Purifying Supervised Representations by Folding Selected Difference Subspaces

Anonymous authors
Paper under double-blind review

Abstract

We design a stronger form of domain alignment and bring it inside supervised data, addressing generalization itself — far from its classical habitat of unsupervised domain adaptation. The logic: first select a difference worth erasing — the choice dictated by the generalization we want and by what the data supports; then partition the data accordingly (multiple appearance variants of the same supervision semantics); then align — a discriminator drives the representation through the corresponding transform, so that the representation's semantic coverage extends over the whole region swept by that transform, which can be far larger than the support of the two aligned domain distributions. No additional data is needed; the product is a purified representation that no longer carries the selected difference, together with real generalization gains. We instantiate this as co-manifold alignment (CM): the reference domain keeps its label and keeps the discriminator at work, while variant domains are pushed toward the reference with an anti-label objective. On colored digits with a strong color–label shortcut, CM escapes the shortcut that ERM and DANN consolidate (rule-reversed unseen accuracy 91.5% vs. 80.1%/81.3%); on a channel combination that never appeared in training (pure-blue digits), both baselines degrade — ERM from overfitting, DANN from unanchored alignment — while CM stays accurate. On monocular depth under weather variants — the evaluation scene held out entirely, three weather types never trained, supervision a plain log-depth loss — the objective improves $d1_6$ from 73.3% to 78.8% and closes the train–held-out generalization gap from -18 points to ≈ 0 or beyond. These results provide behavioral support for our coverage account: aligning a selected relation inside supervised data improves generalization along the folded difference direction; the gain comes from the space swept by the alignment transform, not from the two aligned domains.

1 Introduction

Deep networks generalize when the training data cover the relevant feature space. The usual reading of "more data helps" is more sample points; we argue the operative quantity is coverage: when the model's capacity outgrows the coverage of the data, the covered set fragments into isolated islands, the model interpolates freely between them, and generalization deteriorates — this is overfitting read as a coverage phenomenon. Data augmentation is the standard countermeasure, but augmentation sprinkles points along a few hand-picked, discrete directions.

We study a different lever. First select a difference to erase; partition the data into appearance domains accordingly; then align — a discriminator drives the representation through the transform that makes the domains indistinguishable. The point is not the alignment of two domain distributions: in a high-dimensional representation space, two empirical distributions can degenerate into fragmented, overfitting clusters that do not naturally maintain any spanning coverage. What matters is the coverage of the difference subspace: the representation's semantic span comes to cover the whole region swept by the alignment transform. Closing a book, what we gain is the space swept by its pages, not the two covers.

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.

This view differs from domain-adversarial training as used in unsupervised domain adaptation (UDA) in both design and intent. UDA alignment sees samples only; it matches marginals across a source–target split for label transfer. We align inside labeled data, along a relation we select, as a mechanism for purification and generalization. The difference is also mechanical. Bidirectional fooling on a shared encoder lets both domains drift, and with nothing anchoring the game, the discriminator is no longer pressured once the first separable criterion is folded, and the game stalls. CM anchors the game: the reference domain keeps its label, and its gradient flows into the encoder and maintains — even strengthens — the reference representation, keeping the discriminator at work, while each variant domain is pushed toward being judged as the reference. Empirically, the chase keeps acting long after DANN's adversarial pressure has vanished (Section 3).

Contributions:

- A framing of alignment inside supervised data as selectable-difference folding, with a coverage account of why the gain exceeds the aligned distributions (Section 2).
- CM, an anchored one-sided chasing objective, together with a small tool set (adaptive gradient cap, per-relation discriminators, discriminator refresh) that makes the chase sustainable (Section 2).
- Evidence on three tasks: colored digits with a strong color–label shortcut, a color-coverage task with one color per domain, and monocular depth across weather variants including truly unseen weather types (Section 3).

2 Method

2.1 Selectable differences and partitions

Given a labeled dataset, we select one factor of variation to erase (the color of the digit; the weather of the frame) and partition the data into appearance domains $\{a, b_1, \dots\}$ along it. All domains share one encoder E and one task head h trained with the supervised loss $\mathcal{L}_{\text{task}}$; alignment acts on the representation $z = E(x)$ only.

