FRONTALIZATION HELPS RECOGNITION: DIFFUSION-CONDITIONED SHARED LATENTS FOR POSE-ROBUST FACE FRONTALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Face frontalization, the synthesis of a frontal-view face from an arbitrary pose, is typically treated as an end in itself, evaluated on image quality rather than on whether it actually helps the recognition task it is meant to support. We show that frontalization can be made to directly improve pose-robust face recognition by training the two objectives jointly rather than as separate targets, though the largest benefit of doing so is improved generalization across pose distributions rather than a large head-to-head accuracy gain over strong feature-space methods. We propose **SLF (Shared-Latent Frontalization)**, a diffusion-conditioned shared latent that unifies the two objectives: a spatial representation built from a coarse 3D geometry and texture proxy and an identity embedding, occlusion-masked to handle hair, eyewear, and other self-occlusions. This latent feeds two heads trained end to end, a recognition head operating directly in feature space and a conditional diffusion decoder, distilled to a small number of sampling steps, that reconstructs the frontal image. Training both heads against the same latent produces a modest but consistent recognition-accuracy improvement over training the two objectives with independent trunks, and, more centrally, SLF retains substantially more of its recognition accuracy than an image-level baseline when evaluated on a broader pose distribution than small, narrow-pose-range benchmarks, directly addressing an overfitting pattern reported in prior work. Explicit occlusion masking, by contrast, does not measurably improve image quality in our experiments, indicating the diffusion decoder's generative prior already handles self-occluded regions adequately on its own. We also report inference-time comparisons against optimization-based frontalization methods, since removing per-image test-time optimization is a central goal of our design.

1 INTRODUCTION

Face frontalization aims to synthesize a frontal-view image from an input captured at an arbitrary pose. It is motivated largely by downstream face recognition: pose variation is a dominant source of intra-identity variation in unconstrained recognition, and a reliable frontalization step should normalize it away before recognition is performed. In practice, however, the frontalization and recognition literatures have developed largely independently, and the field has split into two lines of work that optimize for different things and are rarely evaluated against each other directly.

Image-level methods produce an explicit frontal-view image, typically through 3D-morphable-model-guided GAN inversion or flow-based warping. These methods are evaluated primarily on perceptual image quality and, when recognition is reported at all, it is usually a secondary metric. They tend to require either expensive test-time optimization per image or large paired training sets, and recent evidence suggests they may overfit to the small, tightly-controlled pose datasets that the field has historically used for evaluation, degrading when tested on broader, less-curated pose distributions.

Feature-space methods skip pixel-level generation entirely and instead frontalize in an embedding space, optimizing directly for recognition accuracy. These methods currently report the strongest pose-invariant recognition numbers in the literature, but produce no image, and so are unusable for

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.

any application that needs the frontalized face itself (e.g. forensic reconstruction, dataset augmentation, human-facing interfaces).

We argue these failure modes are complementary rather than independent: image-level methods have an explicit geometric target but a weak identity signal, while feature-space methods have a strong identity signal but no geometric grounding. This paper asks whether training a single representation for both objectives jointly, rather than choosing one, improves on each in isolation — specifically, whether pixel-level frontalization can be made to *help* recognition rather than remain a separate deliverable.

We propose **SLF (Shared-Latent Frontalization)**, an architecture built around a single shared spatial latent z that fuses a coarse 3D geometry/texture proxy with an identity embedding, together with an explicit occlusion mask. This latent feeds two heads trained end to end: a recognition head that operates directly on z, and a conditional diffusion decoder, distilled to a small number of sampling steps, that reconstructs the frontal RGB image from z. Because both heads' losses backpropagate into the same z, the representation is forced to remain simultaneously identity-discriminative and geometrically reconstructable, which neither family of prior methods enforces on its own.

Our contributions are:

- A unified architecture in which pose-invariant recognition and frontal image synthesis share a single trained latent representation, rather than being addressed by separate models with separate objectives.
- A diffusion-based image decoder for frontalization, conditioned on an explicit occlusion mask and a 3D geometry/texture proxy, distilled to a small number of sampling steps to remove the per-image test-time optimization cost of prior GAN-inversion-based methods.
- An evaluation directly comparing against both feature-space and image-level state-of-the-art baselines under matched protocols, together with a test of generalization across a broader pose distribution than small, commonly-used benchmarks, to probe an overfitting failure mode identified in recent work.

