CVPR 2026

Rel-Zero: Harnessing Patch-Pair Invariance for Robust Zero-Watermarking Against AI Editing

Robust Zero-Watermarking via Stable Relational Geometry

Pengzhen Chen[1,2,3] Yanwei Liu[1,3†] Xiaoyan Gu[1,2,3†] Xiaojun Chen[1,2,3]
Wu Liu[4] Weiping Wang[1]

[1] Institute of Information Engineering, Chinese Academy of Sciences
[2] School of Cyber Security, University of Chinese Academy of Sciences
[3] State Key Laboratory of Cyberspace Security Defense
[4] University of Science and Technology of China

† Corresponding authors

Core idea

Do not embed a signal into pixels. Extract the watermark from stable patch-pair relations already present in the image.

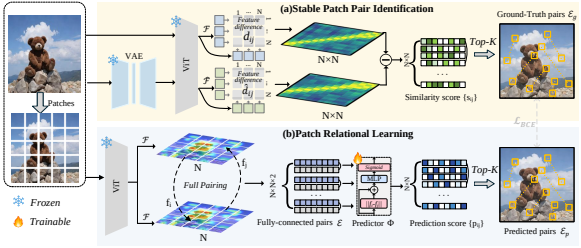
Rel-Zero

Robust zero-watermarking against AI editing

One-sentence idea

Do not embed a signal into pixels. Extract a watermark from stable patch-pair relations already present in the image.

- **Goal**: robust authentication with zero visual distortion
- **Key signal**: stable inter-patch relational geometry
- **Watermark**: Top-K stable patch-pair indices



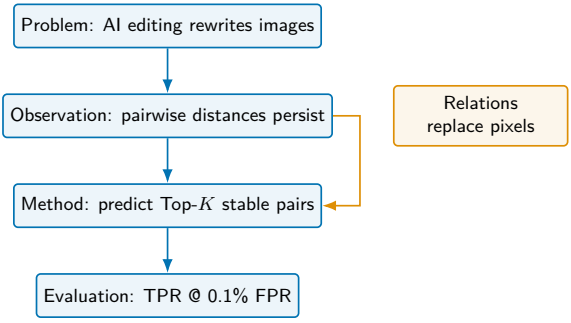
Framework overview: predict stable relations and verify by index overlap.

Talk Roadmap

Main thread

1. Why embedding and classical zero-watermarking are both fragile
2. Empirical finding: patch-pair distances survive editing
3. How Rel-Zero trains, generates, and verifies watermarks
4. Results: robustness, uniqueness, and ablations

- Left column: core message and equations
- Right column: paper figures, tables, or compact diagrams



Background: Generative Editing Threatens Image Provenance

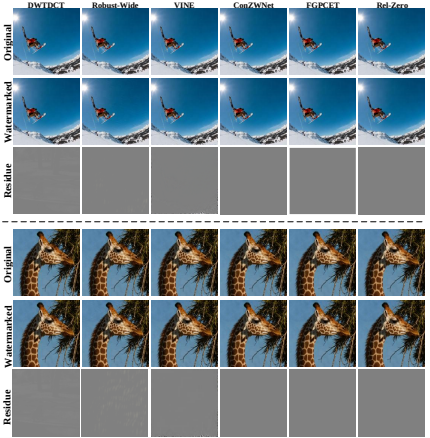
New threat

Diffusion editing, instruction-guided editing, and inpainting make image content easy to rewrite naturally.

- Appearance, texture, and local semantics can change drastically
- Copyright and provenance verification become harder
- High-fidelity domains cannot tolerate visible perturbations

Watermarking target

Resist global and local AI edits without degrading image quality.



Embedding methods leave residual artifacts; zero-watermarking preserves pixels.

