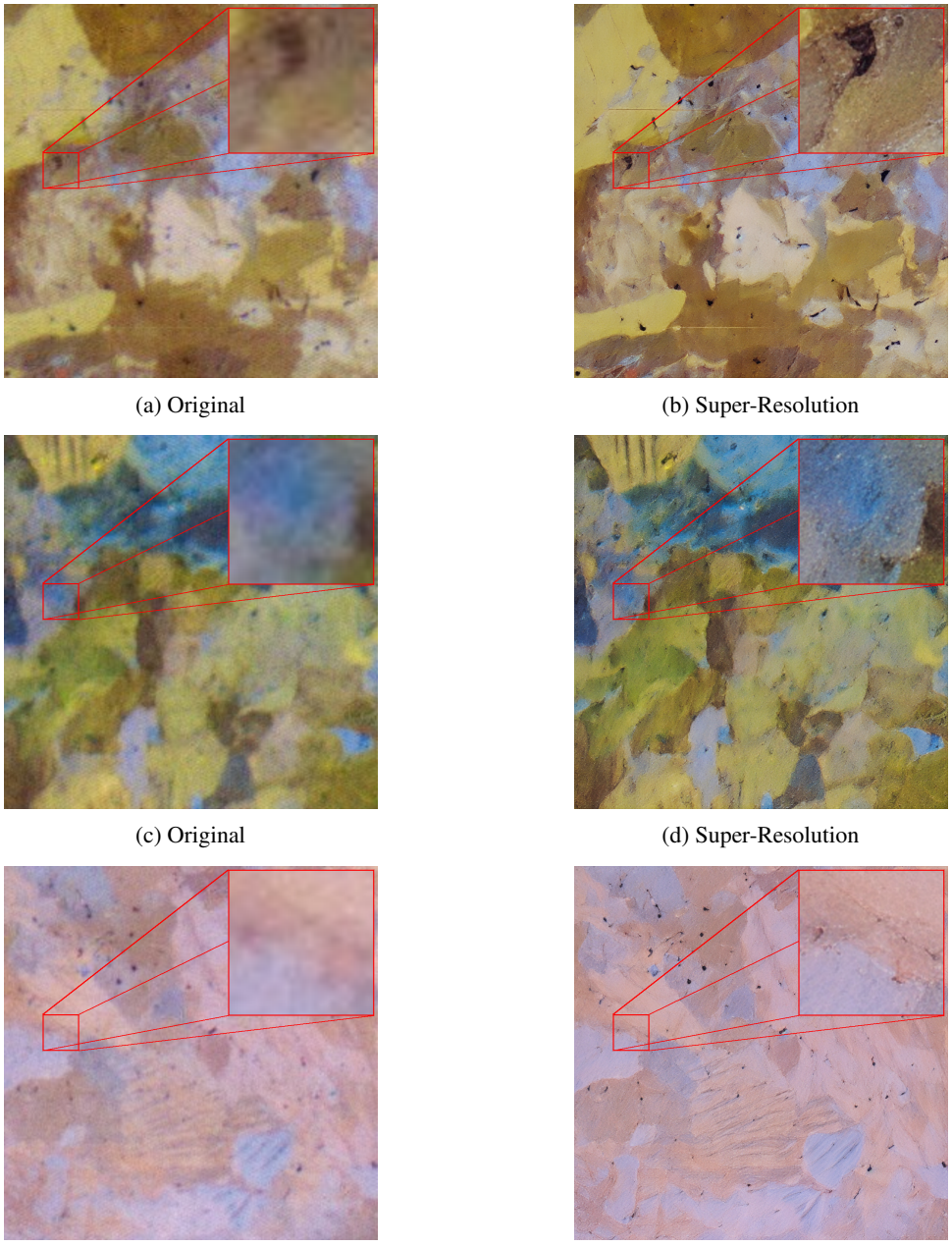
$$\mathrm{TPR}(\tau) = \frac{\mathrm{TP}(\tau)}{\mathrm{TP}(\tau) + \mathrm{FN}(\tau)}, \qquad \mathrm{FPR}(\tau) = \frac{\mathrm{FP}(\tau)}{\mathrm{FP}(\tau) + \mathrm{TN}(\tau)}.$$

ROC-AUC is the area under the ROC curve:

$$\mathrm{AUC} = \int_0^1 \mathrm{TPR}(\mathrm{FPR})\, d\,\mathrm{FPR} \ \approx\ \sum_{k=1}^{K} \big(\mathrm{FPR}_k - \mathrm{FPR}_{k-1}\big)\, \tilde{\mathrm{TPR}}(\mathrm{FPR}_k).$$

It complements AP by summarizing sensitivity–specificity tradeoffs when calibrated scores are available.

B Appendix: Qualitative Results



(a) Original (b) Super-Resolution

(c) Original (d) Super-Resolution

(e) Original (f) Super-Resolution

Figure 5: **Original vs. Super-Resolution images.** Each row shows an original metallographic image (left) and its corresponding super-resolved version (right) generated using OSEDiff. Note the enhanced resolution, high quality reconstruction, and preserved grain boundaries in the right column.