SAME MASK, CONFLICTING CONTEXTS: LEARNING TO LOCALIZE AI EDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowing that an image has been edited does not specify where it was edited. We study how this distinction should shape the adaptation of spatial visual encoders. Opposite-status context training (OSCT) keeps the recipient's mask fixed while changing the image that supplies attention: in an auxiliary forward, edited recipients read keys and values from unedited donors, and vice versa, with donors drawn from different parent images. Natural and altered-context forwards jointly train the same PE-Spatial localizer; deployment remains an ordinary single-image forward. On a 37,162-image modern-edit evaluation, five 384-update continuations from one shared supervised checkpoint improve edited-image macro F1 by 1.32 points and pixel AP by 1.31 points over matched ordinary training, while reducing pristine-image false-positive area from 0.54% to 0.41%. OSCT also improves F1, IoU, and AP on three external localization sets. Parent-matched donor comparisons support opposite-status selection over same-status exchange, and the improvement persists after matched score refitting of both trained endpoints. These results identify the relationship between borrowed context and recipient supervision as a useful training choice for AI-edit localization, without adding an inference module.

1 INTRODUCTION

An image-level judgment can be correct while a localization mask is wrong. A model may recognize that an image contains an AI edit yet implicate untouched regions or miss the modified object. Recent systems such as SIDA and Detective SAM address the need to move beyond image-level detection toward localized forensic predictions (Huang et al., 2025; Lek et al., 2026). We study single-image AI-edit localization: given one RGB image, predict the annotated edit region, with an empty target for unedited images, without access to the original image, editing prompt, or reference images at inference.

This distinction raises a question about how pretrained visual encoders should be adapted. Spatially aligned models such as PE-Spatial provide a strong foundation for dense prediction (Bolya et al., 2025), but spatial features alone do not determine how forensic supervision will use context. Edited and unedited regions within an image share its image-level status while requiring different labels. Such status cues can indicate that an edit exists without specifying its extent, while also helping suppress false alarms on unedited images. Ordinary mask supervision constrains the output but does not uniquely determine the feature dependencies used to predict it, consistent with the broader challenge of shortcut learning (Geirhos et al., 2020). We therefore ask: can training specify not only which mask to predict, but also under which borrowed context that mask remains the target?

We investigate a simple idea: keep the mask fixed while deliberately changing the edit status of the image supplying attention. An edited recipient reads from an unedited donor, and an unedited recipient reads from an edited donor, each drawn from a different underlying original image, or parent. Unlike same-status exchange, which preserves agreement in image-level status, opposite-status exchange deliberately reverses this relationship. Combined with natural training, it subjects the same spatial target to both matching and opposing donor-status conditions. We hypothesize that this task-specific mismatch provides a useful adaptation signal beyond cross-image exchange alone.

We implement this idea as opposite-status context training (OSCT). In an auxiliary forward, recipient queries attend to donor keys and values in PE-Spatial's final blocks, retaining their residual paths and targets. Natural and exchanged-context forwards jointly optimize the existing segmentation loss.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

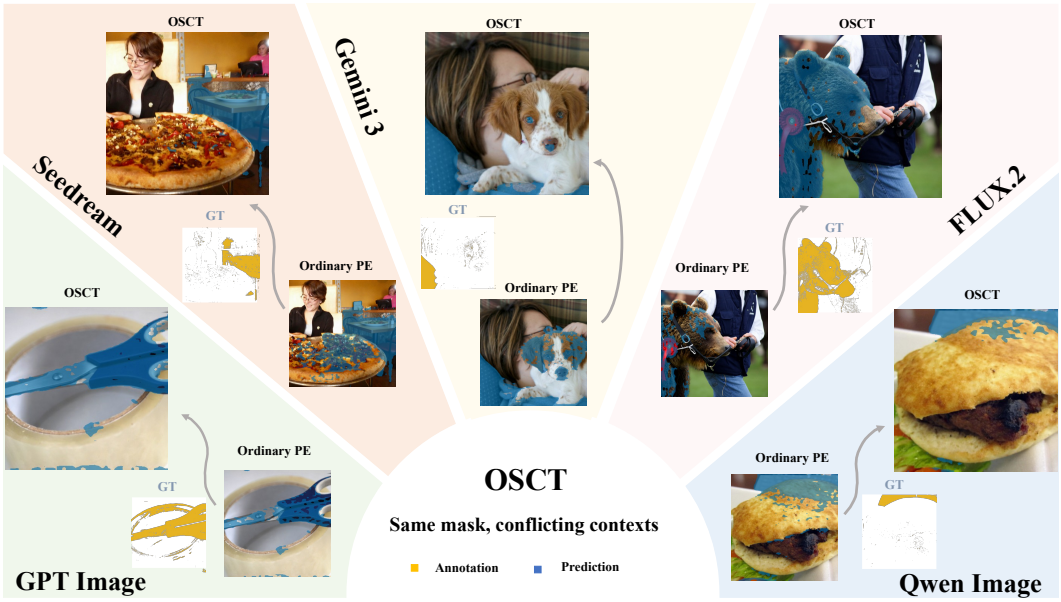

Figure 1: **Single-image AI-edit localization across five editors.** Selected examples show OSCT (large), Ordinary PE (small), and public annotations (GT). Blue denotes predicted edits and gold the annotated region. Arrows connect predictions for the same image.

Inference retains ordinary self-attention, with the baseline's head, parameter count, and single-image computation. The donor supplies a different training condition, not a different answer.

We study OSCT as continued adaptation of one strong, supervised PE-Spatial-G14-448 localizer. Ordinary training, same-status exchange, and OSCT share initialization, data, parameterization, losses, resolution schedules, and training budgets. The exchange policies select different status versions of the same donor parent, directly testing donor selection. Across five continuation seeds, OSCT improves edited-image F1 over both controls on the modern and external evaluation scopes. These controls compare routing policies at a fixed initialization and training budget.

Evaluation uses public annotations in a modern-editing scope excluding known modern downstream-training parents and three external localization collections. OSCT improves F1, IoU, and pixel AP over matched ordinary continuation throughout these comparisons. In the modern evaluation, it improves F1 by 1.32 points and pixel AP by 1.31 points, while reducing pristine-image false-positive area from 0.54% to 0.41%. In the modern and external system-comparison tables, the resulting localizer obtains the highest F1 and IoU among the evaluated public checkpoints. The advantage persists when both trained endpoints receive the same score refit, with unchanged single-image deployment.

