R^2M: Real-Aware Residual Model Merging for Robust and Generalizable Deepfake Detection

Jinhee Park[1,2*], Guisik Kim[2*], Choongsang Cho[2], and Junseok Kwon[1]

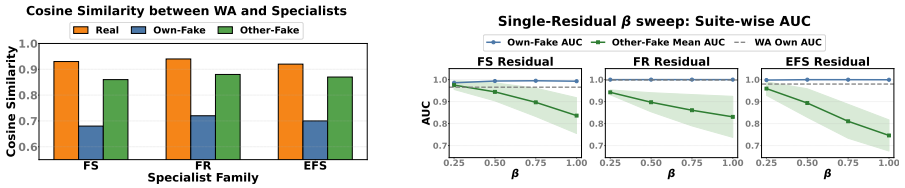
[1] Department of Computer Science and Engineering, Chung-Ang University, Korea
{iv4084em,jskwon}@cau.ac.kr
[2] Korea Electronics Technology Institute (KETI), Korea
specialre@naver.com, ideafisher@keti.re.kr

Abstract. Deepfake generators evolve rapidly, making exhaustive data collection and repeated retraining impractical. Unlike generic multi-task settings, deepfake specialists share a common binary objective (Real vs. Fake) and mainly differ in generator-specific artifacts. However, naive parameter arithmetic can induce unintended decision-boundary shifts, causing unstable ranking behavior and degraded AUC under domain shift. We propose R^2M, a training-free merging framework that decomposes task vectors into a shared component and generator-specific residuals, linking parameter-space updates to logit-space behavior. Offline spectral construction identifies shared and residual subspaces, and online routing selects residuals via first-order gradient-residual alignment, predicting per-sample logit updates under linearization. R^2M is both efficient and interpretable: expensive computations are performed offline, while online inference requires only a single forward–backward pass with lightweight inner-product routing. The same formulation enables diagnostic analysis through margin and routing statistics. Experiments demonstrate consistently strong performance across in-domain, cross-domain, and unseen settings, highlighting R^2M as an interpretable and scalable approach to training-free deepfake model merging.

1 Introduction

Deepfakes have evolved from simple face splicing to photorealistic synthesis driven by modern generative models [5,18,32,40]. Recent systems preserve identity while enabling precise control over lip movements and facial expressions, and high-quality content can now be generated through simple prompts on widely accessible generative platforms and APIs [27,29]. While these advances democratize content creation, they amplify the spread of harmful and deceptive media, including financial fraud, copyright infringement, and political disinformation. These risks underscore the urgent need for robust deepfake detection methods. Because deepfake generation algorithms evolve rapidly, exhaustively collecting per-generator data and retraining a unified detector is fundamentally unsustainable. Even with sufficient data, jointly training on heterogeneous forgeries introduces gradient interference [36, 47]. Conversely, maintaining a dedicated specialist for each generator incurs substantial operational overhead [35, 44].

* Equal contribution.



(a) Weight averaging preserves Real while suppressing Own-Fake alignment.

(b) Residual gains are selective and conflicting.

Fig. 1: Empirical observations motivating selective specialization in deepfake merging. (a) Weight averaging (WA) preserves shared Real structure while attenuating generator-specific fake cues. Cosine similarity between WA and each specialist in the pre-logit space remains high on *Real*, drops on *Own-Fake*, and is relatively higher on *Other-Fake*. (b) Injecting a single specialist residual improves performance on its corresponding forgery family over WA (dashed), but degrades others, exposing an inherent trade-off under unconditional merging. Solid curves show AUC of $\theta_{\mathrm{WA}} + \beta(\theta_i - \theta_{\mathrm{WA}})$ as β varies. Dashed horizontal lines indicate the WA baseline AUC on each dataset. The green shaded band shows the spread of AUC across non-target forgery families (from the lowest to the highest).

To address these challenges, we adopt a *model merging* paradigm for deepfake detection. Concretely, we first fine-tune specialist detectors on generator-specific data and subsequently merge them directly in parameter space to obtain a single unified model. This strategy enables rapid incorporation of new generators without revisiting previous training data or performing costly joint retraining [14, 15, 41]. To the best of our knowledge, this is the first systematic investigation of model merging for deepfake detection.

Observation 1. To validate this design choice, we analyze how a weight-averaged (WA) model [15] relates to its specialists. We construct three specialists corresponding to distinct forgery families in DF40 [44]—FaceSwap (FS), FaceReenactment (FR), and Entire Face Synthesis (EFS)—all finetuned from the same pretrained backbone. A WA model is obtained by averaging their parameters, without additional retraining. We measure feature similarity between specialists and the WA model across Real, Own-Fake(generated by the specialist's corresponding family), and Other-Fake images (generated by the remaining families). Interestingly, a consistent pattern emerges across all families. On Real images, specialists exhibit high similarity to the WA model, indicating that weight averaging preserves the shared Real representation. In contrast, similarity drops markedly on Own-Fake images, revealing that generator-specific signals—strongly encoded in individual specialists—are attenuated in the WA model. In contrast, similarity on Other-Fake images remains substantially higher than on Own-Fake images, as they lack the specialist's dominant fake cues. As illustrated in Fig. 1(a), this behavior implies that WA preferentially retains the shared Real structure while suppressing generator-specific Fake residuals. This indicates that the components suppressed by WA encode generator-specific signals discarded during averaging.

Observation 2. To understand the role of parameters discarded by weight averaging, let θ_i denote the parameters of specialist i and θ_{WA} the weight-

averaged model. We define the residual as $r_i = \theta_i - \theta_{\text{WA}}$ and construct interpolated models $\theta(\beta) = \theta_{\text{WA}} + \beta r_i$, where $\beta \in \mathbb{R}$ controls the residual strength. We evaluate these models across the three forgery families. As shown in Fig. 1(b), increasing β consistently improves performance on the corresponding family relative to WA (dashed), while degrading performance on the others. Although $\beta = 1$ recovers the original specialist, even moderate β values exhibit pronounced trade-offs. This behavior indicates that residuals encode family-specific gains that conflict under unconditional merging. Consequently, no single global parameter setting can simultaneously realize all residual benefits, suggesting that specialization should be applied selectively rather than uniformly.

Proposed R^{2}M. Building on these observations, we tailor merging to the structural properties of deepfake detection. Unlike conventional multi-task settings with disjoint label spaces, all specialists here operate on the same binary label space (Real vs. Fake). In practice, Real cues are largely stable and shared across datasets, whereas Fake cues are generator-specific and heterogeneous. Observations 1 and 2 therefore suggest that Real-aligned directions should be preserved globally, while Fake-specific directions should be activated conditionally to prevent cross-family interference. Accordingly, an appropriate merging strategy should: (i) preserve the shared Real structure, (ii) retain complementary generator-specific Fake knowledge, and (iii) apply Fake-specific gains selectively based on the input. We realize this via *Real-aware Residual Model Merging (R^{2}M)*. We first extract a shared Real-aligned component by applying low-rank factorization (*e.g.* SVD) to the specialists' task vectors, interpreting the dominant directions as a core Real subspace. Each specialist update is then decomposed into a Real-aligned component and a residual encoding generator-specific Fake evidence. To enable efficient conditional specialization, we derive an input-conditional routing rule via a first-order margin approximation. As a result, R^{2}M selects residuals through a single-gradient routing mechanism, eliminating multiple forward/backward evaluations of specialist models while preserving generator-specific discrimination.