2.2 Co-manifold alignment

A discriminator D is trained to optimally separate the reference domain a from a variant domain b on z, with the usual discriminator loss on detached features, $\mathcal{L}_D = \text{BCE}(D(\bar{z}_a), 1) + \text{BCE}(D(\bar{z}_b), 0)$. The generator-side objective is asymmetric. With $z = E(x)$ and supervised loss $\mathcal{L}_{\text{task}}$:

$$\mathcal{L}_G = \mathcal{L}_{\text{task}} + \lambda_t \Big(\underbrace{\text{BCE}(D(z_a), 1)}_{\ell_a \text{ anchor}} + \underbrace{\text{BCE}(D(z_b), 1)}_{\ell_b \text{ chase}} \Big), \quad \lambda_t = \min\Big(1,\ m(t)\, \frac{\|\nabla_E \mathcal{L}_{\text{task}}\|}{\|\nabla_E(\ell_a + \ell_b)\|}\Big), \tag{1}$$

- ℓ_a (anchor term): samples of a are pushed to be classified as a — the discriminator gradient flows into the encoder unchanged. This keeps the a side discriminable and keeps D at work: once the current criterion is folded, the encoder must supply D with new features of the a-domain samples.
- ℓ_b (chase term): samples of b are pushed to be classified as a (an anti-label objective — gradient reversal, or the direct form written in Eq. 1). On colored digits, gradient reversal and the direct form are interchangeable; on the depth task we use the non-saturating direct form — the gradient is largest where D is most confident and decays where the fooling succeeds.

Thus b continually chases whatever feature of a is not yet aligned; aligning one feature makes ℓ_a drive D's attention to the next. The ideal endpoint of the chase is co-manifold: no discriminator in D's parameter class separates the domains. Only a multi-layer neural discriminator makes this goal meaningful: matching means or low-order moments constrains only the matched statistics, and two high-dimensional manifolds can share all of them while remaining broadly separable.

Why anchoring matters. DANN-style bidirectional fooling makes both domains drift on a shared encoder, and the game stalls once the first criterion is aligned. Empirically, DANN's best checkpoint arrives earliest of all methods and its performance does not improve with longer training; CM's chase keeps producing gains (Section 3).

2.3 Tools that sustain the chase

Three optional tools, used only where needed (Eq. 1 shows the general form with the cap active; without it $\lambda_t \equiv 1$): (i) an adaptive cap rescales the adversarial gradient to at most $m(t)\|\nabla_E \mathcal{L}_{\text{task}}\|$ with m decaying convexly from a cap value to 1 — the constraint engages only when the adversarial term would overwhelm the task gradient; (ii) per-relation discriminators: each variant domain gets its own D against the reference; (iii) discriminator refresh: when a discriminator's gradient stays near zero, restart the chase by resetting it in place (revival) or by adding a fresh discriminator while keeping the old ones (accumulation, dadd) — the two variants perform comparably in our experiments.

2.4 The coverage view

Reading the mechanism as coverage: alignment provides continuity of the same supervision semantics along the domain-difference direction, joining the domains' isolated samples into a swept sheet. Unlike augmentation — a few discrete, hand-specified directions — the folding transform is learned and semantics-preserving. Coverage expands at two levels: each new criterion D finds refines the space spanned by the alignment transform, and each newly designed view of D — each exploitable relation in the data — brings a new region of continuity. A purified representation is one whose coverage is not torn by the selected difference. Removal and extension act on different components and do not trade off: what the folding removes is the domain-discriminative direction itself; what extends is the region over which the representation of the same supervision semantics remains valid.

3 Experiments

We ask two questions: (1) does folding remove the selected difference from the representation in a way that helps on unseen appearances of that difference? (2) how do the dynamics differ from ERM and DANN? Unless noted, all results are mean±std over 5 seeds.

3.1 Colored digits

Design. MNIST digits are colored so that color only marks the domain, glyph only marks the label — the two factors are cleanly separated. We report two settings: a strong-shortcut task, which adds a 98% color–parity correlation as an explicit shortcut trap (the lighter of each domain's two hues marks even digits, so the cue that transfers to unseen hues is the lightness ordering, not hue identity; the reversed rule inverts it), and a color-coverage task with exactly one color per domain. Evaluation uses held-out samples recolored with unseen hues, under both the consistent and the rule-reversed labeling rule; a method that has folded color keeps its accuracy under reversal; a shortcut-dependent method drops. The encoder is a small CNN (3 conv layers + FC, $z \in \mathbb{R}^{128}$, LayerNorm (Ba et al., 2016) at the representation), the head linear, D a 3-layer MLP; batch 32, 20 epochs, Adam (Kingma & Ba, 2015).