Our contributions are threefold:

- A recipient-supervised context-training rule. We introduce OSCT, which combines different-parent, opposite-status key/value routing with natural mask supervision while preserving ordinary single-image inference.
- Evidence for task-conditioned donor selection. Parent-matched comparisons favor opposite-status pairing over same-status exchange in five-seed mean F1 across four scopes, establishing a benefit beyond cross-image exchange alone.
- Strong localization performance across evaluation scopes. OSCT improves over matched ordinary continuation and achieves the highest F1 and IoU among evaluated public checkpoints on the modern and external sets.

2 Related Work

From image detection to edit localization. Generated-image detectors study transfer across generators through augmentation, pretrained features, and reconstruction cues (Wang et al., 2020; Ojha et al., 2023; Wang et al., 2023b); generator diversity is also central to recent detection benchmarks (Zhu et al., 2023; Park & Owens, 2025). Localization additionally requires assigning evidence to individual regions. Established approaches learn manipulation traces (Wu et al., 2019; Cozzolino & Verdoliva, 2020), compression artifacts (Kwon et al., 2021), or noise and boundary cues (Chen et al., 2021; Guillaro et al., 2023). Attention and spatial correlations already support forensic localization (Hu et al., 2020; Hao et al., 2021; Wang et al., 2022; Liu et al., 2022). More recent methods sparsify attention (Su et al., 2025), adapt visual foundations (Yu et al., 2026; Lek et al., 2026), model energy anomalies (Wang et al., 2026), or combine forensic and semantic experts (Nuzhdin et al., 2026). SIDA and PIXAR connect localization with multimodal interpretation (Huang et al., 2025; Shang et al., 2026). OSCT studies a training rule within a strong existing localizer.

Spatial representations and their adaptation. DINOv2, DINOv3, and SigLIP 2 develop reusable visual representations with dense-task transfer (Oquab et al., 2023; Siméoni et al., 2025; Tschannen et al., 2025). PE specifically distinguishes representations suited to image-level and spatial tasks (Bolya et al., 2025). Spatial usefulness also depends on computation: SCLIP and ClearCLIP modify late CLIP operations for dense inference (Wang et al., 2024; Lan et al., 2024). Our question concerns how an existing spatial encoder is supervised during adaptation. OSCT changes its auxiliary training context while retaining ordinary self-attention at inference.

Representation exchange and recipient supervision. Mixup, CutMix, and Manifold Mixup combine observations or features with mixed targets (Zhang et al., 2018; Yun et al., 2019; Verma et al., 2019); MixStyle and CrossNorm vary cross-sample statistics while preserving targets (Zhou et al., 2021b; Tang et al., 2021). Other uses include real/fake feature recombination for deepfake detection (Lin et al., 2024) and stylized-feature cross-attention for medical segmentation (Xu & Zhang, 2024). OSCT routes spatial keys and values across different parents, selecting donor edit status under recipient-mask supervision. FOCAL learns image-relative forensic representations (Wu et al., 2023); interchange intervention training connects internal exchange to behavioral supervision (Geiger et al., 2022). Appendix F provides further context.

3 Method

Overview. OSCT varies training context while retaining the recipient's localization target. A shared localizer combines natural self-attention with an auxiliary forward in which late-layer queries read keys and values from different-parent, opposite-status donors (Figure 2). Both forwards supervise the recipient's own mask; deployment uses the same model with ordinary single-image self-attention.

3.1 Recipient-supervised localization

A training example is (x_i, m_i, g_i, p_i), where x_i is an RGB image, m_i is its annotated binary edit mask, $g_i \in \{0, 1\}$ indicates whether it contains an edit, and p_i identifies its underlying original image, or parent. An unedited example has $g_i = 0$ and $m_i = \mathbf{0}$; an edited example can contain both edited and untouched regions. In the paired training subset, each parent provides an edited version (x_p^e, m_p^e) and an unedited version $(x_p^r, \mathbf{0})$. Status and parent identity determine the training-time routing assignments.

Our localizer uses PE-Spatial-G14-448 (Bolya et al., 2025) and a token-wise residual MLP head. The encoder has a patch size of 14 and token width of 1536. The head maps spatial tokens, L2-normalized and scaled by $\sqrt{1536}$, to logits through a linear branch plus a $1536 \to 128 \to 1$ branch with GELU (Hendrycks & Gimpel, 2016). We adapt the LayerNorm (Ba et al., 2016) parameters in every transformer block, rank-4 LoRA (Hu et al., 2022) on query/value projections, and the head, while freezing the remaining parameters. All variants continue the same supervised checkpoint, whose preparation is described in Appendix E.

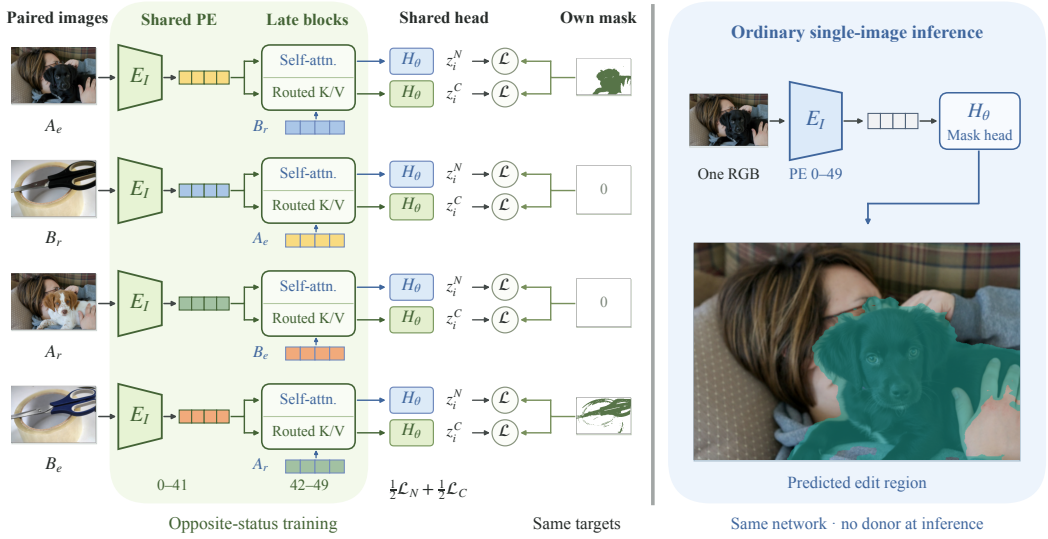

Figure 2: **Opposite-status context training.** Edited (e) and real (r) versions of two parents illustrate routing. Two full forwards share PE and the mask head: natural self-attention and an auxiliary forward whose blocks 42–49 read different-parent, opposite-status K/V. Queries, residuals, and targets stay with the recipient; donor states evolve with gradients. Both losses receive weight $1/2$. Inference uses ordinary self-attention and one RGB image. Photographs illustrate sample roles.