Contributions. Our contributions are summarized as follows: (1) We present the first systematic study of model merging for deepfake detection, identifying a structural asymmetry: Real representations are largely shared across specialists, whereas Fake cues are generator-specific and conflict under unconditional merging. (2) We provide a geometric interpretation linking parameter-space factorization to feature-space structure, explaining why weight averaging preserves Real alignment and how R^{2}M decouples Real and Fake effects. (3) We propose an efficient input-conditional routing mechanism derived from a first-order margin approximation, enabling selective residual application with a single backward pass.(4) Extensive experiments on DF40 and cross-dataset benchmarks demonstrate improved generalization and robustness to unseen forgery families.

2 Related works

Deepfake detection. Deepfake technology has advanced rapidly over the past decade. Early methods focused on localized facial manipulations [4, 19, 23], while

GAN-based generators substantially improved realism [8,31,39], followed by diffusion models that further enhanced fidelity and controllability [32]. Beyond face replacement, modern talking-head and reenactment systems preserve identity while enabling natural control over expressions, speech, and head motion [12,17, 25,26]. Commercial platforms such as Veo, Kling, and Wan have further lowered the barrier to high-quality synthetic video generation. While these advances support creative applications, they also introduce significant privacy and safety risks. On the detection side, prior work spans blending-based augmentation [20,22,35], frequency-domain analysis [16, 38, 48], and generalization to *unseen* forgeries [7, 10, 45]. Representative approaches include self-blending for pseudo-fake generation [35], shared forgery feature extraction [45], high-frequency modeling [38], and temporally aware curriculum learning [22]. Adaptor-based designs improve efficiency [10], and vision–language signals have also been explored [37]. More recently, [43] employ an SVD-based decomposition within a vision foundation model to separate pretrained semantic structure from forgery-specific cues, freezing principal components while adapting the orthogonal residual subspace.

Model merging. Model merging combines task-specific experts into a single model by operating directly in parameter space, typically without additional training. A basic strategy is weight averaging (WA) [15], which averages parameters across experts. In large-scale fine-tuning, WA underpins "model soups" and often improves accuracy and robustness without increasing inference cost [41]. Beyond uniform averaging, task arithmetic represents each task as a task vector—the difference between fine-tuned and pretrained weights—and modifies model behavior through vector addition or subtraction; combining multiple task vectors enables multi-task capability [14]. To mitigate interference, TIES-Merging trims small updates, resolves sign conflicts, and merges only sign-consistent parameters, improving multitask performance across modalities [42]. More recently, WA has been revisited from the task-vector perspective. CART centers task vectors around the weight average and applies low-rank approximation to reduce cross-task interference [6]. Orthogonal approaches such as *Fisher-weighted averaging* incorporate local curvature information during merging [24]. Recent work also explores combining model merging with test-time adaptation, where task coefficients or model parameters are adjusted during inference using gradient-based objectives [46].

Mixture-of-Experts routing. Mixture-of-Experts (MoE) models scale capacity by activating only a subset of expert modules per input using a learned router [34], with later variants simplifying routing and improving training stability [11]. Subsequent work improves routing stability and load balancing, including expert-choice routing variants [49]. Recent efforts also explore converting dense checkpoints into MoE-style models for efficient deployment [50].

Positioning of Our Work. To the best of our knowledge, model merging has not been systematically studied for deepfake detection. Existing merging approaches are largely task-agnostic, treating task updates symmetrically. In contrast, deepfake detection exhibits a structural asymmetry: Real representations are largely shared across specialists, whereas Fake cues are generator-specific.

R^{2}M exploits this property by preserving a shared Real component while selectively activating generator-specific residuals, yielding a unified detector that remains composable as new forgery families are introduced.

Our approach is complementary to training-time generalization and subspace factorization methods for AI-generated image detection [43], but targets a different deployment setting: training-free merging of independently fine-tuned specialists. It is also conceptually related to Mixture-of-Experts routing and test-time adaptation, yet R^{2}M neither learns a router nor updates parameters at inference; gradients are used only as an analytic signal for input-conditional residual selection.

3 Method

Let $f_\theta(x) \in \mathbb{R}^2$ denote the logits of a binary deepfake detector parameterized by θ for an input image sample x, corresponding to the classes Real and Fake. We define the classification margin as

$$z(\theta, x) = f_\theta^{\text{fake}}(x) - f_\theta^{\text{real}}(x), \tag{1}$$

so that $z(\theta, x) > 0$ indicates Fake and $z(\theta, x) < 0$ indicates Real. This margin formulation is natural for deepfake detection, which is typically evaluated using ranking-based metrics such as AUC. Let θ_0 denote a pretrained backbone parameter and $\theta_i = \theta_0 + \Delta_i$ denote a specialist parameter trained on forgery family i, where Δ_i is its task vector. We decompose each task vector as $\Delta_i = c + r_i$, where c represents a shared component and r_i captures family-specific residual updates. This decomposition exposes the shared and family-specific structure of task vectors, making (P1) and (P2) analyzable through explicit geometric constraints. We consider a merged model of the form $\theta^*(x) = \theta_0 + c + \beta r_{i(x)}$, where $i(x)$ denotes an input-dependent routing rule selecting a residual component, and β controls the global residual strength. For convenience, we define the shared model: $\theta_c := \theta_0 + c$. We then define the residual-induced margin shift as

$$\Delta z_i(x) = z(\theta_c + \beta r_i, x) - z(\theta_c, x), \tag{2}$$

which quantifies how the residual component r_i perturbs the classification margin relative to the shared model. Finally, let $\mathcal{D}$ denote the overall data distribution, which consists of Real and Fake samples. We denote by $\mathcal{D}_{\text{real}}$ the marginal distribution of Real samples, and by $\mathcal{D}_{\text{fake}}^{(i)}$ the conditional distribution of Fake samples from forgery family i. Then, a desirable merged model should satisfy the following properties:

(P1) Real invariance (residual robustness). $\mathbb{E}_{x\sim\mathcal{D}_{\text{real}}}\big[|\Delta z_i(x)|\big] \le \epsilon$, for each residual r_i, where $\epsilon > 0$ is a small tolerance controlling Real robustness. This condition ensures that residual updates do not significantly perturb the margin on Real samples, preserving the shared Real decision structure and keeping residual effects fake-specific.

(P2) Fake selective gain (own-family amplification). For Fake samples from family i, $\mathbb{E}_{x\sim\mathcal{D}_{\text{fake}}^{(i)}}\big[\Delta z_i(x)\big] \ge m$, and for $j \ne i$, $\mathbb{E}_{x\sim\mathcal{D}_{\text{fake}}^{(i)}}\big[\Delta z_j(x)\big] \le m'$, where $m > m' \ge 0$ are desired margin gains for own- and cross-family samples,

respectively. This enforces that each residual selectively increases the margin for its corresponding forgery family, while exerting limited influence on others.

Design Objective. Our goal is to design a decomposition and routing mechanism that approximately satisfies (P1) and (P2).

3.1 Real–Fake Structured Decomposition

We aim to construct a decomposition $\Delta_i = c + r_i$ that approximately satisfies (P1) real invariance and (P2) fake selective gain. We first characterize the underlying geometric conditions required for these properties, and then show that they naturally admit a spectral solution.