Strong shortcut. With the color–parity correlation at 98% (anchor domain 20%, each variant domain 5% of the 60,000 training images), Figure 1 tracks rule-reversed unseen accuracy per epoch. ERM and DANN peak at epoch 1–2 and then degrade as the shortcut consolidates; CM keeps climbing and stays high. Final-epoch reversed accuracy: ERM 80.1±6.3%, DANN 81.3±5.7%, CM **91.5**±5.2%; CM's final epoch already exceeds the baselines' best epoch (89.8%/89.6%), and its best is 95.6±1.3%. The shortcut drop (consistent minus reversed) ends at 14.0/13.3 points for the baselines vs. 4.7 for CM — a 3× gap.

Color coverage. One color per domain: anchor white (20% of the training images), aligned variants pure red and pure green (q% each, q varied), evaluated on pure blue plus gray

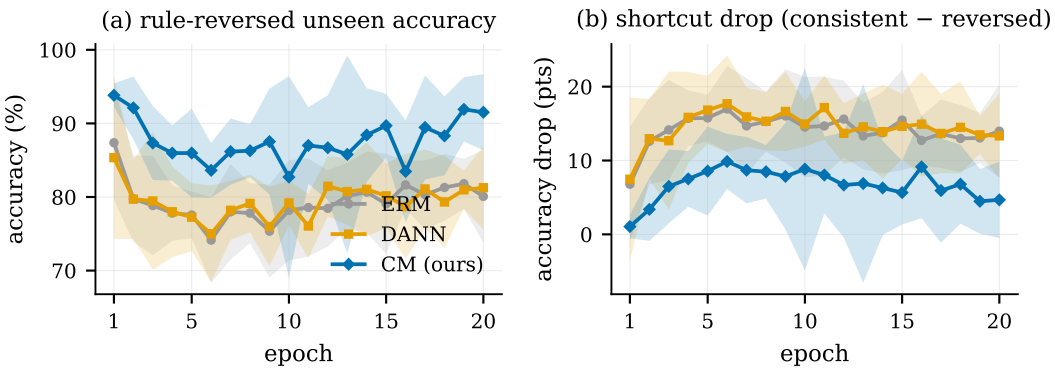

Figure 1: Strong-shortcut task (98% color–parity correlation): per-epoch rule-reversed unseen accuracy (a) and shortcut drop (b), mean±std over 5 seeds. Baselines peak at epoch 1–2 and degrade as the shortcut consolidates; CM is immune.

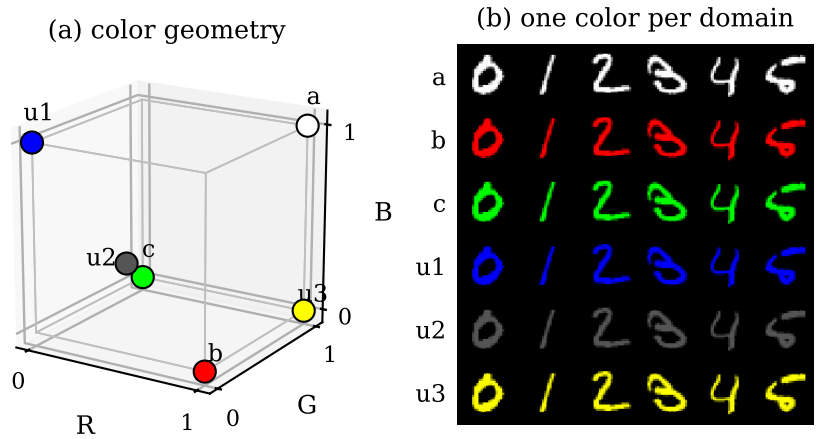

Figure 2: Color-coverage design: (a) the six domain colors in RGB space; blue (u1) is the test color whose channel combination never appears in training. (b) Example images, one color per domain.

and yellow (Figure 2). Blue is the probe that matters: it is a color channel combination that never appears in training (blue high while red and green are low; white activates all channels jointly, red and green activate theirs singly). Gray and yellow lie inside the cone spanned by the training colors and stay accurate under all methods at every scale (5-seed means $\geq 95\%$); we therefore report blue throughout. Figure 3 varies q. CM dominates in the 1%–10% window (e.g., q=5, i.e. 3000 images per variant color: final-epoch blue accuracy 91.9% vs. ERM 81.6%, DANN 69.7%); at 0.2% the red/green samples are too few for ERM to overfit those channels, and ERM catches up. DANN is last at every scale under both final and best protocols — without an anchor, alignment is a net negative on this task. ERM's failure mode is instructive: overfitting to the training data, it learns the seen color combinations as rules, and its accuracy on blue digits degrades even though the glyph itself was learned — the coverage of the glyph has shrunk to the training colors.

