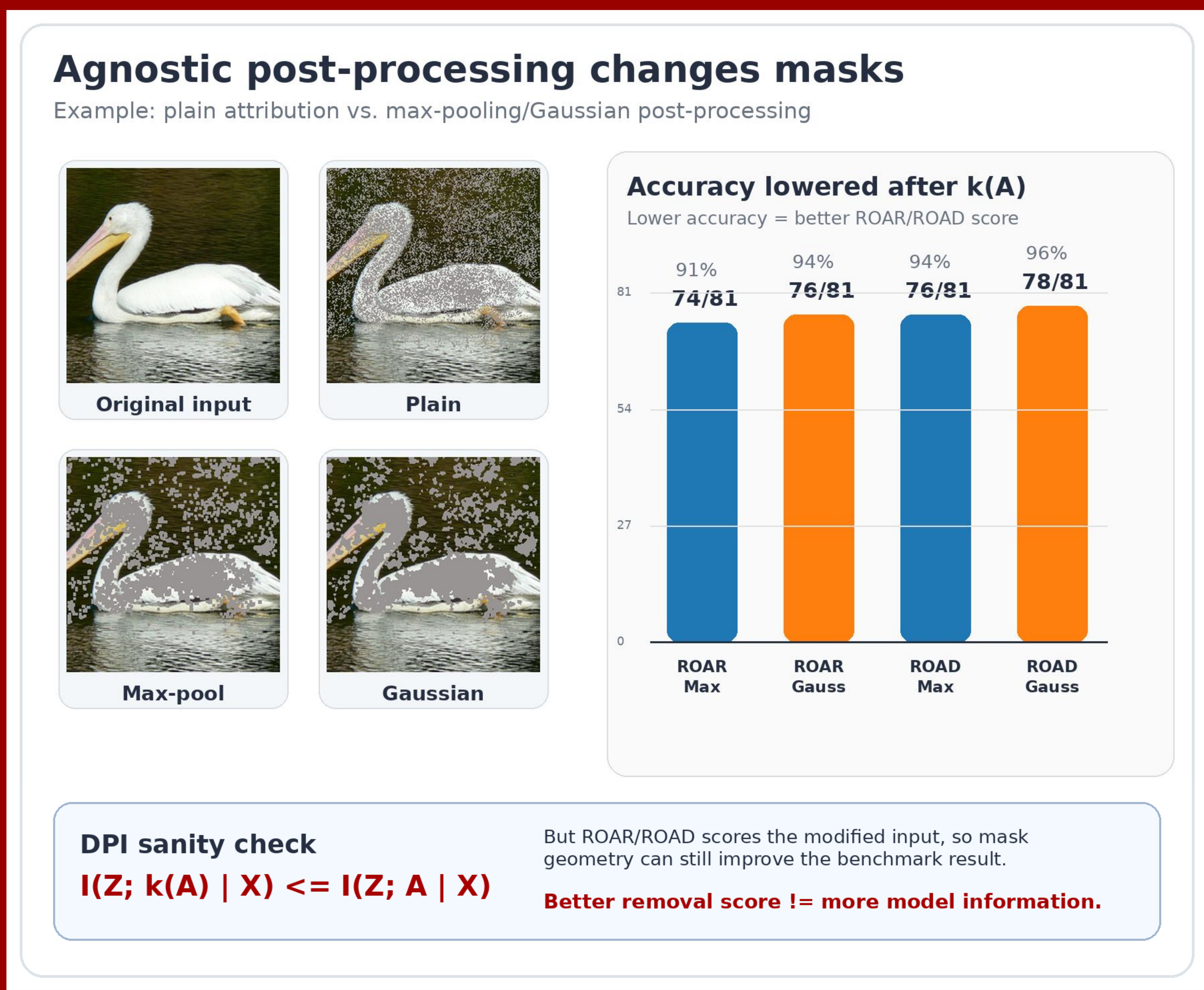


ROAR/ROAD can reward mask geometry, not model understanding

On Pitfalls of RemOve-And-Retrain: A Data Processing Inequality Perspective

Junhwa Song[1], Keumgang Cha[2], Junghoon Seo[1] · [1]PIT-IN Corp. · [2]SI Analytics Co., Ltd.