Let $E_{I,\theta}$ and H_θ denote the encoder and head, with θ collecting all adapted parameters. The ordinary prediction is

$$z_i = f_\theta(x_i) = \mathrm{Up}[H_\theta(E_{I,\theta}(x_i))]\,,$$

where Up interpolates spatial logits to the target grid. All images and forward conditions share these parameters.

3.2 Opposite-status donors, unchanged targets

Why opposite status? Same-status exchange changes the image supplying context but preserves agreement in edit status. OSCT reverses that relationship while retaining the target: an edited recipient predicts its own region despite an unedited donor, and an unedited recipient predicts an empty mask despite an edited donor.

Within each paired-source group, we choose donors satisfying

$$p_{\pi(i)} \neq p_i, \qquad g_{\pi(i)} = 1 - g_i. \tag{1}$$

In the row order of Figure 2, a group contains two distinct parents from the same source:

$$X = (x^e_A, x^r_B, x^r_A, x^e_B), \qquad M = (m^e_A, \mathbf{0}, \mathbf{0}, m^e_B).$$

Using one-based indices, $\pi_{\mathrm{opp}} = (2, 1, 4, 3)$, corresponding to $A_e \leftrightarrow B_r$ and $A_r \leftrightarrow B_e$. The assignment is fixed throughout the auxiliary forward. Each image serves as both recipient and donor; parents need no spatial or semantic correspondence.

For the same-status control, $\pi_{\mathrm{same}} = (4, 3, 2, 1)$ selects the alternative status version of the same donor parent for every recipient. This tests the donor-selection policy with the cross-parent assignment held fixed. Identity routing is denoted by $\mathrm{id}(i) = i$.

3.3 LAYER-WISE KEY/VALUE ROUTING

The encoder's blocks are indexed $0, \ldots, 49$. We restrict routing to $\mathcal{S} = \{42, \ldots, 49\}$, so each image first passes through a natural self-attention prefix. Blocks 0–41 retain ordinary self-attention in both forwards, while remaining trainable through the same adaptation parameters.

Let $b \in \{N, C\}$ index the natural and auxiliary forwards. At block ℓ, $Q_i^{b,\ell}$, $K_i^{b,\ell}$, and $V_i^{b,\ell}$ are computed from image i's current state using the encoder's original projections, normalization, and positional processing. The source index is

$$\rho_b^\ell(i) = \begin{cases} \pi(i), & b = C \text{ and } \ell \in \mathcal{S}, \\ i, & \text{otherwise.} \end{cases}$$

For one head of width d_h, attention becomes

$$O_i^{b,\ell} = \mathrm{softmax}\left(\frac{Q_i^{b,\ell} \left(K^{b,\ell}_{\rho_b^\ell(i)}\right)^\top}{\sqrt{d_h}} \right) V^{b,\ell}_{\rho_b^\ell(i)}. \tag{2}$$

The output grid stays with the recipient. Equation (2) produces one output for each recipient query. Routing changes the aggregated key/value sequence while preserving recipient token indices. The head therefore predicts on the recipient's spatial grid, supervised by m_i, and each query can attend to the donor's full sequence.

Only the batch indices supplying keys and values change. Multi-head merging, output projection, LayerScale (Touvron et al., 2021), residual addition, and feed-forward updates follow the original block. Queries and residual streams remain assigned to the recipient. Each branch computes queries from its own evolving state.

Donors evolve at every routed block. All queries, keys, and values are computed from the batch's current states before gathering donor K/V and updating recipients synchronously. No sample is updated in place before another reads its donor. Gradients traverse both paths; later donor states may contain information received at earlier blocks. Algorithm 1 specifies the update order.

3.4 JOINT OPTIMIZATION AND SINGLE-IMAGE DEPLOYMENT

Let $F_\theta(X; \pi)$ denote joint-batch prediction with assignment π applied only in the designated late blocks. For a prepared source group (X_s, M_s), we obtain

$$Z_s^N = F_\theta(X_s; \mathrm{id}), \qquad Z_s^C = F_\theta(X_s; \pi_s).$$

Both predictions are supervised directly by the same recipient masks M_s using the shared segmentation objective.

We retain the baseline segmentation objective. Source groups contain both edited and unedited examples, indexed by $\mathcal{E}$ and $\mathcal{R}$. Their loss is

$$\begin{aligned} \mathcal{L}_{\mathrm{seg}}(Z, M) = {} & \frac{1}{2|\mathcal{E}|} \sum_{i \in \mathcal{E}} \ell_{\mathrm{balBCE}}(z_i, m_i) + \frac{1}{2|\mathcal{R}|} \sum_{i \in \mathcal{R}} \ell_{\mathrm{negBCE}}(z_i) \\ & + \frac{1}{|\mathcal{E}|} \sum_{i \in \mathcal{E}} \ell_{\mathrm{Dice}}(\sigma(z_i), m_i). \end{aligned} \tag{3}$$

Balanced BCE gives equal weight to foreground and background pixel averages within an edited image; negative BCE averages the negative-label loss over an unedited image; soft Dice is computed per edited image. Terms are averaged within image groups. Appendix B gives the shared per-image definitions and numerical conventions.

For fixed source weights $w_s \geq 0$, $\sum_s w_s = 1$, the joint objective is

$$\begin{aligned} \mathcal{L}(\theta) = \sum_s w_s \Big[& \frac{1}{2} \mathcal{L}_{\mathrm{seg}}(F_\theta(X_s; \mathrm{id}), M_s) \\ & + \frac{1}{2} \mathcal{L}_{\mathrm{seg}}(F_\theta(X_s; \pi_s), M_s) \Big]. \end{aligned} \tag{4}$$

Paired modern-source groups use $\pi_s = \pi_{\text{opp}}$; earlier-source replay uses $\pi_s = \text{id}$, remaining natural in both forwards. Appendix B specifies source weights and the continuation recipe.

Both full forwards share prepared inputs and parameters without caching the prefix. Their gradients accumulate across sources before one optimizer update. Ordinary continuation uses identity routing twice; the same-status control changes only auxiliary donor assignments. All variants match initialization, data, adaptation parameters, losses, resolutions, update counts, and forward-pass counts.

Inference. One RGB image passes once through the full encoder with ordinary self-attention and the same head. From the resulting logits z, we obtain $P = \sigma(z)$ and $\widehat{m}(u) = \mathbf{1}[z_u \geq 0]$. No donor, paired original, edit-status label, or additional inference module is required.