3.2 Monocular depth under weather variants (VKITTI)

Design. VKITTI2 (Cabon et al., 2020) frames come in a clean clone version and five weather variants. Training sees clone+rain+sunset only, and evaluation runs on an entirely held-out scene (Scene02); fog/morning/overcast never participate in training, so our main metric $d1_6$ — mean δ_1<1.25 (Eigen et al., 2014) over all six variants on held-out frames — directly tests unseen-scene and unseen-weather generalization. Supervision is a plain masked L1 loss on

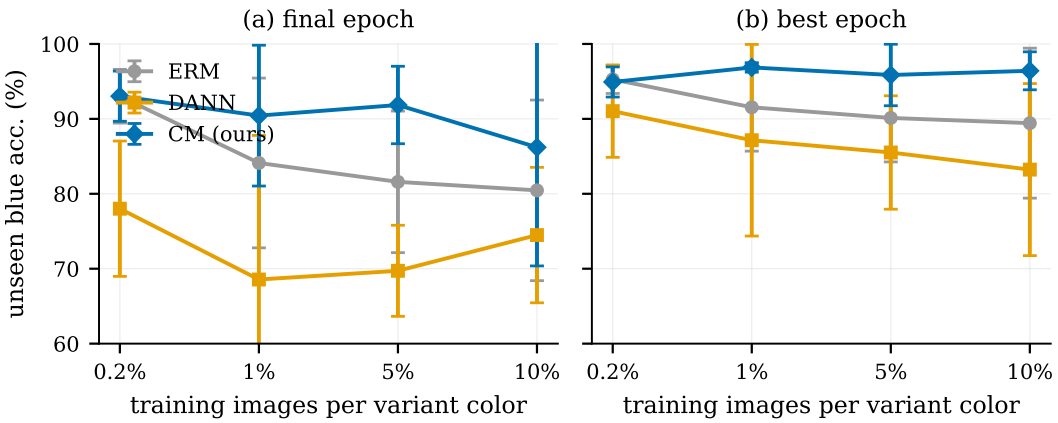

Figure 3: Color-coverage task: unseen-blue accuracy vs. variant scale (5 seeds). (a) final epoch, (b) best epoch. CM dominates in the 1%–10% window; DANN is last at every scale.

log-depth. The model is a pretrained ResNet-101 (He et al., 2016) encoder-decoder with a spatial z (256 channels); each variant domain gets its own patch discriminator against clone; all adversarial arms use the direct anti-label form and the same adaptive cap 2→1; 30 epochs. Baselines are joint training on the three training domains (joint; the ERM counterpart on this task) and DANN. We denote the anchored CM arm la and the anchor-removed arm nola.

Results (Tables 1–2, Figure 4): DANN attains its best checkpoint earliest of all methods, and its performance then decays; overall it is indistinguishable from joint (73.7±1.8 vs. 73.3±0.9). The anchored CM (la) keeps the adversary at work and exceeds both baselines (75.4±1.4). The best results, however, come from removing the anchor-side feature enhancement (nola) and keeping the chase alive mechanically: revival 78.8±2.4 and dadd 78.6±1.3, comparable to each other. Anchor and refresh serve the same end — a discriminator that stays alive and keeps finding new criteria — and the task semantics dictate the choice. On the low-rank digit task the anchor alone suffices: the encoder keeps supplying new a-domain features without disturbing digit identity, and no tool from Section 2.3 is needed. On this high-rank, per-sample task the anchor's amplification of domain-discriminative features fights the supervised semantics — the pressure that keeps D alive also perturbs the depth representation (Section 6) — so we remove it; the price is that the chase dies once one criterion is folded, which is exactly what the threshold-triggered refresh restores.