1 TL;DR

MAIN 3 TAKEAWAYS

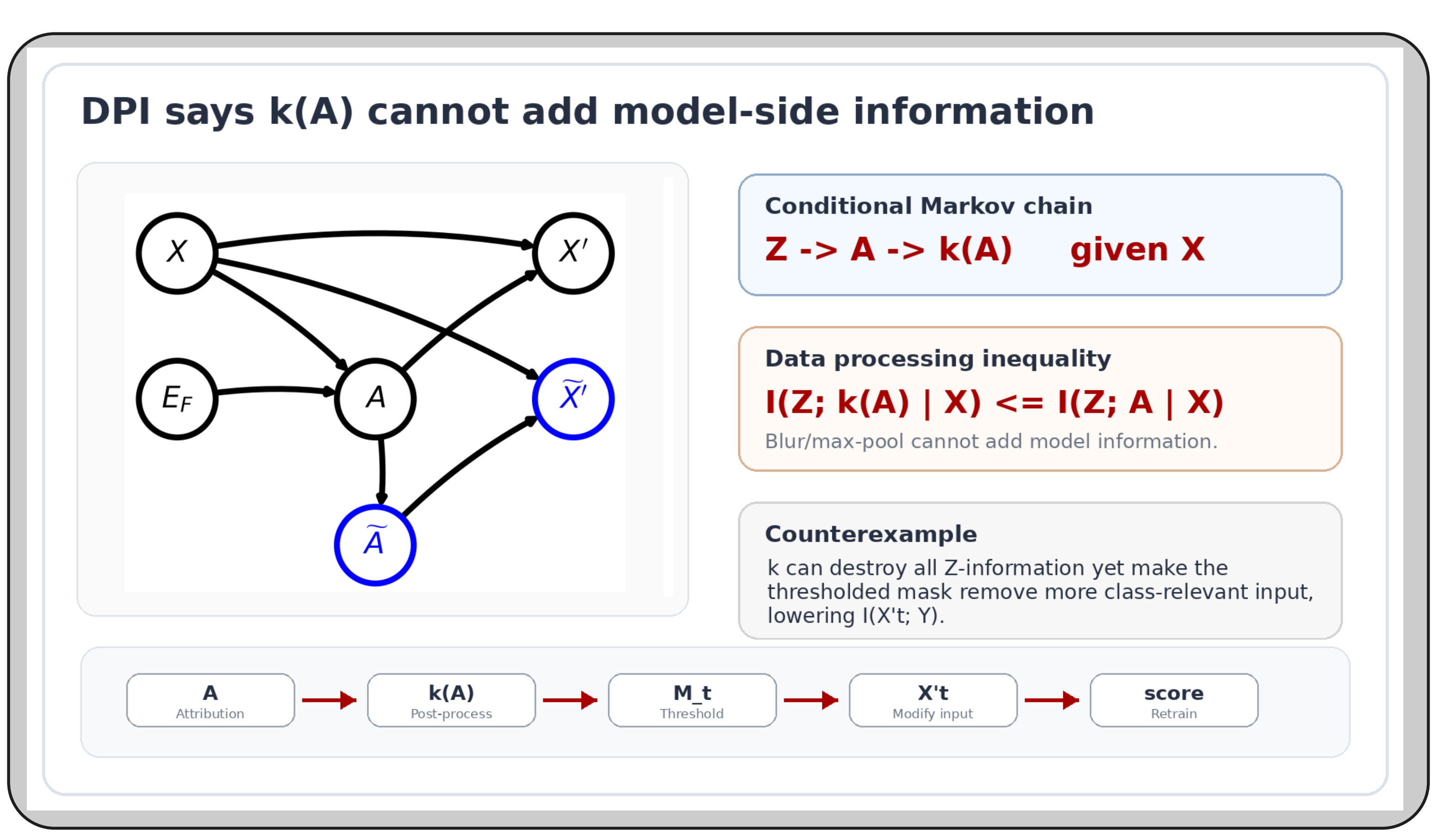
- Agnostic blur/max-pool cannot add model information.
- Yet it lowers ROAR/ROAD accuracy in most comparisons.
- ROAR/ROAD can reward low-TV mask geometry.

2 Background / Motivating Question

- ROAR removes high-attribution features and retrains; lower accuracy is interpreted as a better attribution.
- Low accuracy can arise from mask geometry, not from more information about the decision function.
- Can an agnostic k(A) improve scores without improving interpretability?