4 Experiments

We test localization quality, transfer across public collections, and the contribution of donor status beyond ordinary training or score refitting. Figure 3 compares complete masks from OSCT, matched Ordinary PE, and two public systems: the shared error coloring reveals reduced excess foreground and recovered edit regions in the displayed cases.

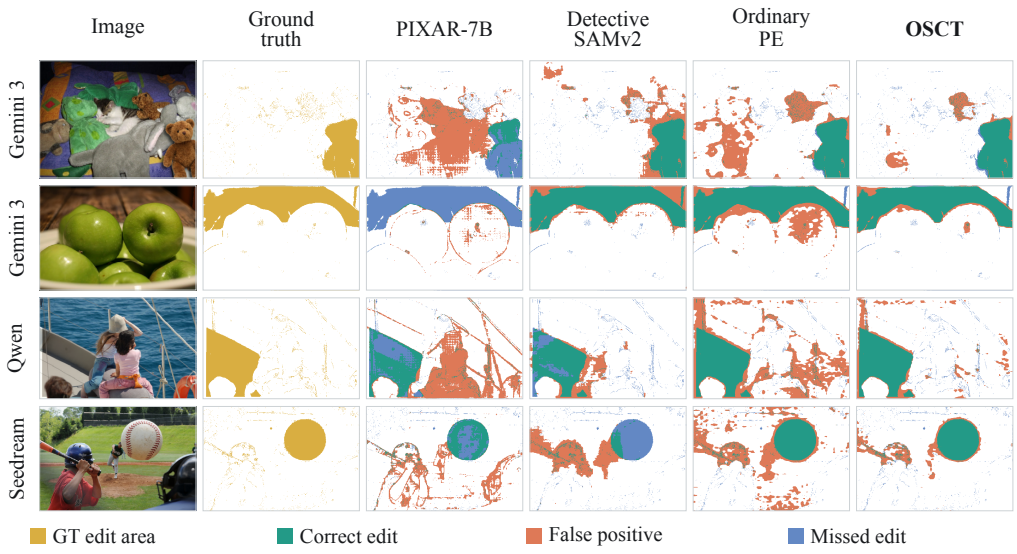

Figure 3: **Qualitative system comparison.** Selected examples from a fixed continuation checkpoint. Gold denotes annotations; predictions use teal for true positives, coral for false positives, blue for false negatives, and white for true background. All methods show the complete image extent. Ordinary PE and OSCT share the training setup.

4.1 Experimental setup

Data and metrics. The modern scope contains 37,162 PIXAR images (Shang et al., 2026): 31,110 edited and 6,052 real, after excluding known modern downstream-training parents. External evaluation uses parent-filtered CocoGlide (Guillaro et al., 2023), AutoSplice JPEG Q100 (Jia et al., 2023), and a fixed SID-Set test subset (Huang et al., 2025). The donor comparison uses these same four scopes. Appendices A and E specify cohort definitions, parent matching, and target assets.

F1 and IoU are averaged per edited image; pixel AP averages edited images containing both label classes. Real FP area averages the predicted-positive pixel fraction on real images. All entries are percentages. PE uses a zero-logit mask threshold and continuous scores for AP. Trainable PE rows report mean $\pm$ SD across five continuation seeds from one fixed supervised initialization; external SDs are in Table 6. Public checkpoints and frozen initialization controls are point estimates.

Training and comparisons. Seven public method families use identical evaluation rows and the pinned inference configurations in Appendix E.5. Ordinary PE, same-status exchange, and OSCT share initialization, data, adaptation parameters, losses, resolutions, 384 updates, and two forwards per update. Both exchange policies use the same donor parent. All runs use the fixed final checkpoint. Inputs are 896 pixels every fourth update and 448 otherwise; PE evaluation uses 896 pixels and returns logits to the annotation grid. Appendices B and E specify training and initialization selection.

Table 1: **Modern AI-edit localization.** Image-macro metrics on 37,162 images excluding known modern downstream-training parents. Real FP area measures false-positive pixel area. PE continuations report five-seed mean ± SD. Bold marks the best column mean.

Method	F1 ↑	IoU ↑	AP ↑	Real FP area ↓
TruFor	11.13	7.10	41.10	3.07
SparseViT	6.79	4.63	41.57	0.64
FLAME	18.70	11.76	39.76	11.77
FUSED	23.67	15.38	37.64	9.96
DINOv3-IML	12.77	8.66	45.51	0.89
DetectiveSAMv2	40.37	26.82	46.42	33.18
PIXAR-7B (raw)	34.87	22.67	47.75	10.30
PIXAR-7B (predicted gate)	27.97	18.45	44.13	1.16
PE initialization	40.56	28.18	51.91	0.81
Ordinary PE	41.26 ± 0.42	27.81 ± 0.34	52.17 ± 0.47	0.54 ± 0.07
OSCT	**42.58** ± 0.41	**29.33** ± 0.33	**53.48** ± 0.45	**0.41** ± 0.06

4.2 Modern AI-edit localization

OSCT improves F1 by 1.32, IoU by 1.52, and AP by 1.31 points over Ordinary PE, while reducing Real FP area from 0.54% to 0.41% (Table 1). Both thresholded masks and pixel ranking improve at the reported operating point. From a shared 40.56-F1 initialization, ordinary continuation reaches 41.26 and OSCT reaches 42.58.

Among the evaluated systems, OSCT leads all four metrics in this scope. Its F1 exceeds DetectiveSAMv2 by 2.21 points and raw PIXAR-7B by 7.71 points, using ordinary single-image inference without a class gate.

Table 2: **External localization.** Edited-image means (%); Continuation rows report five-seed means; SDs are in Table 6. †Overlapping modern-training parents excluded. AutoSplice uses JPEG Q100; SID uses a fixed subset. Bold marks the highest column mean.