A decomposition at the same best checkpoint (train side: 100 training frames; held-out side: 100 fixed Scene02 frames; both averaged over the three seen variants; two seeds per arm; absolute values in Table 3) shows: joint has memorized the training frames (train δ_1 93%, held-out 75%, gap −18 points), while CM variants compress the gap to ≈ 0 or flip its sign (revival ≈ 0; nola +9, la +13, dadd +13 points). The flipped signs are not an artifact of a damaged training fit: every adversarial arm's held-out accuracy exceeds joint's in absolute terms. Training frames are exactly what the alignment dismantles, and held-out frames collect the generalization benefit of the purified representation.

4 Analysis

Baselines consolidate the shortcut; CM escapes it. On the strong-shortcut task, ERM and DANN reach their best rule-reversed accuracy at epoch 1–2 and then degrade: training past the first epochs is actively harmful under a 98% color–parity correlation. CM escapes shortcut reliance, and its final epoch still exceeds the baselines' best epoch. The behavioral evidence is consistent with the color direction being folded: accuracy barely drops under rule reversal.

Table 1: VKITTI2 weather-variant depth, best-checkpoint $d1_6$ (mean±std, 5 seeds; evaluation on a held-out scene, averaging six variants including three truly unseen weather types).

method	$d1_6$ best (%)
joint	73.3±0.9
DANN	73.7±1.8
CM (la)	75.4±1.4
CM (nola)	76.6±1.7
CM (nola + revival)	**78.8**±2.4
CM (nola + dadd)	**78.6**±1.3

Table 2: Per-variant breakdown of best-checkpoint δ_1 (%, 5-seed means, held-out-scene frames). seen = clone/rain/sunset (trained), unseen = fog/morning/overcast (never trained). Gains concentrate on morning/overcast and the seen variants; on fog, whose difference from clone lies outside the folded difference, all adversarial arms fall below joint.

	seen			unseen					
method	clone	rain	sunset	fog	morning	overcast	seen-3	unseen-3	$d1_6$
joint	75.0	75.4	74.9	69.1	74.5	71.1	75.1	71.6	73.3
DANN	76.5	81.3	76.3	60.5	78.4	69.1	78.0	69.3	73.7
CM (la)	79.8	82.6	77.1	58.1	82.0	72.9	79.8	71.0	75.4
CM (nola)	80.1	80.7	78.7	65.5	80.9	73.7	79.9	73.4	76.6
CM (nola+revival)	82.3	83.0	80.1	66.8	83.1	77.4	81.8	75.8	78.8
CM (nola+dadd)	82.4	86.6	81.9	57.7	83.8	79.2	83.7	73.5	78.6

Overfitting fragments coverage; alignment maintains it. On the color-coverage task, ERM learns the seen color combinations as rules: the glyph itself is learned, yet accuracy on the unseen combination (blue) still degrades — the coverage of the glyph has shrunk to the training colors. The anchored chase keeps the representation covering the color direction, and unseen-blue accuracy stays high across the 1%–10% regime. DANN, without an anchor, is last at every scale: bidirectional fooling does not maintain coverage either.

Joint training exhibits an illusory high training-set fit. On depth, joint training reaches the highest training-frame δ_1 (93%) — and loses 18 points on held-out frames of the held-out scene. The adversarial arms fit the training frames less well (58–89%) but generalize as well as or better than they fit. What the adversarial alignment dismantles on the training frames is exactly the domain component; only what survives the folding of the domain-difference space can serve as the representation.

Alignment direction determines generalization direction. The per-variant breakdown (Table 2) shows that the gains follow the semantic structure of the selected difference. Aligning sunset to clone folds mainly illumination differences, and morning — the unseen variant dominated by the same kind of illumination difference — receives the strongest unseen-variant gain (joint 74.5 → revival 83.1 / dadd 83.8). Fog — whose difference from clone does not resemble the illumination difference we fold — does not exceed joint under any adversarial arm. Our alignment does not improve every unseen domain — generalization extends along the folded difference direction.