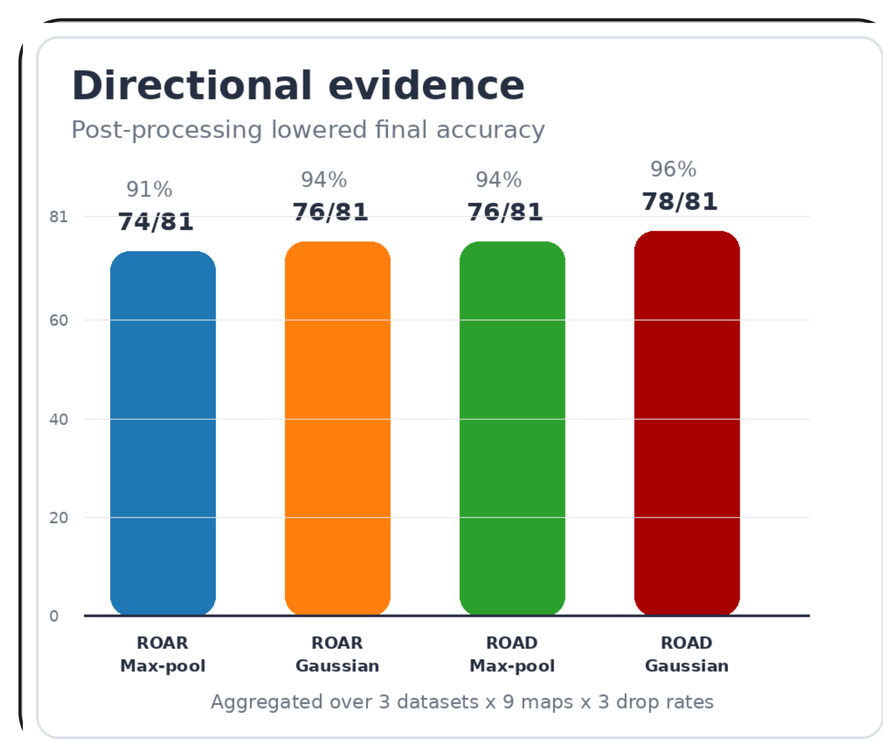
3 Theory: DPI sanity check

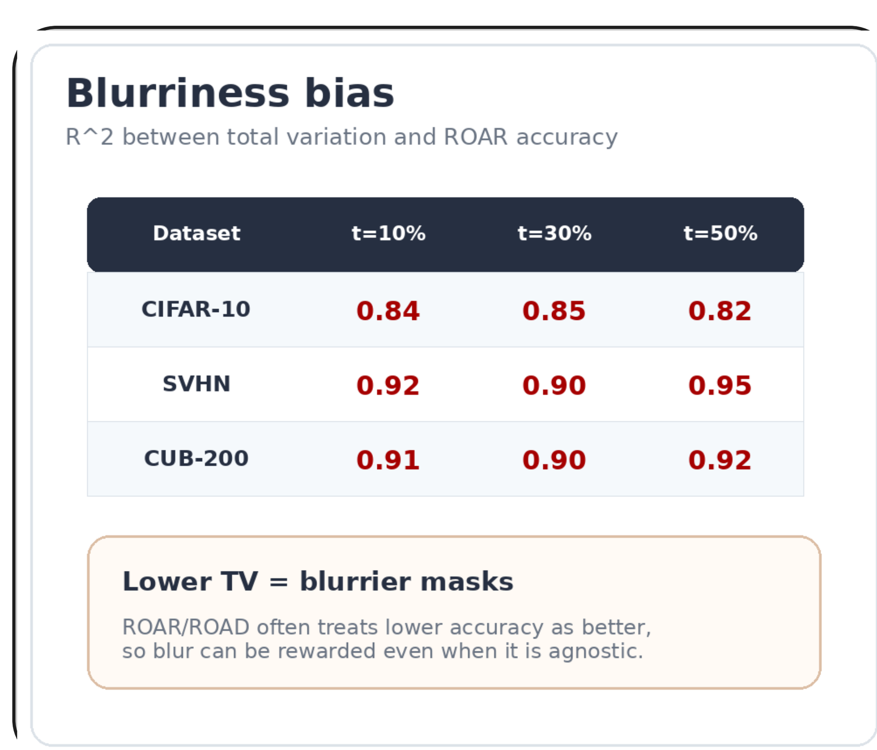
- For fixed X, agnostic post-processing forms Z -> A -> k(A).
- DPI: I(Z; k(A) I X) <= I(Z; A I X), so k cannot add model-side information.
- Still, thresholding k(A) changes the mask and can lower I(X't; Y).



4 Experiments: post-processing wins

DATA: CIFAR-10, SVHN, CUB-200
MODELS: ResNet18 / ResNet50
k(A): max-pool 3x3 or Gaussian sigma=1
MAPS: 9 maps x 3 drop rates
LOWER ACCURACY = better ROAR/ROAD score





Blurriness bias

R^2 between total variation and ROAR accuracy

Dataset	t=10%	t=30%	t=50%
CIFAR-10	0.84	0.85	0.82
SVHN	0.92	0.90	0.95
CUB-200	0.91	0.90	0.92

Lower TV = blurrier masks

ROAR/ROAD often treats lower accuracy as better, so blur can be rewarded even when it is agnostic.

- Lower total variation (TV) means blurrier masks.
- TV predicts ROAR accuracy with R^2 = 0.82-0.95.
- GC^2 is less affected because its maps have lower resolution.
- Mask smoothness is a benchmark confound.

5 Discussion

- Do not rank attribution methods solely by ROAR or ROAD.
- Report mask TV, mask visuals, ablations, and sanity checks with removal scores.
- Image results may not transfer unchanged to text, graph, or time-series settings.
- Prefer evaluations that directly target decision-function information.

References

- Adebayo, Julius, et al. "Sanity checks for saliency maps." In NeurIPS. 2018.
- Rong, Y., Leemann, T., Borisov, V., Kasneci, G., and Kasneci, E. A consistent and efficient evaluation strategy for attribution methods. In ICML, pp. 18770–18795. PMLR, 2022.

Mechanistic Interpretability Workshop, ICML 2026

sjh@pitin-ev.com