Optimal Residual Geometry under Real–Fake Constraints. For the shared model $\theta_c = \theta_0 + c$, the residual-induced margin shift $\Delta z_i(x)$ is defined in Eq. (2). Applying a first-order Taylor expansion, $z(\theta + \delta, x) \approx z(\theta, x) + \nabla_\theta z(\theta, x)^\top \delta$, we obtain

$$\Delta z_i(x) \approx \beta \, \nabla_\theta z(\theta_c, x)^\top r_i = \beta \langle g(x), r_i \rangle, \tag{3}$$

where $g(x) := \nabla_\theta z(\theta_c, x)$ denotes the gradient of the margin evaluated at the shared model. Under this linearization, the effect of a residual direction r_i is governed by its alignment with the gradient evaluated at the shared model θ_c. In other words, residual-induced margin shifts depend on how strongly r_i aligns with the local geometry of the margin at θ_c.

To analyze **(P1)** in a tractable manner, we consider the expected squared margin shift on Real samples. Minimizing the squared shift provides a smooth surrogate for controlling $|\Delta z_i(x)|$. Using the first-order approximation derived in Eq. (3), $\Delta z_i(x)^2 \approx \beta^2 \langle g(x), r_i \rangle^2 = \beta^2 r_i^\top g(x) g(x)^\top r_i$. Taking expectation over Real samples yields $\mathbb{E}_{x \sim \mathcal{D}_{\text{real}}} \left[\Delta z_i(x)^2 \right] \approx \beta^2 \, r_i^\top \Sigma_R \, r_i$, where Σ_R is the real gradient covariance matrix, *i.e.* $\Sigma_R = \mathbb{E}_{x \sim \mathcal{D}_{\text{real}}} \left[g(x) g(x)^\top \right]$. Thus, Real sensitivity is governed by the quadratic form induced by Σ_R. Let the eigendecomposition be $\Sigma_R = U_R \Lambda_R U_R^\top$. Directions associated with large eigenvalues correspond to parameter perturbations that strongly affect the Real margin.

To satisfy **(P1)** under the squared-shift surrogate, we require each residual to have limited Real sensitivity:

$$r_i^\top \Sigma_R r_i \leq \epsilon_R, \qquad \epsilon_R > 0. \tag{4}$$

A sufficient condition is to remove the dominant eigenspace of Σ_R, motivating the projection step introduced next.

For Fake samples from family i, $\mathbb{E}_{x \sim \mathcal{D}_{\text{fake}}^{(i)}} \Delta z_i(x) \approx \beta \, r_i^\top \mathbb{E}_{x \sim \mathcal{D}_{\text{fake}}^{(i)}} [g(x)]$. Let $\mu_{F,i} := \mathbb{E}_{x \sim \mathcal{D}_{\text{fake}}^{(i)}} [g(x)]$. Then **(P2)** promotes residual directions that increase $r_i^\top \mu_{F,i}$, *i.e.* directions aligned with the mean Fake gradient of family i. Combining this with the Real-invariance constraint in Eq. (4) yields the following constrained trade-off:

$$\max_{r_i} \quad r_i^\top \mu_{F,i} \quad \text{s.t.} \quad r_i^\top \Sigma_R r_i \leq \epsilon_R, \tag{5}$$

which formalizes the requirement that residual directions should align with fake-specific gradients while avoiding dominant real-sensitive modes.

Spectral Interpretation. The constrained optimization in Eq. (5) admits a closed-form solution characterized by the spectrum of Σ_R.

Proposition 1 (Optimal Residual under Real–Fake Trade-off). *Let $\Sigma_R = U\Lambda U^\top$ be the eigendecomposition of the real gradient covariance matrix, and let $\mu_{F,i}$ denote the fake-gradient mean of family i. Consider the constrained optimization problem $\max_r \; r^\top \mu_{F,i} \quad s.t. \quad r^\top \Sigma_R r \le \epsilon_R$. An optimal solution is given by $r_i^\star \propto U\Lambda^\dagger U^\top \mu_{F,i} = \sum_{k=1}^{d} \frac{u_k^\top \mu_{F,i}}{\lambda_k} u_k$, where λ_k and u_k denote the eigenvalues and eigenvectors of Σ_R, and $\Lambda^\dagger$ is the Moore–Penrose pseudoinverse of Λ.*

Implication. The optimal residual scales each fake-gradient component by the inverse of its corresponding Real-sensitivity eigenvalue. Consequently, directions associated with large λ_k (*i.e.* dominant Real-sensitive modes) are suppressed, while alignment with fake-specific gradients is retained. The proof is provided in the Supplementary Material.

While Proposition 1 characterizes the optimal direction for a single family r_i, practical merging requires a shared low-rank residual structure across families. We therefore seek a rank-k subspace $W \in \mathbb{R}^{d\times k}$ such that each residual r_i lies in span(W). A natural surrogate for jointly preserving fake-specific alignment is the aggregated projection energy:

$$\sum_{i=1}^{T} \|W^\top \mu_{F,i}\|^2 = \mathrm{Tr}(W^\top M_F W), \quad M_F := \sum_{i=1}^{T} \mu_{F,i}\mu_{F,i}^\top, \tag{6}$$

which measures how well the subspace W captures the family-wise Fake-gradient means. Using the trace formulation in Eq. (6), we next characterize the optimal low-rank subspace under the Real-sensitivity constraint.

Proposition 2 (Optimal Low-Rank Residual Subspace). *Let $\Sigma_R \succeq 0$ denote the real-gradient covariance matrix, and let M_F be defined in Eq.* (6)*. To obtain a shared rank-k residual structure, we consider the constrained spectral problem: $\max_{W^\top W = I_k} \; \mathrm{Tr}(W^\top M_F W) \quad s.t. \quad \mathrm{Tr}(W^\top \Sigma_R W) \le \epsilon_R$. Then, any optimal solution $W^\star$ spans the top-k generalized eigenspace of (M_F, Σ_R). Equivalently, $W^\star$ is obtained by spectral decomposition of the whitened matrix $\Sigma_R^{-1/2} M_F \Sigma_R^{-1/2}$.*

Implication. The optimal residual subspace arises from a generalized spectral trade-off between fake-specific alignment and Real robustness. Whitening by $\Sigma_R^{-1/2}$ attenuates directions that strongly influence Real margins, while M_F promotes directions shared across Fake families. Consequently, the learned subspace captures fake-discriminative variations that lie in the complement of dominant Real-sensitive modes. The proof is provided in the Supplementary Material.