	CocoGlide†			AutoSplice Q100			SID subset		
Method	F1	IoU	AP	F1	IoU	AP	F1	IoU	AP
TruFor	36.10	29.32	69.64	65.46	54.39	**94.38**	29.56	21.80	47.52
SparseViT	38.94	32.89	69.29	38.53	29.14	84.95	19.52	15.49	49.50
FLAME	59.97	49.05	72.55	64.46	52.04	83.37	60.11	51.52	69.92
FUSED	63.13	51.16	71.48	58.00	44.90	69.43	31.39	22.53	42.90
DINOv3-IML	67.90	59.67	87.56	50.71	39.14	88.25	41.28	34.70	67.45
DetectiveSAMv2	61.58	50.61	71.28	72.02	59.28	89.43	61.82	50.80	78.36
PE initialization	85.25	78.05	93.88	78.05	67.41	88.94	67.67	57.98	79.87
Ordinary PE	85.61	77.74	94.13	78.43	67.12	89.18	68.04	57.72	80.07
OSCT	**87.47**	**79.97**	**95.33**	**80.21**	**69.63**	90.86	**69.49**	**59.13**	**81.11**

4.3 External localization

OSCT increases F1 over Ordinary PE by 1.86 points on CocoGlide, 1.78 on AutoSplice, and 1.45 on SID (Table 2). All nine F1, IoU, and AP entries improve. AP gains of 1.20, 1.68, and 1.04 points show that the effect extends beyond crossing a fixed mask threshold.

OSCT leads F1 and IoU among the evaluated methods on all three sets and AP on two; TruFor has the highest AutoSplice AP. OSCT also exceeds the shared initialization by 2.22, 2.16, and 1.82 F1 points, respectively. These results require no score refitting.

Table 3: **Parent-matched donor policies.** F1 (%), mean ± SD over five continuation seeds, using exactly the four scopes of Tables 1–2.

Donor policy	Modern	CocoGlide	AutoSplice	SID
Ordinary PE	41.26 ± 0.42	85.61 ± 0.44	78.43 ± 0.51	68.04 ± 0.78
Same-status	40.79 ± 0.43	85.58 ± 0.45	78.31 ± 0.53	67.98 ± 0.79
OSCT	**42.58** ± 0.41	**87.47** ± 0.46	**80.21** ± 0.53	**69.49** ± 0.79

4.4 Does the donor's edit status matter?

Table 3 compares opposite-status and same-status versions of a fixed donor parent. OSCT improves over same-status exchange by 1.79 F1 points on the modern scope, 1.89 on CocoGlide, 1.90 on AutoSplice, and 1.51 on SID. Same-status exchange does not improve mean F1 over Ordinary PE in these scopes. With adaptation capacity, training budget, and donor-parent assignments fixed, the mean advantage across all four scopes supports status-conditioned donor selection beyond cross-image exchange alone.

4.5 Equal-area spatial selection

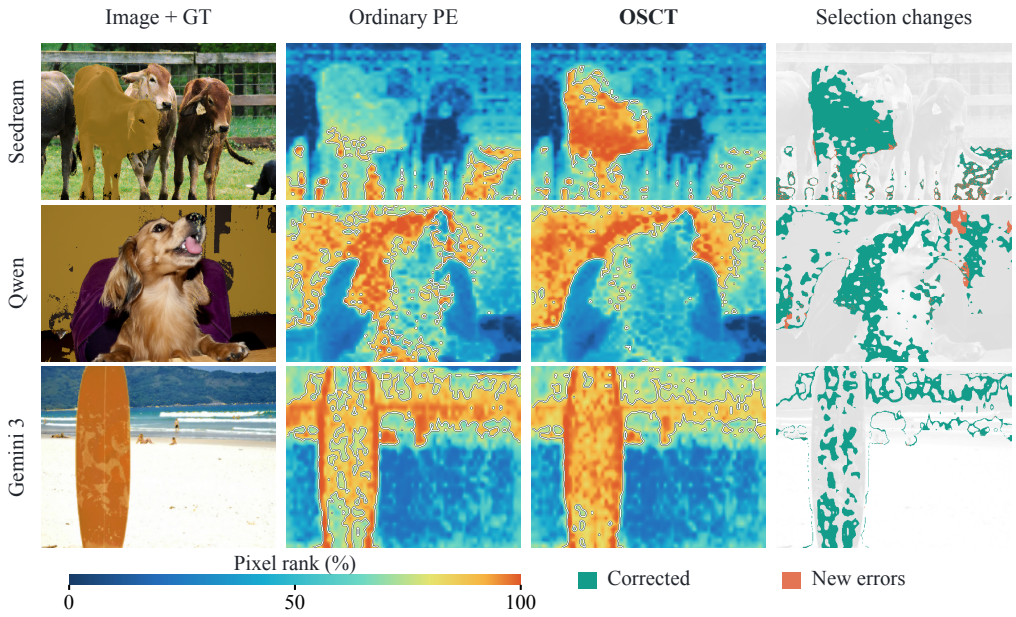

Figure 4: **Equal-area spatial selection.** Selected examples from a fixed continuation checkpoint. Both models select the top 20% of full-image pixels before the common crop. Gold denotes annotations; heatmaps show within-image percentiles; white contours mark selections. Teal marks corrected decisions, coral new errors, and pale gray unchanged decisions.

Figure 4 compares where the models allocate the same foreground budget: each selects its top 20% of full-image pixels, before applying the common display crop. OSCT corrects Ordinary PE's spatial selections in the displayed cases. The final column distinguishes corrected decisions from newly introduced errors, making the selection changes directly visible. The full-evaluation AP gains in Tables 1 and 2 provide the corresponding aggregate ranking evidence.

4.6 BEYOND SCORE REFITTING

Table 4: **Score adaptation before and after continuation.** Same scope as Table 1. Refit freezes each trained endpoint and fits the same source-only score model; OSCT itself uses no refit. †Positive-affine refitting preserves per-image AP, so its mean and SD are inherited. Bold marks the best column mean.

Model / score processing	F1 ↑	IoU ↑	AP ↑	Real FP area ↓
PE initialization	40.56	28.18	51.91	0.81
+ opposite-context calibration	42.04	28.02	51.91	1.28
Ordinary PE	41.26 ± 0.42	27.81 ± 0.34	52.17 ± 0.47	0.54 ± 0.07
OSCT	42.58 ± 0.41	29.33 ± 0.33	**53.48** ± 0.45	0.41 ± 0.06
Ordinary PE + refit	40.74 ± 0.41	27.96 ± 0.34	52.17 ± 0.47†	0.49 ± 0.07
OSCT + refit	**43.62** ± 0.40	**29.72** ± 0.32	**53.48** ± 0.45†	**0.34** ± 0.05

Table 4 separates learning a localizer from adjusting its output scores. Each endpoint refit freezes its model and fits the same three-parameter positive-affine function on matched training-source data, independently per seed and without test labels. Both trained models receive this adjustment, testing whether OSCT's advantage remains after the baseline can change its operating point.

Frozen adaptation and learned localization. At initialization, opposite-context calibration raises F1 from 40.56 to 42.04, while AP remains 51.91 and Real FP area rises from 0.81% to 1.28%. Full OSCT training instead reaches 42.58 F1, 29.33 IoU, and 53.48 AP, with 0.41% Real FP area. Its gain therefore includes improved ranking and lower real-image false-positive area, beyond the effect of this frozen score control. Appendix D gives the remaining initialization-level controls.