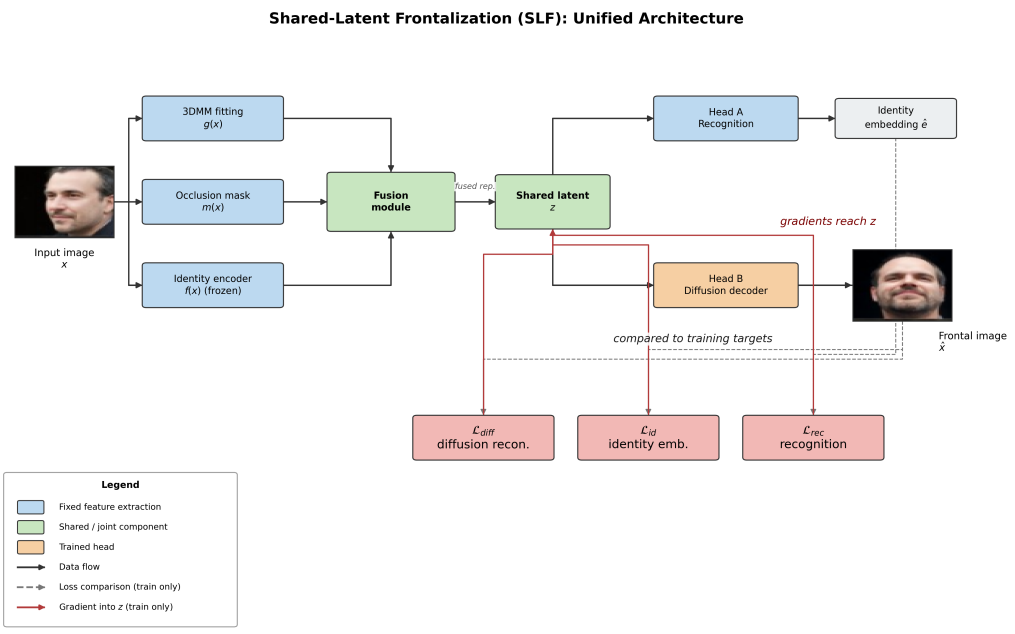


Figure 1: Overview of the proposed architecture. A shared latent z, fused from a coarse 3D geometry/texture proxy and an identity embedding, feeds a recognition head and a conditional diffusion decoder trained jointly.

2 Related Work

Image-level face frontalization. Early frontalization methods combined 3DMM fitting with GAN-based texture completion, using the fit to establish geometry and a generative prior to fill occluded regions (Yin et al., 2017). Flow-based warping instead learns a dense correspondence field to a canonical frontal pose without an explicit 3D fit, and serves as the state-of-the-art image-level baseline in recent comparative work (Wei et al., 2020). GAN-inversion methods project the input into a pretrained generator's (e.g. StyleGAN2) latent space and iteratively optimize it to match visible regions while hallucinating occluded ones, typically at the cost of per-image test-time optimization (Gecer et al., 2021); more recent variants apply parameter-efficient transfer learning to reduce this cost, or target extreme, self-occluding poses directly via statistical texture modeling (Zeng et al., 2025).

Feature-space and recognition-oriented frontalization. A separate line of work skips pixel synthesis and instead learns to transform pose-variant feature embeddings toward a pose-invariant (canonical) feature space, optimizing recognition metrics directly. Recent work in this direction reports pose-invariant recognition accuracy exceeding image-level methods across standard verification benchmarks, using an attention-guided fine-tuning stage on top of a pretrained identity encoder (Stanishev et al., 2025). Related work explicitly comparing "defrontalization" against image-level baselines found that gains hold on larger datasets but not on small, narrow-pose-range benchmarks at extreme angles, raising the possibility that image-level methods evaluated only on such benchmarks overfit to them (Mesec & Jović, 2025).