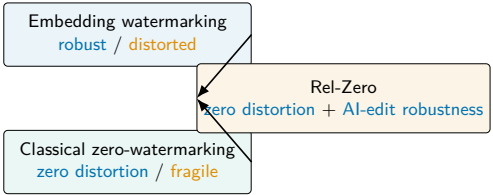
The Tension in Existing Watermarking Paradigms

Embedding watermarking

- Writes a signal into pixels or frequency bands
- Strong robustness often requires stronger perturbation
- Poor fit for medical, autonomous driving, or high-value imagery

Zero-watermarking

- Registers an external fingerprint without modifying pixels
- Classical features depend on absolute appearance
- Generative rewriting can destroy these descriptors



Shift from absolute features to relational structure

Key Observation: Patches Change, Pairwise Distances Persist

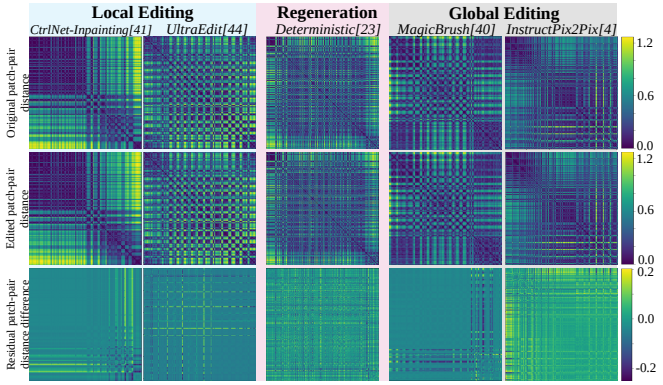
Empirical finding

Editing changes local appearance, yet many patch-pair distances remain approximately invariant.

$$d_{ij}^{\text{before}} = \|v_i - v_j\|_2$$

$$d_{ij}^{\text{after}} = \|v'_i - v'_j\|_2$$

- Distant patch pairs tend to remain distant after editing
- Relations are more stable than absolute RGB values



RGB patch-pair distances preserve structure across several edit models.

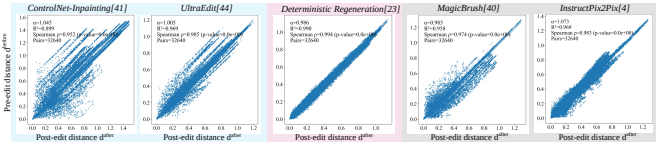
Quantification: Near-Linear Distance Preservation

Linear hypothesis

If editing is structure-preserving, pairwise distances should follow a global similarity transform.

$$d_{ij}^{\text{after}} \approx \alpha d_{ij}^{\text{before}} + \beta$$

- $\alpha \approx 1$, $\beta \approx 0$
- High R^2 and Spearman correlation
- Stable rank/order enables index-based watermarking



Distance-distance scatter plots show strong pre/post-edit predictability.

Residual View: Most Relation Changes Stay Near Zero

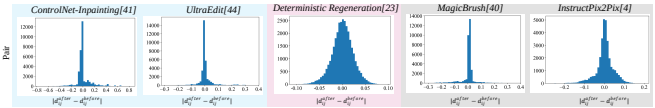
Residual

$$r_{ij} = |d_{ij}^{\text{after}} - d_{ij}^{\text{before}}|$$

- Most r_{ij} values concentrate near 0
- No strong systematic bias
- A stable subset can serve as relational anchors

From observation to method

Treat a stable relation set $\mathcal{E}$ as the image's own watermark signature.



Residual distribution: stable patch pairs provide redundant authentication signals.

Why Can Relations Stay Stable?

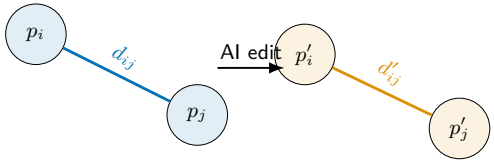
Two explanations

1. **Fidelity constraints**: editing models preserve non-target structure
2. **Low-dimensional edits**: style, tone, and texture shifts are nearly global

$$v_i' \approx Av_i + b$$

$$v_i' - v_j' \approx A(v_i - v_j)$$

- Translation term b cancels in differences
- Global scaling preserves many relation rankings