3.2 Spectrally-Motivated Practical Construction

Propositions 1 and 2 establish that the optimal residual structure is governed by a generalized spectral problem involving the Real-gradient covariance Σ_R and the Fake second-moment matrix M_F. In particular, the optimal residual subspace $W^\star$ corresponds to the top-k eigenspace of the whitened matrix $\Sigma_R^{-1/2} M_F \Sigma_R^{-1/2}$. Directly computing Σ_R and M_F is infeasible in high-dimensional parameter

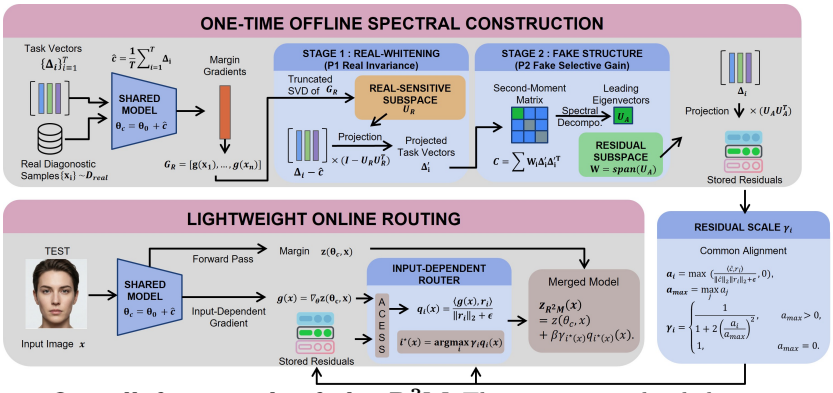


Fig. 2: Overall framework of the R²M The process is divided into two main phases: (Top) **One-time Offline Spectral Construction:** This phase identifies the optimal residual structure through two stages: *Stage 1 (Approximate Real-Whitening)* suppresses dominant Real-sensitive modes to ensure P1; *Stage 2 (Spectral Approximation)* extracts the leading Fake-structured components from projected task vectors to satisfy P2. (Bottom) **Lightweight Online Routing:** For a test sample x, a single backward pass obtains the margin gradient $g(x)$, allowing the input-dependent router to select the most aligned residual $r_{i^*(x)}$ and approximate the merged margin z_{merged}.

space. We therefore construct an efficient low-rank approximation that preserves the same spectral geometry while remaining tractable.

Conceptually, Proposition 2 reveals that our method performs PCA on Fake gradients after removing the geometry induced by Real gradients. The whitening step *(Stage 1)* suppresses Real-sensitive directions, and the subsequent spectral decomposition *(Stage 2)* extracts dominant Fake-structured components in this normalized space. In practice, this corresponds to projecting out dominant Real modes and then performing spectral analysis on the residual task vectors. Overall, the proposed two-stage procedure yields a truncated and sample-efficient approximation to the theoretically optimal generalized spectral solution.

Stage 1: *(Approximate Real-Whitening via Subspace Removal).* Proposition 1 indicates that optimal residual directions should attenuate the dominant eigendirections of Σ_R. Let $g(x) = \nabla_\theta z(\theta_c, x)$ denote the margin gradient evaluated at the shared model. Using a small set of Real diagnostic samples, we construct the gradient matrix $G_R = [\,g(x_1), \ldots, g(x_n)\,], \quad x_j \sim \mathcal{D}_{\text{real}}$. Since $\Sigma_R \approx \frac{1}{n} G_R G_R^\top$, its principal eigenspace can be approximated via truncated SVD of G_R. Let U_R denote the top-r_0 left singular vectors. We first estimate the common component $\hat{c} = \frac{1}{T}\sum_{i=1}^{T} \Delta_i$ and then suppress real-sensitive modes by projecting each centered task vector onto the orthogonal complement: $\Delta_i' = (I - U_R U_R^\top)(\Delta_i - \hat{c})$. This operation approximates whitening by suppressing the dominant eigenspace of Σ_R. Rather than explicitly rescaling by $1/\lambda_k$, we remove the principal Real-sensitive modes, which suffices to enforce (P1) in practice.

Stage 2: *(Spectral Approximation of Fake Second-Moment Structure).* Proposition 2 shows that the optimal residual subspace is determined by the spectral structure of $M_F = \sum_i \mu_{F,i}\mu_{F,i}^\top$. In practice, the true Fake-gradient means $\mu_{F,i}$ are unknown and expensive to estimate precisely. Task vectors Δ_i, being opti-

Algorithm 1 R^{2}M Offline Construction and Calibrated Routing

Input: pretrained parameter θ_0, specialist task vectors $\{\Delta_i\}_{i=1}^{T}$, Real diagnostic samples, ranks r_0, r_A, residual strength β.
Offline construction:

1. Compute the common vector $\hat{c} = \frac{1}{T}\sum_{i=1}^{T} \Delta_i$ and shared model $\theta_c = \theta_0 + \hat{c}$.
2. Estimate the Real-sensitive subspace U_R from diagnostic gradients G_R.
3. Project centered task vectors away from Real-sensitive modes: $\Delta_i' = (I - U_R U_R^\top)(\Delta_i - \hat{c})$.
4. Compute the top-r_A eigenspace U_A of $C = \sum_i w_i \Delta_i' \Delta_i'^\top$ and form residuals $r_i = U_A U_A^\top \Delta_i'$.
5. Compute the positive common alignment $a_i = \max\left(\frac{\langle \hat{c}, r_i\rangle}{\|\hat{c}\|_2 \|r_i\|_2 + \epsilon}, 0\right)$ and $a_{\max} = \max_j a_j$.
6. Store residual scales $\gamma_i = \begin{cases} \frac{1}{1+2(a_i/a_{\max})^2}, & a_{\max} > 0, \\ 1, & a_{\max} = 0. \end{cases}$

Online routing:

1. For test input x, compute $z(\theta_c, x)$ and $g(x) = \nabla_\theta z(\theta_c, x)$.
2. Score each residual by $q_i(x) = \frac{\langle g(x), r_i\rangle}{\|r_i\|_2 + \epsilon}$.
3. Route to $i^\star(x) = \arg\max_i \gamma_i q_i(x)$ and output $z_{\mathrm{R^2M}}(x) = z(\theta_c, x) + \beta\, \gamma_{i^\star(x)} q_{i^\star(x)}(x)$.

mized to minimize family-specific Fake losses, naturally tend to align with $\mu_{F,i}$. We therefore use the projected task vectors Δ_i' as proxies for Fake-gradient directions. To approximate M_F, we construct a weighted second-moment matrix in task space: $C = \sum_{i=1}^{T} w_i\, \Delta_i' \Delta_i'^\top$, where weights w_i capture the degree of family-specific gradient alignment estimated from diagnostic samples. We then compute the top-r_A eigenvectors of C, denoted by U_A. The final subspace $W = \mathrm{span}(U_A)$ serves as a low-rank surrogate for the leading eigenspace of $\Sigma_R^{-1/2} M_F \Sigma_R^{-1/2}$.

3.3 Practical Implementation

As summarized in Fig. 2 and Algorithm 1, R^{2}M consists of a *one-time offline spectral construction* followed by *lightweight online routing*. The offline phase estimates the common vector $\hat{c}$, removes Real-sensitive directions, extracts low-rank Fake-structured residuals, and stores a deterministic calibration scale for each residual branch. The online phase then uses a single margin-gradient computation to select one calibrated residual and update the common margin.

The calibration step accounts for approximation error in the practical construction. Since Real whitening is estimated from finite samples and Fake moments are replaced by task-vector proxies, a residual may still retain *common-aligned leakage*. Such a branch can re-amplify shared evidence rather than contribute complementary Fake-specific evidence. We therefore shrink only residuals with positive alignment to the common vector. This correction is fixed after offline construction and does not introduce an additional learning objective.

The scale γ_i in Algorithm 1 acts as an inverse reliability penalty: the most common-aligned residual receives bounded shrinkage, while orthogonal or oppositional residuals remain unchanged. Since $\{\gamma_i\}$ are computed and stored offline,

inference requires only one gradient computation and T residual inner products per sample.