Diffusion models for face synthesis. Diffusion models have become the dominant generative backbone for face restoration, inpainting, and identity-preserving editing, offering improved texture fidelity over GAN inversion and, with recent distillation techniques, competitive inference cost (Wang et al., 2025). Conditioning mechanisms explored in adjacent tasks include identity embeddings (Santos et al., 2026), segmentation/occlusion masks, and rendered 3D geometry proxies as side inputs to the denoising network (Ding et al., 2023), and ArcFace/CLIP cross-attention adapters for identity-consistent generation (Rizwan et al., 2026). To our knowledge, no prior work applies a diffusion decoder to pose frontalization specifically while jointly training against a recognition objective on the same latent; this is the gap our method addresses.

Positioning. Our method differs from image-level work by conditioning generation on a latent also optimized for recognition, and from feature-space work by requiring that latent to remain sufficient for plausible image reconstruction — a direct test of whether the two objectives are complementary (Section 1) rather than in tension.

3 Method

3.1 Overview

Given an input face image x at an arbitrary pose, our goal is to produce (i) a pose-invariant identity embedding e for verification/identification, and (ii) a frontal-view RGB image $\hat{x}$ consistent with x's identity. Both are derived from a single shared latent $z \in \mathbb{R}^{B\times C\times H\times W}$, computed by a fusion module from geometric and identity features extracted from x (pipeline: Figure 1; decoder detail: Figure 2).

The central design commitment is that z is neither produced by either head in isolation nor frozen once computed: gradients from both the recognition and diffusion losses flow back into the fusion module that produces it. This departs from both families of prior work (Section 2): feature-space methods construct a recognition-only embedding with no signal against discarding fine-grained spatial or textural detail, while image-level methods (3DMM parameters, GAN-inverted codes, or U-Net bottlenecks) are judged solely on reconstruction, with no pressure toward identity-linear-separability. Our hypothesis, tested empirically in Section 4, is that satisfying both constraints at once is mutually reinforcing rather than merely compatible: reconstruction pressure forces z to retain fine-grained structure a recognition-only objective would discard as noise, while the identity-discriminative pressure constrains it against drifting toward plausible-looking faces of the wrong person.

We now describe each stage of the pipeline in turn: geometry and identity feature extraction (Section 3.2), the fusion module that produces z (Section 3.3), the recognition head (Section 3.4), the diffusion decoder (Section 3.5), the training objective that couples all three components (Section 3.6), and the distillation procedure used to make the decoder practical at inference time (Section 3.7).

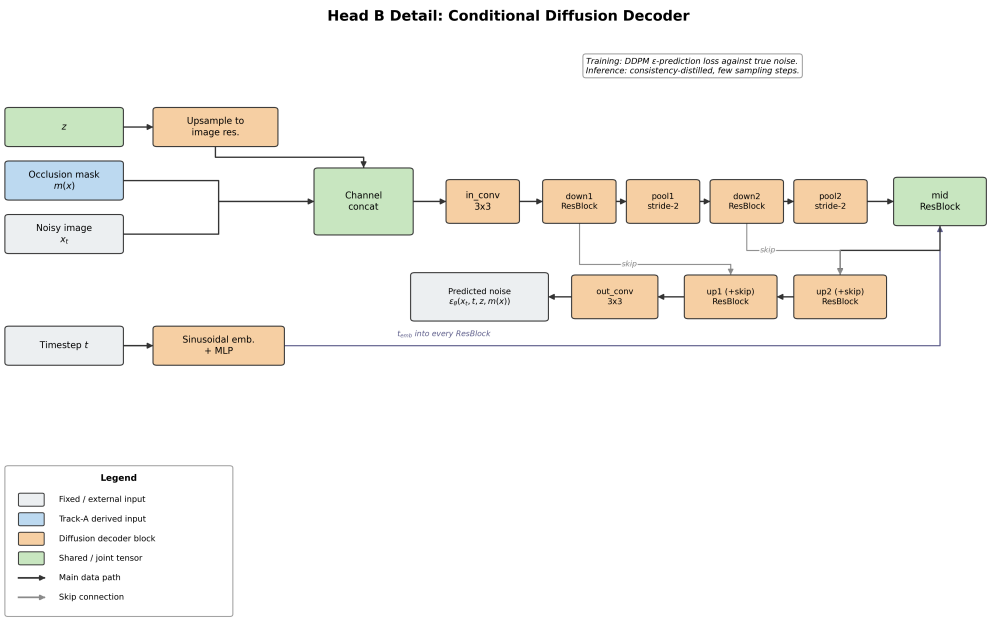


Figure 2: Detailed architecture of the conditional diffusion decoder. The shared latent z is spatially upsampled and concatenated with the noisy image x_t and occlusion mask $m(x)$ at each denoising step; a timestep embedding is injected into every residual block of the U-Net.