5 Related work

Domain alignment for adaptation. DANN (Ganin et al., 2016) and its lineage — conditional variants (Long et al., 2018), moment-matching objectives (Tzeng et al., 2014; Sun & Saenko, 2016), and separate-encoder designs (Tzeng et al., 2017) — align marginals for label transfer in UDA. We share the adversarial tool (Goodfellow et al., 2014) but not the problem: we align inside supervised data along a selected relation, aiming for purification and general-

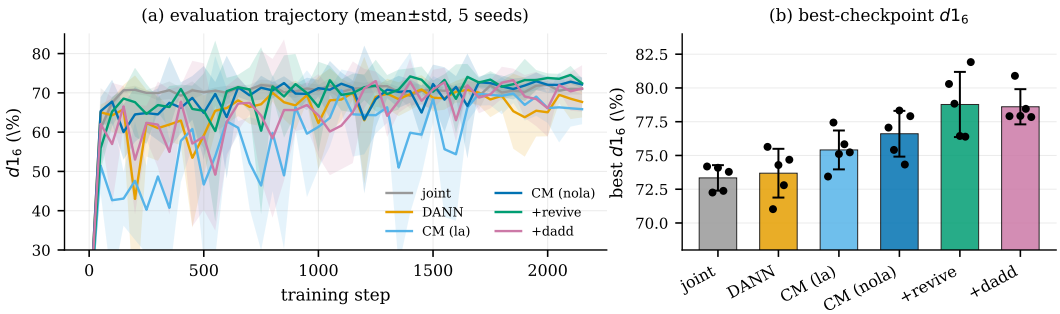

Figure 4: VKITTI2 dynamics (5 seeds). (a) held-out $d1_6$ over training (mean±std); (b) best-checkpoint $d1_6$ per arm, dots = seeds. DANN peaks early and stalls; the refreshed chase (revival, dadd) keeps improving and leads.

ization rather than cross-domain label transfer. Our anchored one-sided objective differs mechanically from bidirectional fooling.

Shortcut learning and robustness. Colored-MNIST-style correlations are the standard probe of shortcut reliance (Arjovsky et al., 2019; Geirhos et al., 2020); our strong-shortcut and color-coverage results add a dynamics view — baselines consolidate the shortcut after epoch 1–2 — and a remedy that removes the shortcut direction from the representation.

Augmentation and coverage. Augmentation covers a few hand-picked directions with discrete points; the folding transform is learned and fills continuously along discriminator-discovered directions.

Domain generalization. Gulrajani & Lopez-Paz (2021) evaluate on held-out domains; our unseen-weather and unseen-color evaluations are in this spirit, with the difference selected and folded at training time, in the hope that the resulting coverage extends to unseen appearances of that difference.

6 Limitations

On high-rank, per-sample semantics the adversarial pressure perturbs the task representation and must be restrained. Alignment may also fail to extend — or even hurt — along difference directions not covered by the trained partition: on fog, whose difference from clone lies outside the folded direction, all adversarial arms fall below joint. The method was conceived to reach co-manifold-level alignment — the two domains indistinguishable under every parameterization of the discriminator — and this ideal state was not achieved. Finally, all three tasks are small-to-mid scale; whether the theory of alignment-extended coverage holds across more tasks and at larger scale remains to be verified.

7 Conclusion

Select a difference, partition the supervised data along it, and let a discriminator drive the representation through the alignment transform: this folds the difference subspace and maintains the coverage of the generalization space rather than merely aligning two distributions. The resulting purified representation escapes shortcut consolidation and overfitting to the seen colors on colored digits, and on depth it generalizes to an unseen scene under unseen weather conditions, at an accuracy that can exceed its own training-frame fit. We see alignment inside supervised data as a general lever for turning relations we have into invariances we want; the space swept by the alignment transform — not the aligned domains — is the real source of the gain. We expect this idea to transfer to broader domains and tasks.

AI use statement

In this work, we used generative AI tools to implement the methods and write the experiment code, and to assist with and polish the manuscript. Additionally, we used generative AI tools to search and summarize relevant literature. The research questions, hypotheses, and experimental designs are the authors' own. We have reviewed all AI-assisted work: all experiment code was executed and its outputs verified. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Reproducibility statement

Implementation details for both tasks — architectures, hyperparameters, data partitions, the checkpoint-selection protocol, and compute — are given in Appendix A. Anonymous source code for the full training and evaluation pipeline is included in the supplementary materials.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. arXiv:1907.02893, 2019.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. arXiv:1607.06450, 2016.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. arXiv:2001.10773, 2020.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In NeurIPS, 2014.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. Journal of Machine Learning Research, 17(59):1–35, 2016.