4 Experiments

Benchmark and protocols (DF40). We evaluated on DF40 [44], a recent deepfake detection benchmark that implements *40* manipulation/generation methods across four *forgery categories*: face swapping (FS), face reenactment (FR), entire-face synthesis (EFS), and face editing (FE). We followed DF40 naming of data "domains" (*e.g.* FaceForensics++/FF [33] and Celeb-DF/CDF [21]); we use FF and CDF as shorthand throughout the experiments. DF40 standardizes three evaluation protocols: (i) *Protocol 1 (cross-forgery, same domain)*: train and test within the same data domain while varying forgery methods; (ii) *Protocol 2 (cross-domain, same forgery)*: train and test on the same forgery category while changing the data domain; (iii) *Protocol 3 (unknown forgery & domain)*: train on seen forgeries/domains and test on *unseen* forgeries and domains to simulate open-set conditions. These protocol definitions, the four-category taxonomy (FS/FR/EFS/FE), and FF/CDF domains were adopted from DF40.

Specialist Model Construction and Training Policy. We trained three specialists, one per forgery category: θ_{FS}, θ_{FR}, and θ_{EFS}, each trained on the union of methods within its category (8 methods per category; 24 total). We also trained a single *all-in-one* model on all 24 methods. All models use CLIP-L/14 [30] as the backbone with `pooler_output` embeddings and a binary linear head trained via cross-entropy. Since our goal is to study training-free model merging for diverse forgery types, we train all specialists under an identical architecture and training recipe to ensure a controlled comparison. All training followed the official DF40 splits, with no data from Protocols 2 or 3 used for training or tuning. Training details and the full method list are provided in the Supplementary Material.

Evaluation metrics. We reported image-level area under the ROC curve (AUC) under all three DF40 protocols and compared merged models against specialists and the all-in-one baseline. All evaluations followed the DF40 train/test splits and protocol definitions. We compared against training-free merging baselines.

Model Merging Hyperparameter Selection. Hyperparameters were selected using Protocol 1 validation only and fixed across Protocols 2 and 3 to prevent cross-protocol leakage. For our method, the hyperparameters consist of two offline ranks and one online scaling factor. r_0 controls the dimension of the Real-sensitive subspace removed in Stage 1 (top-r_0 singular vectors U_R), r_A determines the rank of the residual subspace extracted in Stage 2 (top-r_A eigenvectors U_A of C), and β is the residual scaling coefficient applied at inference time. We searched $r_0 \in \{0, 1, 2\}$ and $r_A \in \{1, 2, 3\}$, and selected β from $\{0, 6, 12, 18, 24, 30\}$. For baseline merging methods, we followed their hyperparameter grids; formulations are detailed in the Supplementary Material. To ensure fair comparison, we applied the same search ranges across all merging methods with the same averaged head. Experiments related to the classifier head are included in the Supplementary Material.

Table 1: DF40 Protocols 1–2 Results (higher is better). Columns are category×domain + domain-wise means. Protocol 1: cross-forgery within FF (in-domain). Protocol 2: cross-domain transfer to Celeb-DF (CDF). Gray cells denote category-matched specialist references, and blue cells denote the jointly trained all-in-one reference. Bold and underline indicate the best and second-best results, respectively, computed only among training-free merging methods.

	Protocol1 - FF				Protocol2 - CDF			
Method	FS	FR	EFS	*Mean*	FS	FR	EFS	*Mean*
Our DF40-compliant training								
Specialist–FS	0.995	0.912	0.766	0.891	0.959	0.690	0.603	0.751
Specialist–FR	0.923	0.999	0.742	0.888	0.505	0.915	0.099	0.506
Specialist–EFS	0.676	0.814	0.999	0.830	0.618	0.621	0.989	0.742
All-in-one	0.962	0.997	0.978	0.979	0.759	0.860	0.366	0.662
Training-free merging baselines								
Weight Averaging	0.968	0.997	0.982	0.982	0.825	0.909	0.767	0.834
Task Arithmetic	0.956	0.995	0.965	0.972	0.741	0.888	**0.926**	0.852
Fisher	0.959	**0.998**	0.949	0.969	0.735	**0.919**	0.442	0.699
CART	0.976	0.994	0.995	0.988	0.851	0.874	0.907	0.877
R^{2}M (ours)	**0.983**	0.995	**0.998**	**0.992**	**0.886**	**0.919**	0.898	**0.901**

Table 2: DF40 Protocol 3 Results (higher is better). *Unseen datasets:* DeepFaceLab [23], HeyGen [1], Midjourney [2], WhichIsReal [3], StarGAN [8], StarGAN v2 [9], StyleCLIP [28], CollabDiff [13].

Method	DeepFace Lab	Hey Gen	Mid journey	WhichIs Real	Star Gan	StarGan v2	Style Clip	Collab Diff	*Mean*
Our DF40-compliant training									
Specialist–FS	0.892	0.573	0.334	0.364	0.887	0.640	0.608	0.596	0.612
Specialist–FR	0.812	0.546	0.603	0.228	0.697	0.433	0.053	0.141	0.439
Specialist–EFS	0.709	0.398	0.561	0.694	0.901	0.777	0.952	0.997	0.749
All-in-one	0.884	0.561	0.177	0.512	0.860	0.714	0.761	0.854	0.665
Training-free merging baselines (same backbone)									
WA	0.930	**0.626**	0.613	0.455	0.953	0.728	0.635	0.848	0.724
TA	0.887	0.541	**0.695**	0.255	0.891	0.579	0.504	0.622	0.622
Fisher	0.896	0.623	0.628	0.377	0.880	0.601	0.346	0.538	0.611
CART	**0.943**	0.608	0.559	0.497	**0.975**	0.780	0.813	0.955	0.766
R^{2}M (ours)	0.909	0.569	0.618	**0.716**	0.958	**0.817**	**0.962**	**0.997**	**0.818**

4.1 Main Results on DF40 (Protocols 1–3)

Tab. 1 summarizes per-category AUCs under Protocols 1 and 2, covering both in-domain (FF) and cross-domain (Celeb-DF) settings.

▶ *Training-free model merging fits deepfake detection.* Even plain *Weight Averaging* achieves a strong FF mean AUC of 0.982, essentially matching the jointly trained *All-in-one* model (0.979). This indicates substantial cross-task parameter sharing in this domain and suggests that weight-space merging is a natural and effective paradigm for deepfake detection.

▶ *CART and R^{2}M retain specialists almost perfectly on FF (saturation).* Both *CART* and *R^{2}M* achieve near-saturated performance on FF (Protocol 1), reaching mean AUCs of 0.988 and 0.992, respectively. Category-wise values closely track the specialists (e.g., FS: 0.976/0.983 vs 0.995; FR: 0.994/0.995 vs 0.999; EFS: 0.995/0.998 vs 0.999). This indicates that the proposed decomposition preserves specialist knowledge instead of washing out task-specific evidence through static averaging. Compared with training-free merging baselines, R^{2}M obtains the best mean AUC, showing that calibrated residual routing can recover category-specific gains without requiring additional training.