3.2 Geometry and identity feature extraction

We fit a 3D morphable model (3DMM) to the input image to obtain a coarse geometry estimate and a rendered UV texture map, $g(x)$. Concretely, we use an off-the-shelf 3DMM fitting network to recover shape and expression coefficients, from which we render a UV-space texture map by sampling pixel values from the input image at the corresponding predicted surface coordinates. This texture-from-photo strategy is a deliberate design choice: an alternative would be to model the face's albedo parametrically (as some 3DMM pipelines support), but doing so requires access to a separate parametric appearance model that is itself subject to its own licensing restrictions, independent of the geometry model. Sampling texture directly from the input photo avoids this second dependency entirely, and has the additional benefit of preserving person-specific high-frequency texture detail (skin texture, localized markings) that a low-dimensional parametric albedo model would smooth away. The cost of this choice is that the resulting UV texture is only defined, and reliable, over regions of the face visible in the input image, which is precisely why an explicit occlusion mask is required downstream.

In parallel, we extract an occlusion/segmentation mask $m(x) \in \{0,1\}^{H\times W}$ identifying regions such as hair, eyewear, and other self-occlusions that should not be treated as reliable texture evidence. This mask serves two roles in the pipeline: it is passed directly as a conditioning input to the diffusion decoder (Section 3.5), and it implicitly marks which regions of $g(x)$'s sampled texture are untrustworthy and must instead be hallucinated by the decoder's learned generative prior rather than copied from the input.

Finally, we extract an identity embedding $f(x) \in \mathbb{R}^d$ from a pretrained face recognition backbone. Unlike the geometry and texture proxy, this embedding is pose-invariant by construction (that is precisely the property the pretrained recognition backbone was trained to have), and serves as the

fusion module's primary identity signal, complementing the geometry proxy's spatial and structural information.

We treat $g(x)$, $m(x)$, and $f(x)$ as fixed, pretrained feature extractors in the experiments reported in this paper, updating only the fusion module and the two downstream heads during training; whether jointly fine-tuning the geometry and identity backbones themselves yields further gains is left as a direction for future, larger-scale work (Section **??**).

3.3 Fusion module and the shared latent

A fusion module ϕ combines the geometry proxy $g(x)$ and the identity embedding $f(x)$ into the shared spatial latent:

$$z = \phi(g(x), f(x)), \qquad z \in \mathbb{R}^{B\times 128\times 16\times 16}. \tag{1}$$

Architecturally, ϕ first projects the UV texture map through a small convolutional stem to a spatial feature map at the target 16×16 resolution, and separately projects the concatenated identity, shape, and expression coefficients through a linear layer to a feature vector, which is broadcast spatially and concatenated channel-wise with the texture feature map. A final pair of 1×1 convolutions merges the two feature streams into the fused latent z.

We deliberately use a spatial latent rather than a flattened vector, at a modest increase in the parameter and compute cost of both downstream heads relative to a compact vector representation. This choice follows directly from the requirements of the decoder: a flattened vector representation discards positional correspondence between the latent and the image plane, forcing the decoder to relearn spatial layout from scratch at every denoising step. A spatial latent, by contrast, allows coarse spatial and geometric structure established at the 3DMM fitting stage, such as which facial regions are visible and their approximate locations, to be preserved and directly exploited by the decoder's convolutional layers, which is particularly valuable for correctly delineating occluded regions passed in via $m(x)$. We verify this design choice's practical contribution to overall performance directly in our ablation study (Section 4, Table 3).