Matched endpoint refitting. Raw OSCT already exceeds refitted Ordinary PE in F1 (42.58 versus 40.74), IoU (29.33 versus 27.96), and AP (53.48 versus 52.17), with lower Real FP area (0.41% versus 0.49%). Refitting raises OSCT's F1 by 1.04 points; refitted Ordinary PE reaches 40.74. After both are refitted, OSCT retains a 2.88-point F1 and 1.76-point IoU advantage, with Real FP area of 0.34% versus 0.49%.

Connection to spatial selection. The refit has a positive slope and an image-wide offset, so it preserves within-image pixel order and AP. It also leaves the top-q sets in Figure 4 unchanged. The ranking gains and displayed selection corrections thus capture a different effect from threshold adjustment, while the matched refit results show that localization improvements remain after the same low-capacity score model is fitted to either endpoint.

5 CONCLUSION

OSCT makes the relationship between borrowed context and recipient supervision an explicit training choice. Different-parent, opposite-status donors supply late-layer keys and values while the recipient's mask remains fixed. On the modern-edit evaluation, OSCT improves F1 by 1.32 points and AP by 1.31 points over matched ordinary training, with lower real-image false-positive area. External results, parent-matched donor controls, and matched endpoint score refits support this training rule. Deployment remains an ordinary single-image forward. The central finding is that a strong supervised PE localizer can improve through continuation that preserves the spatial target while reversing the donor's edit status.

AI USE STATEMENT

Generative AI assistants supported literature discovery, method and experiment design, implementation, analysis, visualization, and manuscript drafting. The authors are responsible for the reported experiments and the final text, figures, and claims.

REPRODUCIBILITY STATEMENT

Section 3 specifies the routing operation and training objective. Section 4.1 and Appendix A distinguish the evaluation scopes and averaging conventions. Appendix B documents the continuation recipe, losses, and frozen score controls; Appendix E specifies initialization selection, source splits, evaluation-cohort definitions, and target assets. Algorithm 1 gives synchronous routing; Appendix E.5 specifies public-system inference. Appendix references and Algorithm 1 refer to the separately submitted supplementary material. We provide the core implementation of OSCT.

REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. URL https://arxiv.org/abs/1907.02893.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. URL https://arxiv.org/abs/1607.06450.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision*, pp. 456–473, 2018. URL https://openaccess.thecvf.com/content_ECCV_2018/html/Beery_Recognition_in_Terra_ECCV_2018_paper.html.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025. URL https://arxiv.org/abs/2504.13181.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. URL https://arxiv.org/abs/2104.14294.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. AdaptFormer: Adapting Vision Transformers for Scalable Visual Recognition. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2205.13535.

Xinru Chen, Chengbo Dong, Jiaqi Ji, Juan Cao, and Xirong Li. Image Manipulation Detection by Multi-View Multi-Scale Supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14185–14193, October 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Chen_Image_Manipulation_Detection_by_Multi-View_Multi-Scale_Supervision_ICCV_2021_paper.html.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention Mask Transformer for Universal Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2112.01527.

Davide Cozzolino and Luisa Verdoliva. Noiseprint: A CNN-Based Camera Model Fingerprint. *IEEE Transactions on Information Forensics and Security*, 15:144–159, 2020. doi: 10.1109/TIFS.2019.2916364. URL https://arxiv.org/abs/1808.08396.

Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. RandAugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 702–703, 2020. URL https://openaccess.thecvf.com/content_CVPRW_2020/html/w40/Cubuk_Randaugment_Practical_Automated_Data_Augmentation_With_a_Reduced_Search_Space_CVPRW_2020_paper.html.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision Transformers Need Registers. *arXiv preprint arXiv:2309.16588*, 2023. URL https://arxiv.org/abs/2309.16588.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv:2010.11929*, 2020. URL https://arxiv.org/abs/2010.11929.

Stephanie Fu, Mark Hamilton, Laura Brandt, Axel Feldman, Zhoutong Zhang, and William T. Freeman. FeatUp: A Model-Agnostic Framework for Features at Any Resolution. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2403.10516.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/4f5c422f4d49a5a807eda27434231040-Abstract.html.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 7324–7338, 2022. URL https://proceedings.mlr.press/v162/geiger22a.html.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bygh9j09KX.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2020. doi: 10.1038/s42256-020-00257-z. URL https://arxiv.org/abs/2004.07780.

Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20606–20615, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Guillaro_TruFor_Leveraging_All-Round_Clues_for_Trustworthy_Image_Forgery_Detection_and_CVPR_2023_paper.html.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=lQdXeXDoWtI.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. Hierarchical Fine-Grained Image Forgery Detection and Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3155–3165, June 2023. URL https://openaccess.thecvf.com/content/CVPR2023/

html/Guo_Hierarchical_Fine-Grained_Image_Forgery_Detection_and_Localization_CVPR_2023_paper.html.

Jing Hao, Zhixin Zhang, Shicai Yang, Di Xie, and Shiliang Pu. TransForensics: Image Forgery Localization With Dense Self-Attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15055–15064, October 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Hao_TransForensics_Image_Forgery_Localization_With_Dense_Self-Attention_ICCV_2021_paper.html.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. *arXiv preprint arXiv:2111.06377*, 2021. URL https://arxiv.org/abs/2111.06377.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016. URL https://arxiv.org/abs/1606.08415.

Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1912.02781.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Xuefeng Hu, Zhihan Zhang, Zhenye Jiang, Syomantak Chaudhuri, Zhenheng Yang, and Ram Nevatia. SPAN: Spatial Pyramid Attention Network for Image Manipulation Localization. In *Computer Vision – ECCV 2020*, 2020. doi: 10.1007/978-3-030-58589-1_19. URL https://arxiv.org/abs/2009.00726.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1501–1510, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/Huang_Arbitrary_Style_Transfer_ICCV_2017_paper.html.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. SIDA: Social media image deepfake detection, localization and explanation with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28831–28841, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Huang_SIDA_Social_Media_Image_Deepfake_Detection_Localization_and_Explanation_with_CVPR_2025_paper.html.

Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A. Efros. Fighting Fake News: Image Splice Detection via Learned Self-Consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 101–117, September 2018. URL https://openaccess.thecvf.com/content_ECCV_2018/html/Jacob_Huh_Fighting_Fake_News_ECCV_2018_paper.html.