▶ *Cross-domain seen transfer (CDF; Protocol 2).* The advantage becomes more pronounced under Protocol 2 (Celeb-DF), where both the domain and forgery

Table 3: Effect of spectral ranks (r_0, r_A) and calibration. We report mean AUCs on sampled subsets of Protocol 1 (FF), Protocol 2 (CDF), and Protocol 3 (Unseen). **Raw** uses unnormalized residual inner products, while **Cal** applies common-alignment calibration. Δ denotes (Cal - Raw).

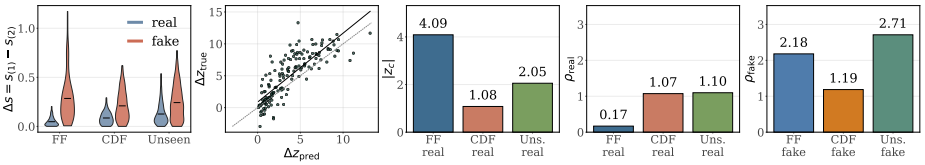
r_0	r_A	FF_{Raw}	FF_{Cal}	Δ	CDF_{Raw}	CDF_{Cal}	Δ	Uns_{Raw}	Uns_{Cal}	Δ
0	1	0.898	0.990	+0.092	0.398	0.830	+0.433	0.696	0.792	+0.096
0	2	0.910	0.974	+0.064	0.499	0.879	+0.380	0.682	0.796	+0.114
0	3	0.913	0.975	+0.062	0.511	0.880	+0.369	0.684	0.798	+0.114
1	1	0.944	0.988	+0.044	0.436	0.851	+0.415	0.705	0.787	+0.082
1	2	0.977	0.987	+0.010	0.588	0.901	+0.314	0.748	0.806	+0.078
1	3	0.984	0.989	+0.005	0.589	0.901	+0.312	0.749	0.812	+0.063
2	1	0.944	0.988	+0.044	0.436	0.851	+0.416	0.708	0.789	+0.081
2	2	0.978	0.987	+0.009	0.588	0.901	+0.313	0.751	0.812	+0.077
2	**3**	0.980	**0.992**	+0.012	0.601	**0.901**	+0.300	0.755	**0.818**	+0.063

distribution differ from the FF training setting. In this regime, individual specialists and static merging baselines exhibit unstable transfer: For example, the FR specialist achieves an AUC of only 0.099 on EFS under CDF, indicating that decision boundaries learned for one manipulation category may not transfer across domain and manipulation shifts. Training-free merging methods mitigate such failures. In contrast, R^2M maintains a robust common component while selectively adding calibrated specialist residuals. As a result, it achieves the highest mean AUC (0.901), suggesting that the proposed routing provides robustness under simultaneous domain and forgery shifts.

Tab. 2 reports per-forgery AUC on *unseen* generators, with the macro mean. ▶ *Merging prevents catastrophic failures and improves macro performance.* The jointly trained *All-in-one* model collapses on several unseen generators (e.g., `MidJourney` AUC = 0.177). This suggests that a decision boundary learned for FR in FF does not transfer to EFS in Unseen domain, indicating significant geometric misalignment across categories and domains. Similarly, the jointly trained All-in-one model, while competitive on FF, drops to 0.366 on EFS under CDF, implying that collapsing heterogeneous forgeries into a single shared boundary can lead to suboptimal compromises under domain shift. In contrast, all training-free merging variants avoid such failures. For example, *Weight Averaging* achieves 0.613 and *CART* 0.559 on `MidJourney`. This shows that merging preserves diverse specialist cues and improves robustness under distribution shift. ▶ *R^2M delivers the strongest overall zero-shot transfer.* Among training-free methods, $\mathbf{R^2M}$ achieves the best macro AUC on unseen generators (0.818), outperforming CART (0.766), Weight Averaging (0.724), and other baselines. R^2M achieves the highest performance on multiple generators, including `WhichIsReal` (0.716), `StarGAN v2` (0.817), `StyleCLIP` (0.962), and `CollabDiff` (0.997). Overall, these results indicate that preserving manipulation-specific residual directions while controlling real-sensitive modes leads to stronger zero-shot generalization across diverse unseen generators.

4.2 Ablation on Spectral Construction and Calibration

Tab. 3 summarizes the mean AUCs across Protocols 1–3 for all (r_0, r_A) configurations, before and after calibration. All results in this rank-grid analysis are obtained with a fixed residual scaling coefficient $\beta = 18$. Therefore, performance differences arise solely from the spectral ranks (r_0, r_A) and the presence or absence of calibration, not from tuning the residual magnitude.



(a) Routing Gap (b) Linearization Check (c) Real Common Margin (d) Real Residual Ratio (e) Fake Residual Ratio

Fig. 3: Empirical analysis of the proposed real–fake decomposition. (a) Calibrated routing confidence gap $\Delta s = s_{(1)} - s_{(2)}$, where $s_i(x) = \gamma_i \langle g(x), r_i \rangle / (\|r_i\|_2 + \epsilon)$ and $s_{(1)} \geq s_{(2)}$ are the top-2 branch scores. (b) Linearization check comparing the predicted residual shift $\Delta z_{\text{pred}} = \beta s_{i^\star}(x)$ with the actual logit shift Δz_{true} after applying the selected calibrated residual update. (c) Common margin magnitude $|z_c|$ on real samples. (d) Median residual-to-common ratio ρ on real samples. (e) Median residual-to-common ratio ρ on fake samples.

Effect of Spectral Ranks (r_0, r_A). Across both Raw and Cal settings, performance generally improves as the spectral ranks increase. Using only one residual direction ($r_A = 1$) is often insufficient, especially under domain shift, while increasing r_A to 3 yields stronger CDF and Unseen performance. Similarly, increasing r_0 from 0 to 2 improves transfer performance, suggesting that removing Real-sensitive directions helps construct robust residual branches. The gains saturate after moderate ranks, with $r_0 = 2$ and $r_A = 3$ achieving the best results.

Effect of Calibration. For every rank setting, Cal consistently outperforms Raw. The improvement is modest on FF, where the test distribution is close to the specialist training domain, but becomes substantial on CDF, where Raw residual routing often drops below 0.60. This indicates that unnormalized residual inner products are sensitive to branch scale and common-aligned leakage under domain shift. Residual normalization and common-alignment calibration mitigate this issue, producing more stable routing and stronger transfer performance. Additional ablations on the residual scaling coefficient β are provided in the Supplementary Material, showing that R^{2}M remains stable over a broad range of residual strengths and is not sensitive to the exact value of β.

4.3 Empirical Analysis of Real–Fake Decomposition

Routing selectivity. We first examine whether the residual branches provide selective evidence for individual inputs. For each sample, we compute the calibrated branch score $s_i(x) = \gamma_i \langle g(x), r_i \rangle / (\|r_i\|_2 + \epsilon)$ and measure the top-2 confidence gap $\Delta s = s_{(1)} - s_{(2)}$. As shown in Fig. 3(a), fake samples have larger routing gaps than real samples across all domains, indicating more selective residual activation for manipulated inputs. The separation is most pronounced on FF, where the real and fake distributions are well matched to the specialist training domains. Under domain shift, the gap between real and fake becomes smaller on CDF and Unseen data, but fake samples still maintain higher routing confidence than real samples. This suggests that residual branches retain manipulation-specific selectivity even beyond the source domain, although domain shift makes routing intrinsically more ambiguous.