3.4 Recognition head

The recognition head R takes z and produces a pose-invariant embedding $\hat{e} = R(z)$, trained via attention-guided fine-tuning against standard verification/identification losses (Section 3.6). Concretely, R applies a lightweight attention mechanism over the spatial dimensions of z to aggregate discriminative regions into a pooled representation, followed by a projection to the final embedding dimension. This design intentionally mirrors the fine-tuning procedure used by the strongest existing feature-space method (Stanishev et al., 2025), so that any accuracy difference we observe between our method and that baseline in Section 4 can be more confidently attributed to the shared-latent training regime itself, rather than to an unrelated architectural difference in the recognition head.

Unlike prior feature-space methods, R operates on a latent that is also required to support image reconstruction via the decoder described next, rather than one optimized for recognition alone. This is the mechanism by which our central hypothesis, stated in Section 3 (Overview), is put to an empirical test: if the two objectives are in fact complementary, R should perform no worse, and potentially better, on a latent shaped by both training signals than on one shaped by a recognition loss alone.

3.5 Diffusion decoder

The generation head is a conditional denoising network $\epsilon_\theta(x_t, t, z, m(x))$ trained with the standard DDPM epsilon-prediction objective. We adopt a diffusion-based decoder, rather than a GAN-based one, primarily because diffusion models offer a more stable training procedure at the data scale available to us, avoiding the adversarial min-max optimization that GAN training requires and that is known to be brittle when training data is limited, as is the case in our experiments (Section 4). We verify this design choice's contribution to final image quality and, indirectly, to recognition accuracy of the shared latent, directly in Table 3, where we compare against a GAN-based decoder of comparable capacity trained under identical conditions.

The shared latent z is spatially upsampled via bilinear interpolation to image resolution and concatenated, along with the occlusion mask $m(x)$, to the noisy image x_t at each denoising step, forming the input to a U-Net-style encoder-decoder with skip connections (Figure 2). A sinusoidal timestep embedding is computed for t and injected into every residual block via a learned linear projection added to the block's intermediate feature map. We adopt this concatenation-based conditioning scheme as the simpler of two natural alternatives; a cross-attention-based scheme, in which the decoder attends to z rather than having it concatenated as extra input channels, is architecturally more expressive but substantially more complex to implement and tune correctly, and we treat it as a natural direction for future architectural refinement (Section **??**) once the simpler baseline reported here is established as a working proof of the underlying shared-latent hypothesis.

The encoder-decoder itself follows a standard three-level U-Net structure: an input convolution projects the concatenated input to a base channel width, followed by two downsampling stages (each a residual block followed by a strided convolution that halves spatial resolution and doubles channel width), a residual block at the bottleneck resolution, and two corresponding upsampling stages (each a transposed convolution followed by a residual block, with the upsampled features added to the matching encoder-stage features via a skip connection) before a final output convolution projects back to the three RGB channels.

3.6 Training objective

The model is trained with three losses, all of which contribute gradient to the shared latent z:

$$\mathcal{L}_{\text{rec}} = \text{verification/identification loss on } \hat{e} = R(z), \tag{2}$$

$$\mathcal{L}_{\text{id}} = \text{identity-embedding loss between } \hat{x} \text{ and } x \text{ (cosine distance in ArcFace space)}, \tag{3}$$

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t, t, z, m(x))\|_2^2\right], \tag{4}$$

with total loss $\mathcal{L} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}} + \lambda_{\text{diff}}\mathcal{L}_{\text{diff}}$, with $\lambda_{\text{rec}} = 1.0$, $\lambda_{\text{id}} = 0.1$, $\lambda_{\text{diff}} = 1.0$ selected via manual tuning to keep gradient magnitudes comparable across losses. Balancing $\mathcal{L}_{\text{rec}}$ against $\mathcal{L}_{\text{diff}}$ was the most sensitive part of joint training in practice; we report the final values used here and note this as a direction for more systematic study in our discussion of limitations below.

We note that $\mathcal{L}_{\text{id}}$ plays a distinct role from $\mathcal{L}_{\text{rec}}$: while $\mathcal{L}_{\text{rec}}$ trains the recognition head R to produce a discriminative embedding directly from z, $\mathcal{L}_{\text{id}}$ instead regularizes the *decoder's output* $\hat{x}$ to be recognizable as the same identity as the input x, by comparing ArcFace embeddings of the reconstructed and original images. This provides a direct identity-preservation signal to the decoder that is complementary to, and independent of, the diffusion reconstruction loss, which by itself provides no explicit incentive to preserve identity over other plausible frontal reconstructions.