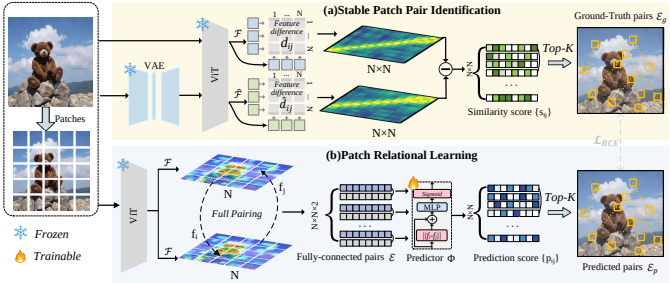
$d_{ij}' \approx \alpha d_{ij} + \beta$
absolute locations shift; internal geometry remains traceable

Method Overview: Relational Zero-Watermarking

Three stages

1. Build supervision from stable patch pairs
2. Learn a pair predictor Φ
3. Generate and verify relation-index watermarks

- Training uses a VAE as an edit surrogate
- Inference only needs ViT + MLP pair scoring
- External storage contains only $\mathcal{E}_p$ indices



Framework: target construction and inference-time verification are decoupled.

Stage 1: Use VAE Reconstruction to Build Stable-Pair Targets

Original and reconstructed features

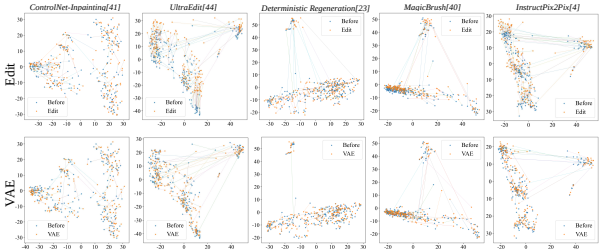
$$\hat{\mathbf{I}} = V_{ae}(\mathbf{I})$$

$$\mathcal{F} = \phi_{\text{vit}}(\mathbf{I}), \quad \hat{\mathcal{F}} = \phi_{\text{vit}}(\hat{\mathbf{I}})$$

$$s_{ij} = \exp(-|d_{ij} - \hat{d}_{ij}|)$$

$$\mathcal{E}_g = \text{TopK}_{i,j}(s_{ij})$$

- VAE acts as a cheap edit surrogate
- Top-K stable pairs become positive labels



Supplementary evidence: VAE and real edits induce similar ViT feature shifts.

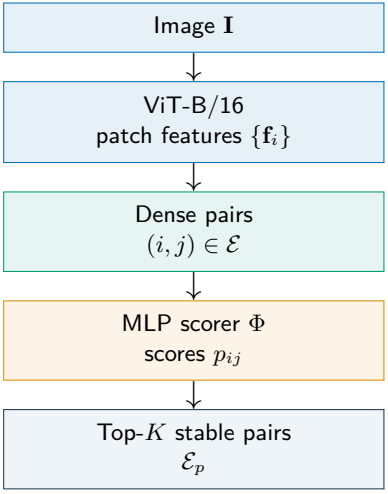
Stage 2: Learn a Patch-Pair Stability Predictor

Pair predictor

$$p_{ij} = \Phi(\mathbf{f}_i, \mathbf{f}_j)$$

$$p_{ij} = \sigma(\psi[\mathbf{f}_i \oplus \mathbf{f}_j \oplus \|\mathbf{f}_i - \mathbf{f}_j\|_2])$$

- Input: ViT patch features
- Output: probability that an edge is stable
- Target: binary labels derived from $\mathcal{E}_g$



Stage 3: Watermark Generation and Verification

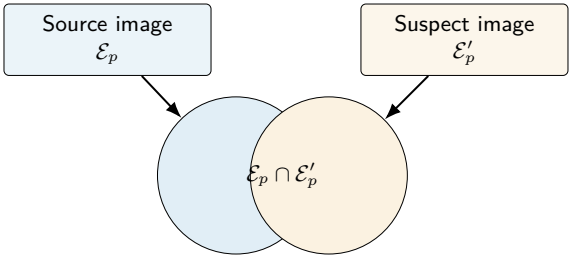
Generation

$$\mathcal{E}_p = \mathrm{TopK}(\Phi(\phi_{\mathrm{vit}}(\mathbf{I})))$$

Verification

$$\eta = \frac{|\mathcal{E}_p \cap \mathcal{E}'_p|}{K}$$

- $\mathcal{E}_p$: registered watermark of the source image
- $\mathcal{E}'_p$: extracted watermark of a suspect image
- Authenticate if $\eta \geq \tau$