Linearization validity. Fig. 3(b) compares the predicted residual shift $\Delta z_{\text{pred}} = \beta s_{i^\star}(x)$ with the actual logit shift obtained by applying the selected residual up-

date. The two quantities are strongly correlated, showing that the gradient inner product is a reliable first-order surrogate for estimating the effect of residual updates. This supports the use of gradient-based residual routing at test time.

Common margin under real-domain shift. Fig. 3(c) reports the common margin magnitude $|z_c|$ on real samples. The common model produces a large real margin on FF, but this margin decreases substantially on CDF and unseen domains. This indicates that cross-domain real distribution shift weakens the shared common detector, making residual calibration more important.

Residual-to-common ratio. Fig. 3(d,e) show the median residual-to-common ratio $\rho = |\beta s_{i^\star}(x)|/(|z_c| + \epsilon)$ for real and fake samples, respectively. On FF, real samples have a small residual ratio, while fake samples have a much larger ratio, suggesting that residual branches primarily contribute manipulation-specific evidence. A similar separation is also observed on Unseen data, where fake samples maintain a substantially larger residual ratio than real samples. This indicates that the residual directions capture fake-specific cues that transfer beyond the training forgery families. On CDF, however, the gap between real and fake ratios becomes smaller because the common real margin collapses. This explains why raw residual routing is unstable under cross-domain transfer: when real and fake residual ratios are less separated, branch scale mismatch and common-aligned leakage can more easily affect routing. Common-alignment calibration mitigate this issue, leading to the large CDF gains observed in Tab. 3.

4.4 Merging Specialists with Heterogeneous Real Domains

In the main experiments above, the merged specialists differ by forgery type (FS, FR, EFS) but are trained under the same FF real domain. Thus, they share a largely common real distribution, and R^{2}M only needs to route among forgery-specific residuals. We further consider a harder stress test where this assumption is relaxed: we merge six specialists, FS/FR/EFS models trained from FF and FS/FR/EFS models trained from CDF. Here, residual branches may carry both forgery-specific evidence and real-domain-specific response scale. As shown in Tab. 4, unrestricted routing over all six residuals preserves CDF performance but degrades FF performance. Unlike a black-box merged detector, R^{2}M exposes the common vector, residual branches, and routing scores, allowing us to diagnose this failure directly.

Table 4: Merging six specialists with heterogeneous real distributions. AUC on FF/CDF domains for individual specialists and merged models. Gray cells indicate the matched specialist reference. R^{2}M-domain uses the same calibrated intra-group routing as R^{2}M after domain grouping.

	FF			CDF		
Model	FS	FR	EFS	FS	FR	EFS
Individual specialists						
FS-FF	0.995	0.912	0.766	0.959	0.696	0.621
FR-FF	0.923	0.999	0.742	0.503	0.923	0.106
EFS-FF	0.676	0.814	0.999	0.635	0.608	0.988
FS-CDF	0.754	0.647	0.616	0.999	0.989	0.982
FR-CDF	0.557	0.849	0.532	0.991	0.999	0.929
EFS-CDF	0.329	0.244	0.923	0.955	0.961	0.999
Merged models						
WA	0.788	0.929	0.910	0.986	0.992	0.979
CART	0.867	0.957	0.954	0.998	0.994	0.996
R^{2}M unres.	0.778	0.922	0.920	0.999	0.978	0.982
domain	**0.929**	**0.988**	**0.979**	**0.999**	**0.997**	**0.997**

Fig. 4 reveals the cause. In Fig. 4(a), CDF-trained residuals activate strongly on FF real samples, while FF-trained residuals activate on CDF real samples. Fig. 4(b) confirms that this response mismatch affects routing: under unre-

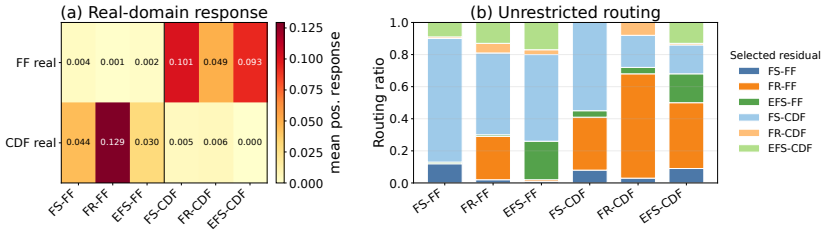


Fig. 4: Analysis of routing under heterogeneous real domains. (a) Mean positive residual response on real samples from each domain. Opposite-domain residuals respond strongly, indicating domain-specific score-scale mismatch. (b) Routing distribution without domain grouping. FF inputs are often routed to CDF residuals, which explains the FF degradation in Tab. 4.

stricted routing, many FF inputs are assigned to CDF-trained residuals (*e.g.* 77% of FS-FF samples are routed to FS-CDF). Thus, the degradation comes from domain-induced score-scale mismatch, not from a lack of useful residuals.

The diagnostic also gives a simple correction. We compute a real-response contrast for each residual, $d_i = (\mathbb{E}_{\text{CDF real}}[q_i] - \mathbb{E}_{\text{FF real}}[q_i])/\text{Std}_{\text{real}}[q_i]$. Residuals with $d_i > 0$ form the FF-trained group, while those with $d_i < 0$ form the CDF-trained group; no fake validation data is used. After selecting the domain-consistent residual pool, we apply the original calibrated three-branch R^{2}M rule, $i^\star(x) = \arg\max_{i \in G} \gamma_i q_i(x)$. This domain-grouped variant recovers FF performance and outperforms CART on all six sampled domains, showing that R^{2}M's interpretable decomposition can turn a failure mode into an effective correction.

5 Conclusion

We study model merging for deepfake detection, where specialists share a binary objective but capture generator-specific artifacts. R^{2}M links parameter-space composition to logit-space behavior, showing that multiple specialist detectors can be consolidated without additional training data. To the best of our knowledge, this is the first systematic study of this setting. On DF40, R^{2}M achieves the best performance among training-free merging methods across in-domain, cross-domain, and unseen-generator evaluations.

Beyond accuracy, R^{2}M exposes an interpretable common–residual decomposition of the merged detector. Its common vector, residuals, and routing scores allow us to diagnose failure modes, as shown in the heterogeneous-domain stress test, and to derive simple extensions such as domain-consistent residual grouping. The main limitation is inference cost: input-level routing requires a gradient computation for each sample. Developing faster routing mechanisms is therefore an important direction for future work.

Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. RS-2024-00456709, No. RS-2025-02218768, No. RS-2022-II220926, No. RS-2026-25525363, No. RS-2021-II211341).