Philip T. Jackson, Amir Atapour-Abarghouei, Stephen Bonner, Toby P. Breckon, and Boguslaw Obara. Style augmentation: Data augmentation via style randomization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 83–92, 2019. URL https://openaccess.thecvf.com/content_CVPRW_2019/html/Deep_Vision_Workshop/Jackson_Style_Augmentation_Data_Augmentation_via_Style_Randomization_CVPRW_2019_paper.html.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual Prompt Tuning. In *European Conference on Computer Vision*, 2022. URL https://arxiv.org/abs/2203.12119.

Shan Jia, Mingzhen Huang, Zhou Zhou, Yan Ju, Jialing Cai, and Siwei Lyu. AutoSplice: A text-prompt manipulated image dataset for media forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 893–903, 2023. URL https://openaccess.thecvf.com/content/CVPR2023W/WMF/papers/Jia_AutoSplice_A_Text-Prompt_Manipulated_Image_Dataset_for_Media_Forensics_CVPRW_2023_paper.pdf.

Shibo Jie and Zhi-Hong Deng. Convolutional Bypasses Are Better Vision Transformer Adapters. *arXiv preprint arXiv:2207.07039*, 2022. URL https://arxiv.org/abs/2207.07039.

Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9012–9020, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Kim_Learning_Not_to_Learn_Training_Deep_Neural_Networks_With_Biased_CVPR_2019_paper.html.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment Anything. *arXiv preprint arXiv:2304.02643*, 2023. URL https://arxiv.org/abs/2304.02643.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (REx). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5815–5826, 2021. URL https://proceedings.mlr.press/v139/krueger21a.html.

Myung-Joon Kwon, In-Jae Yu, Seung-Hun Nam, and Heung-Kyu Lee. CAT-Net: Compression Artifact Tracing Network for Detection and Localization of Image Splicing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 375–384, January 2021. URL https://openaccess.thecvf.com/content/WACV2021/html/Kwon_CAT-Net_Compression_Artifact_Tracing_Network_for_Detection_and_Localization_of_WACV_2021_paper.html.

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. ClearCLIP: Decomposing CLIP Representations for Dense Vision-Language Inference. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2407.12442.

Trung-Nghia Le, Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. OpenForensics: Large-Scale Challenging Dataset for Multi-Face Forgery Detection and Segmentation In-the-Wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10117–10127, October 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Le_OpenForensics_Large-Scale_Challenging_Dataset_for_Multi-Face_Forgery_Detection_and_Segmentation_ICCV_2021_paper.html.

Gert Lek, Nicolas van Schaik, Chaoyi Zhu, Pin-Yu Chen, Robert Birke, and Lydia Chen. Detective SAM: Adaptive AI-image forgery localization. In *International Conference on Learning Representations*, pp. 22463–22490, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/262dd62fd1bbb30d6a6b4d578f5e65ff-Paper-Conference.pdf.

Li Lin, Xinan He, Yan Ju, Xin Wang, Feng Ding, and Shu Hu. Preserving fairness generalization in deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16815–16825, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Lin_Preserving_Fairness_Generalization_in_Deepfake_Detection_CVPR_2024_paper.html.

Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. PSCC-Net: Progressive Spatio-Channel Correlation Network for Image Manipulation Detection and Localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. URL https://arxiv.org/abs/2103.10596.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Xiaochen Ma, Bo Du, Zhuohang Jiang, Xia Du, Ahmed Y. Al Hammadi, and Jizhe Zhou. IML-ViT: Benchmarking Image Manipulation Localization by Vision Transformer. *arXiv preprint arXiv:2307.14863*, 2023. URL https://arxiv.org/abs/2307.14863.

Owen Mayer and Matthew C. Stamm. Forensic Similarity for Digital Images. *arXiv preprint arXiv:1902.04684*, 2019. URL https://arxiv.org/abs/1902.04684.

Elif Nebioglu, Emirhan Bilgiç, and Adrian Popescu. AI-generated image detectors overrely on global artifacts: Evidence from inpainting exchange. *arXiv preprint arXiv:2602.00192*, 2026. URL https://arxiv.org/abs/2602.00192.

Adam Novozamsky, Babak Mahdian, and Stanislav Saic. IMD2020: A Large-Scale Annotated Dataset Tailored for Detecting Manipulated Images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pp. 71–80, March 2020. URL https://openaccess.thecvf.com/content_WACVW_2020/html/w4/Novozamsky_IMD2020_A_Large-Scale_Annotated_Dataset_Tailored_for_Detecting_Manipulated_Images_WACVW_2020_paper.html.

Anton Nuzhdin, Marcel Worring, and Ivona Najdenkoska. FUSED: Forensic-semantic mixture-of-experts for AI inpainting detection and localization. *arXiv preprint arXiv:2608.28302*, 2026. URL https://arxiv.org/abs/2608.28302.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards Universal Fake Image Detectors That Generalize Across Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24480–24489, June 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Ojha_Towards_Universal_Fake_Image_Detectors_That_Generalize_Across_Generative_Models_CVPR_2023_paper.html.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *arXiv preprint arXiv:2304.07193*, 2023. URL https://arxiv.org/abs/2304.07193.

Jeongsoo Park and Andrew Owens. Community Forensics: Using Thousands of Generators to Train Fake Image Detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8245–8257, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Park_Community_Forensics_Using_Thousands_of_Generators_to_Train_Fake_Image_CVPR_2025_paper.html.

Yusu Qian, Eli Bocek-Rivele, Liangchen Song, Jialing Tong, Yinfei Yang, Jiasen Lu, Wenze Hu, and Zhe Gan. Pico-Banana-400K: A large-scale dataset for text-guided image editing. *arXiv preprint arXiv:2510.19808*, 2025. URL https://arxiv.org/abs/2510.19808.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763. PMLR, 2021. URL https://arxiv.org/abs/2103.00020.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision Transformers for Dense Prediction. *arXiv preprint arXiv:2103.13413*, 2021. URL https://arxiv.org/abs/2103.13413.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment Anything in Images and Videos. *arXiv preprint arXiv:2408.00714*, 2024. URL https://arxiv.org/abs/2408.00714.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Generation Models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2023. doi: 10.1145/3576915.3616588. URL https://yangzhangalmo.github.io/papers/CCS23-DEFAKE.pdf.

Xinyi Shang, Yi Tang, Jiacheng Cui, Ahmed Elhagry, Salwa K. Al Khatib, Sondos Mahmoud Bsharat, Jiacheng Liu, Xiaohan Zhao, Jing-Hao Xue, Hao Li, Salman Khan, and Zhiqiang Shen. From masks to pixels and meaning: A new taxonomy, benchmark, and metrics for VLM image tampering. *arXiv preprint arXiv:2603.20193*, 2026. URL https://arxiv.org/abs/2603.20193.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. URL https://arxiv.org/abs/2508.10104.

Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. URL https://arxiv.org/abs/2105.05633.

Lei Su, Xiaochen Ma, Xuekang Zhu, Chaoqun Niu, Zeyu Lei, and Ji-Zhe Zhou. Can we get rid of handcrafted feature extractors? SparseViT: Nonsemantics-centered, parameter-efficient image manipulation localization through spare-coding transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(7):7024–7032, 2025. doi: 10.1609/aaai.v39i7.32754. URL https://ojs.aaai.org/index.php/AAAI/article/view/32754.

Zhiqiang Tang, Yunhe Gao, Yi Zhu, Zhi Zhang, Mu Li, and Dimitris N. Metaxas. CrossNorm and SelfNorm for generalization under distribution shifts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 52–61, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Tang_CrossNorm_and_SelfNorm_for_Generalization_Under_Distribution_Shifts_ICCV_2021_paper.html.

Sunil Thulasidasan, Gopinath Chennupati, Jeff A. Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/36ad8b5f42db492827016448975cc22d-Abstract.html.

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 32–42, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Touvron_Going_Deeper_With_Image_Transformers_ICCV_2021_paper.html.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features. *arXiv preprint arXiv:2502.14786*, 2025. URL https://arxiv.org/abs/2502.14786.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6438–6447, 2019. URL https://proceedings.mlr.press/v97/verma19a.html.

Feng Wang, Jieru Mei, and Alan Yuille. SCLIP: Rethinking Self-Attention for Dense Vision-Language Inference. In *European Conference on Computer Vision*, 2024. URL https://arxiv.org/abs/2312.01597.

Junke Wang, Zuxuan Wu, Jingjing Chen, Xintong Han, Abhinav Shrivastava, Ser-Nam Lim, and Yu-Gang Jiang. ObjectFormer for Image Manipulation Detection and Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2364–2373, June 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Wang_ObjectFormer_for_Image_Manipulation_Detection_and_Localization_CVPR_2022_paper.html.

Sai Wang, Ye Zhu, Ruoyu Wang, Amaya Dharmasiri, Olga Russakovsky, and Yu Wu. DETER: Detecting Edited Regions for Deterring Generative Manipulations. *arXiv preprint arXiv:2312.10539*, 2023a. URL https://arxiv.org/abs/2312.10539.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-Generated Images Are Surprisingly Easy to Spot... for Now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8695–8704, June 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Wang_CNN-Generated_Images_Are_Surprisingly_Easy_to_Spot..._for_Now_CVPR_2020_paper.html.

Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. OpenSDI: Spotting diffusion-generated images in the open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Wang_OpenSDI_Spotting_Diffusion-Generated_Images_in_the_Open_World_CVPR_2025_paper.pdf.

Yiming Wang, Baiqi Wu, Qingming Li, Jiahao Chen, Tong Zhang, and Shouling Ji. Order within chaos: Capturing intrinsic energy anomalies for AI-manipulated image forgery localization. *arXiv preprint arXiv:2606.02178*, 2026. URL https://arxiv.org/abs/2606.02178.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for Diffusion-Generated Image Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22445–22455, October 2023b. URL https://openaccess.thecvf.com/content/ICCV2023/html/Wang_DIRE_for_Diffusion-Generated_Image_Detection_ICCV_2023_paper.html.

Thomas Wimmer, Prune Truong, Marie-Julie Rakotosaona, Michael Oechsle, Federico Tombari, Bernt Schiele, and Jan Eric Lenssen. AnyUp: Universal Feature Upsampling. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.12764.

Haiwei Wu, Yiming Chen, and Jiantao Zhou. Rethinking Image Forgery Detection via Contrastive Learning and Unsupervised Clustering. *arXiv preprint arXiv:2308.09307*, 2023. URL https://arxiv.org/abs/2308.09307v1.

Yue Wu, Wael Abd-Almageed, and Prem Natarajan. BusterNet: Detecting Copy-Move Image Forgery with Source/Target Localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 168–184, September 2018. URL https://openaccess.thecvf.com/content_ECCV_2018/html/Rex_Yue_Wu_BusterNet_Detecting_Copy-Move_ECCV_2018_paper.html.

Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. ManTra-Net: Manipulation Tracing Network for Detection and Localization of Image Forgeries With Anomalous Features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9543–9552, June 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Wu_ManTra-Net_Manipulation_Tracing_Network_for_Detection_and_Localization_of_Image_CVPR_2019_paper.html.

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2006.09994.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2105.15203.

Peilin Xiong, Honghui Yuan, Junwen Chen, and Keiji Yanai. BRIDGE: Background routing and isolated discrete gating for coarse-mask local editing. *arXiv preprint arXiv:2605.07846*, 2026. URL https://arxiv.org/abs/2605.07846.

Yuheng Xu and Taiping Zhang. Boundless across domains: A new paradigm of adaptive feature and cross-attention for domain generalization in medical image segmentation. *arXiv preprint arXiv:2411.14883*, 2024. URL https://arxiv.org/abs/2411.14883.

Jieming Yu, Qiuxiao Feng, Zhuohan Wang, and Xiaochen Ma. DINOv3 beats specialized detectors: A simple foundation model baseline for image forensics. *arXiv preprint arXiv:2604.16083*, 2026. URL https://arxiv.org/abs/2604.16083.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6023–6032, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Yun_CutMix_Regularization_Strategy_to_Train_Strong_Classifiers_With_Localizable_Features_ICCV_2019_paper.html.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2303.15343.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1710.09412.

Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. UltraEdit: Instruction-based fine-grained image editing at scale. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/05a30a0fc9e6bacdd3abd4ca8508a9e6-Abstract-Datasets_and_Benchmarks_Track.html.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract Free Dense Labels from CLIP. In *European Conference on Computer Vision*, 2022. URL https://arxiv.org/abs/2112.01071.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT Pre-Training with Online Tokenizer. *arXiv preprint arXiv:2111.07832*, 2021a. URL https://arxiv.org/abs/2111.07832.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with MixStyle. In *International Conference on Learning Representations*, 2021b. URL https://arxiv.org/abs/2104.02008.

Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis. Learning Rich Features for Image Manipulation Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1053–1061, June 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Zhou_Learning_Rich_Features_CVPR_2018_paper.html.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f4d4a021f9051a6c18183b059117e8b5-Abstract-Datasets_and_Benchmarks.html.