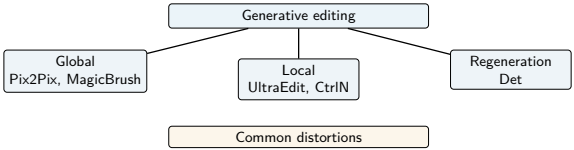
Experimental Setup

Data and implementation

- Train: COCO
- Eval: 10,000 images from UltraEdit and MagicBrush
- Image size: 224×224
- ViT-B/16, Stable Diffusion v1.4 VAE
- Patch size: 16×16, $N = 196$, $K = 50$

Metrics

Fidelity: PSNR / SSIM / LPIPS
Robustness: TPR @ 0.1% FPR



Main Results: Robustness Under Zero Distortion

Core result

Rel-Zerosubstantially improves zero-watermark robustness against generative edits while leaving image pixels untouched.

- SSIM = 1.000, LPIPS = 0.000
- MagicBrush: 95.63
- UltraEdit: 96.55
- ControlNet-Inpainting: 97.43

Unit: TPR @ 0.1% FPR (%)

Method	Det	Pix2Pix	Magic	Ultra	CtrlN
DWT-DCT	0.09	0.04	0.05	0.32	0.56
Robust-Wide	90.41	97.23	81.97	80.45	82.11
VINE	**99.98**	**97.46**	94.58	**99.96**	93.04
ConZWNet	0.10	0.02	0.01	5.13	2.41
FGPCET	1.13	0.54	0.11	7.25	3.22
Rel-Zero	85.13	89.65	**95.63**	96.55	**97.43**

Method	PSNR	SSIM	LPIPS
Robust-Wide	41.93	0.9908	0.0034
VINE	37.34	0.9934	0.0063
Rel-Zero	–	**1.000**	**0.000**

Visualization: Similar Relation Watermarks After Editing

How to read the figure

Compare extracted relational watermarks before and after editing.

- Global editing tends to be more destructive
- Local editing often preserves background relations
- Authentication relies on redundant stable pairs

Intuition

Not every patch must survive; enough pair relations must still match.

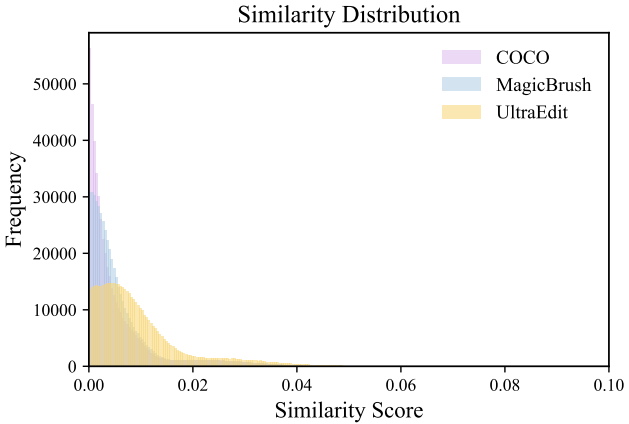
Pre-edit and post-edit relational watermark visualization.

Uniqueness: Different Images Rarely Collide

Cross-image similarity

$$\eta_{a,b} = \frac{|\mathcal{E}_p(\mathbf{I}_a) \cap \mathcal{E}_p(\mathbf{I}_b)|}{K}$$

- Measured on COCO, UltraEdit, and MagicBrush
- Most overlaps between different images are near zero
- The watermark is image-specific, not a generic template



Similarity distribution between watermarks from different images.