References

1. Heygen. https://www.heygen.com/, accessed 2025-09-25
2. Midjourney. https://www.midjourney.com/, accessed 2025-09-25
3. Which face is real? https://www.whichfaceisreal.com/, accessed 2025-09-25
4. Faceswap. https://github.com/deepfakes/faceswap (2019)
5. Chen, R., Chen, X., Ni, B., Ge, Y.: Simswap: An efficient framework for high fidelity face swapping. In: ACM MM. pp. 2003–2011 (2020)
6. Choi, J., Kim, D., Lee, C., Hong, S.: Revisiting weight averaging for model merging. arXiv preprint arXiv:2412.12153 (2024)
7. Choi, J., Kim, T., Jeong, Y., Baek, S., Choi, J.: Exploiting style latent flows for generalizing deepfake video detection. In: CVPR. pp. 1133–1143 (2024)
8. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: CVPR. pp. 8789–8797 (2018)
9. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. In: CVPR. pp. 8188–8197 (2020)
10. Cui, X., Li, Y., Luo, A., Zhou, J., Dong, J.: Forensics adapter: Adapting clip for generalizable face forgery detection. In: CVPR. pp. 19207–19217 (2025)
11. Fedus, W., Zoph, B., Shazeer, N.: Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research **23**(120), 1–39 (2022)
12. Guo, J., Zhang, D., Liu, X., Zhong, Z., Zhang, Y., Wan, P., Zhang, D.: Liveportrait: Efficient portrait animation with stitching and retargeting control. arXiv preprint arXiv:2407.03168 (2024)
13. Huang, Z., Chan, K.C.K., Jiang, Y., Liu, Z.: Collaborative diffusion for multi-modal face generation and editing. In: CVPR. pp. 6080–6090 (2023)
14. Ilharco, G., Ribeiro, M.T., Wortsman, M., Gururangan, S., Schmidt, L., Hajishirzi, H., Farhadi, A.: Editing models with task arithmetic. arXiv preprint arXiv:2212.04089 (2022)
15. Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., Wilson, A.G.: Averaging weights leads to wider optima and better generalization. arXiv preprint arXiv:1803.05407 (2018)
16. Jeong, Y., Kim, D., Ro, Y., Choi, J.: Frepgan: robust deepfake detection using frequency-level perturbations. In: AAAI. pp. 1060–1068 (2022)
17. Li, J., Zhang, J., Bai, X., Zheng, J., Ning, X., Zhou, J., Gu, L.: Talkinggaussian: Structure-persistent 3d talking head synthesis via gaussian splatting. In: ECCV. pp. 127–145 (2024)
18. Li, L., Bao, J., Yang, H., Chen, D., Wen, F.: Faceshifter: Towards high fidelity and occlusion aware face swapping. arXiv preprint arXiv:1912.13457 (2019)
19. Li, L., Bao, J., Yang, H., Chen, D., Wen, F.: Advancing high fidelity identity swapping for forgery detection. In: CVPR. pp. 5074–5083 (2020)
20. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., Guo, B.: Face x-ray for more general face forgery detection. In: CVPR. pp. 5001–5010 (2020)
21. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-df: A large-scale challenging dataset for deepfake forensics. In: CVPR. pp. 3207–3216 (2020)
22. Lin, Y., Song, W., Li, B., Li, Y., Ni, J., Chen, H., Li, Q.: Fake it till you make it: Curricular dynamic forgery augmentations towards general deepfake detection. In: ECCV. pp. 104–122 (2024)

23. Liu, K., Perov, I., Gao, D., Chervoniy, N., Zhou, W., Zhang, W.: Deepfacelab: Integrated, flexible and extensible face-swapping framework. Pattern Recognition **141**, 109628 (2023)
24. Matena, M.S., Raffel, C.A.: In: Merging models with fisher-weighted averaging. pp. 17703–17716 (2022)
25. Mukhopadhyay, S., Suri, S., Gadde, R.T., Shrivastava, A.: Diff2lip: Audio conditioned diffusion models for lip-synchronization. In: WACV. pp. 5292–5302 (2024)
26. Nirkin, Y., Keller, Y., Hassner, T.: Fsgan: Subject agnostic face swapping and reenactment. In: ICCV. pp. 7184–7193 (2019)
27. Park, J., Owens, A.: Community forensics: Using thousands of generators to train fake image detectors. In: CVPR. pp. 8245–8257 (2025)
28. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: Styleclip: Text-driven manipulation of stylegan imagery. In: ICCV. pp. 2085–2094 (2021)
29. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: A lip sync expert is all you need for speech to lip generation in the wild. In: ACM MM. pp. 484–492 (2020)
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)
31. Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D.: Encoding in style: a stylegan encoder for image-to-image translation. In: CVPR. pp. 2287–2296 (2021)
32. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
33. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. In: ICCV. pp. 1–11 (2019)
34. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538 (2017)
35. Shiohara, K., Yamasaki, T.: Detecting deepfakes with self-blended images. In: CVPR. pp. 18720–18729 (2022)
36. Standley, T., Zamir, A., Chen, D., Guibas, L., Malik, J., Savarese, S.: Which tasks should be learned together in multi-task learning? In: ICML. pp. 9120–9132 (2020)
37. Sun, K., Chen, S., Yao, T., Zhou, Z., Ji, J., Sun, X., Lin, C.W., Ji, R.: Towards general visual-linguistic face forgery detection. In: CVPR. pp. 19576–19586 (2025)
38. Tan, C., Zhao, Y., Wei, S., Gu, G., Liu, P., Wei, Y.: Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In: AAAI. pp. 5052–5060 (2024)
39. Thies, J., Zollhöfer, M., Nießner, M.: Deferred neural rendering: Image synthesis using neural textures. Acm Transactions on Graphics **38**(4), 1–12 (2019)
40. Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., Ortega-Garcia, J.: Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion **64**, 131–148 (2020)
41. Wortsman, M., Ilharco, G., Gadre, S.Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A.S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., et al.: Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In: ICML. pp. 23965–23998 (2022)
42. Yadav, P., Tam, D., Choshen, L., Raffel, C.A., Bansal, M.: In: Ties-merging: Resolving interference when merging models. pp. 7093–7115 (2023)

43. Yan, Z., Wang, J., Wang, Z., Jin, P., Zhang, K.Y., Chen, S., Yao, T., Ding, S., Wu, B., Yuan, L.: Effort: Efficient orthogonal modeling for generalizable ai-generated image detection. arXiv preprint arXiv:2411.15633 **2** (2024)
44. Yan, Z., Yao, T., Chen, S., Zhao, Y., Fu, X., Zhu, J., Luo, D., Wang, C., Ding, S., Wu, Y., et al.: Df40: Toward next-generation deepfake detection. Advances in Neural Information Processing Systems **37**, 29387–29434 (2024)
45. Yan, Z., Zhang, Y., Fan, Y., Wu, B.: Ucf: Uncovering common features for generalizable deepfake detection. In: ICCV. pp. 22412–22423 (2023)
46. Yang, E., Wang, Z., Shen, L., Liu, S., Guo, G., Wang, X., Tao, D.: Adamerging: Adaptive model merging for multi-task learning. arXiv preprint arXiv:2310.02575 (2023)
47. Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., Finn, C.: Gradient surgery for multi-task learning. In: NIPS. pp. 5824–5836 (2020)
48. Zhou, J., Li, Y., Wu, B., Li, B., Dong, J., et al.: Freqblender: Enhancing deepfake detection by blending frequency knowledge. In: NeurIPS. pp. 44965–44988 (2024)
49. Zhou, Y., Lei, T., Liu, H., Du, N., Huang, Y., Zhao, V., Dai, A.M., Le, Q.V., Laudon, J., et al.: Mixture-of-experts with expert choice routing. Advances in Neural Information Processing Systems **35**, 7103–7114 (2022)
50. Zhu, X., Guan, Y., Liang, D., Chen, Y., Liu, Y., Bai, X.: Moe jetpack: From dense checkpoints to adaptive mixture of experts for vision tasks. Advances in Neural Information Processing Systems **37**, 12094–12118 (2024)