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, 2:665–673, 2020.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In NeurIPS, 2014.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In ICLR, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In ICLR, 2015.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In NeurIPS, 2018.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In ECCV Workshops, 2016.

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. In arXiv:1412.3474, 2014.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In CVPR, 2017.

Table 3: Train/held-out decomposition at the best checkpoint (absolute δ_1, %; two seeds per arm, means). train = 100 training frames, held-out = 100 Scene02 frames, both averaged over the three seen variants.

method	train	held-out	gap
joint	92.8	74.5	−18.3
DANN	81.7	78.8	−2.9
CM (la)	66.7	80.1	+13.4
CM (nola)	70.7	80.1	+9.4
CM (nola + revival)	82.3	81.3	−1.0
CM (nola + dadd)	70.4	83.6	+13.3

A Implementation details

Colored digits. MNIST train images are partitioned into a rich anchor domain a (20% of the 60,000 training images) and two scarce variant domains b, c (q% each: q=5 in the strong-shortcut task; $q \in \{10, 5, 1, 0.2\}$ in the color-coverage task; 5% $\approx$ 3000 images per color). Each domain is assigned its own color set: in the shortcut task, two nearby hues per domain, the lighter marking even digits with 98% consistency — hue identity differs across domains and between train and test, so the only transferable color cue is the within-domain lightness ordering, which the reversed rule inverts; in the coverage task, exactly one color per domain. Test images are recolored with three unseen hue sets (u_1, u_2, u_3) and evaluated under the consistent and the reversed parity rule; the coverage task additionally evaluates single unseen colors (blue, gray, yellow). Encoder: 3 conv layers $(32, 64, 128)$ stride 2, FC to $z \in \mathbb{R}^{128}$ with LayerNorm; linear head; D is a 3-layer MLP (256 hidden). Structured batches of 32 (4 digits × 2 domains × 4 samples per class–domain cell), identical across all three methods — the methods differ only in the loss, and MNIST train images outside the partition are unused by every method. Adam lr 10^{-3}, 20 epochs. DANN uses its standard 20% warmup ramp; CM needs no warmup in this task. LayerNorm ablation: without it, all three methods collapse on the color-coverage task's blue evaluation (ERM 12%, DANN 12%, CM 24%; single-seed control), while on the strong-shortcut main grid CM's shortcut drop without LayerNorm is 0.7 points (2.0 with; 3 seeds) — the fold itself does not depend on LayerNorm (the shortcut direction is still removed); what depends on the normalized representation geometry is maintaining coverage along the folded direction. Both main-text digit experiments use LayerNorm.

VKITTI2 depth. Frames with all six render variants (clone, rain, sunset in training; fog, morning, overcast never trained). Training frames come from four scenes (Scene01/06/18/20; 1,893 frames); Scene02 (233 frames) is entirely held out for evaluation. Supervision is dense simulated depth, masked L1 on log-depth. Encoder-decoder: ImageNet-pretrained ResNet-101 with a spatial representation z of 256 channels (12×40); the decoder is three convolution blocks (256→128→64→32) with bilinear upsampling and a single-channel head — z is the only encoder–decoder link (no skip connections). One patch discriminator (384 wide, 3 deep; 3×3 hidden convolutions, 1×1 head, one logit per spatial position) per variant domain against clone. Batches mix 26 frames per domain (78 images); Adam lr 5×10^{-4}, 30 epochs, bf16 autocast, cap 2→1 (Eq. 1), i.e. $m(t) = 1 + (2-1)\max(0, 1 - t/(0.8T))^2$: a convex decay from 2 to 1, completed at 80% of training. Refresh trigger: a variant domain's discriminator is restarted (revival) — or a fresh one is appended (dadd) — whenever its adversarial gradient norm stays below 0.1 at two consecutive log points (100 steps). Every 50 steps we evaluate on 100 held-out frames and checkpoint whenever $d1_6$ improves; reported numbers use the best checkpoint. Checkpoint selection and reporting thus share the same 100 held-out frames (early stopping on the evaluation set); the protocol is identical for all arms, and on colored digits we additionally report the final epoch. For the train/held-out decomposition we additionally forward 100 training frames through the best checkpoint under the same protocol; absolute per-arm values are given in Table 3.

Compute. Single RTX 4090; a 20-epoch colored-digit run takes minutes; a 30-epoch depth run ≈1 h.