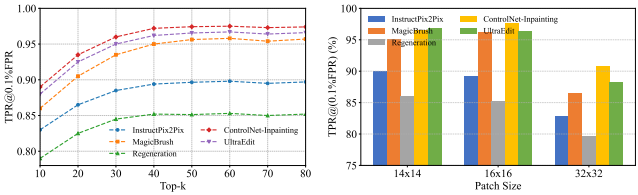
Parameter Analysis: K and Patch Count

Top-K pairs

- Larger K provides more relational redundancy
- Performance saturates after $K = 50$

Patch count

- Too coarse patching weakens fine-grained relations
- 14×14 and 16×16 are both stable



Relational redundancy and patch granularity jointly shape robustness.

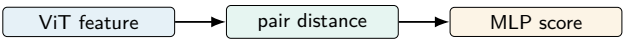
Ablation: ViT Features and a Simple MLP Work Best

Design takeaway

ViT patch features capture fine-grained relational cues; adding Transformer/GAT layers does not help.

- CNN backbones reduce robustness
- Attention layers may blend patch representations
- Pair-distance discrimination benefits from direct pair features

Model configuration	TPR @ 0.1% FPR
Ours: ViT + MLP	**97.43**
ViT → ResNet-18	84.13
ViT → ResNet-50	85.21
MLP → Transformer + MLP	92.11
MLP → GAT + MLP	94.45



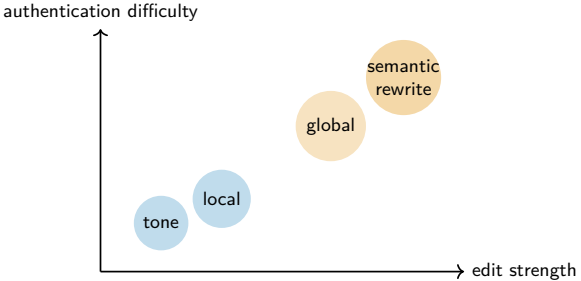
Value and Boundary Conditions

Strengths

- No pixel modification: zero distortion by design
- No edit-model inference during verification
- External watermark indices can be encrypted

Limitations

- Large semantic reconstruction remains harder
- Strong geometric transforms may require alignment
- Thresholds depend on the target FPR calibration



Rel-Zero works best when relational geometry is preserved

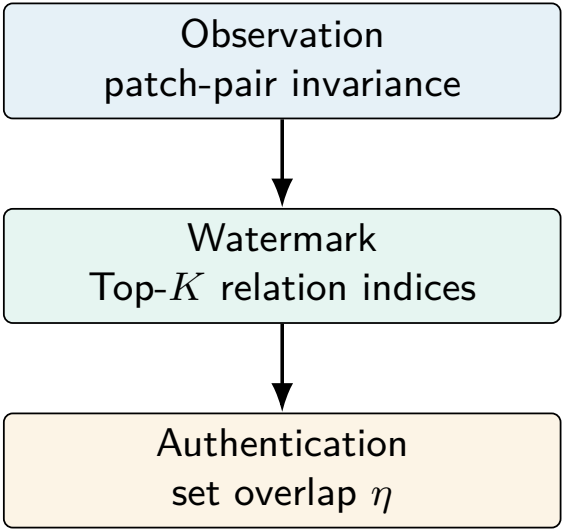
Summary

Three contributions

1. Identifies patch-pair distance preservation under generative editing
2. Builds a zero-watermark from stable relation indices
3. Improves robustness over prior zero-watermarking baselines

Takeaway

Rel-Zeroprotects internal relational geometry rather than pixel-level appearance.



References

P. Chen, Y. Liu, X. Gu, X. Chen, W. Liu, and W. Wang.
Rel-Zero: Harnessing Patch-Pair Invariance for Robust Zero-Watermarking Against AI Editing.
CVPR, 2026.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer.
High-Resolution Image Synthesis with Latent Diffusion Models.
CVPR, 2022.

T. Brooks, A. Holynski, and A. A. Efros.
InstructPix2Pix: Learning to Follow Image Editing Instructions.
CVPR, 2023.

L. Zhang, A. Rao, and M. Agrawala.
Adding Conditional Control to Text-to-Image Diffusion Models.
ICCV, 2023.

S. Lu, Z. Zhou, J. Lu, Y. Zhu, and A. W.-K. Kong.
Robust Watermarking Using Generative Priors Against Image Editing.
arXiv, 2024.