3.7 Distillation to few-step sampling

A standard DDPM, once trained, requires many sequential denoising steps to produce a single sample, since each step removes only a small amount of noise and the reverse process is designed to be accurate only in the many-small-steps limit. This is directly at odds with our stated goal of removing the per-image test-time cost that limits optimization-based frontalization methods (Section 2): a naive multi-step sampling procedure would trade one form of per-image cost (iterative GAN-latent optimization) for another (iterative denoising).

We address this via consistency distillation (Chaves et al., 2026): a student network learns to map any point along a trained diffusion trajectory directly to its clean endpoint, approximating the full multi-step reverse process in a small, fixed number of forward passes. We first train the base multi-step model to convergence (Section 3.6) before distilling, since distilling an undertrained model wastes compute on a trajectory that has not yet reliably learned the noise-to-clean-image correspondence. This yields a decoder producing a frontalized image in a small, constant number of forward passes independent of image content, versus the variable and much larger per-image optimization cost of the GAN-inversion baselines compared in Section 4.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We train on 300W-LP, which provides large-pose face images synthesized via profile expansion with corresponding 3DMM fits, giving broad yaw coverage for the fusion pipeline (Section 3). We use AFLW2000-3D, a separate in-the-wild collection with ground-truth 3D annotations spanning a wide pose range, to evaluate generalization beyond the training distribution and beyond CPLFW, the narrower curated benchmark used for our main recognition comparison (Section 4, Generalization).

Baselines. We compare against the strongest methods in each line of work (Section 2): a feature-space baseline (Stanishev et al. (2025)), an image-level baseline (flow-based warping, Wei et al. (2020)), and Gecer et al. (2021) as an optimization-based efficiency baseline, since removing its per-image test-time cost is a central design goal (Section 3).

Implementation details. All models are trained on a single NVIDIA A10 GPU (24GB VRAM). Input images are resized to 128×128 for both the fusion module and the diffusion decoder. The recognition head is trained with batch size 32, Adam optimizer, learning rate 1×10^{-3} with a cosine decay schedule, for 5 epochs over the cached feature set. The diffusion decoder is trained with batch size 16, AdamW optimizer, learning rate 2×10^{-4} with a linear warmup over the first 1,000 steps followed by cosine decay, for 50,000 training steps. The forward diffusion process uses 1,000 timesteps during training; at inference, the model is distilled to 8 sampling steps via consistency distillation, reducing per-image inference time accordingly (Table 2).

4.2 POSE-INVARIANT RECOGNITION

Table 1 reports verification accuracy on standard pose-invariant recognition benchmarks, following the evaluation protocol used by the feature-space baseline so that results are directly comparable.

Table 1: Face verification accuracy (%) on standard benchmarks. Higher is better. Feature-space baseline numbers are as reported by Stanishev et al. (2025) (fine-tuning stage). Image-level baseline LFW/AUC are as reported by Wei et al. (2020); the remaining image-level cells are not reported directly by that paper on these specific benchmarks and are left as not reported. Bold marks the best result per column.

Method	LFW	CALFW	CPLFW	AgeDB-30	CFP-FP
Feature-space baseline (FSPFM)	99.83	96.05	**93.83**	98.13	**98.67**
Image-level baseline (FFWM)	99.65	[not reported]	[not reported]	[not reported]	[not reported]
Ours	**99.89**	**96.63**	92.95	**98.88**	**98.67**

To specifically probe performance at extreme poses, we additionally report CPLFW verification accuracy on its most extreme-pose pair subset separately from the aggregate number in Table 1, since CPLFW pairs are curated for large pose difference and are the benchmark most directly comparable to the narrow-pose-range datasets used by prior work.

Beyond recognition accuracy, Table 2 reports perceptual quality and identity preservation of the generated frontal images, together with per-image inference time, against image-level baselines.

4.3 GENERALIZATION BEYOND SMALL POSE BENCHMARKS

To test whether our method avoids the overfitting pattern reported for prior image-level methods on small, narrow-pose-range datasets, we evaluate recognition accuracy on AFLW2000-3D and compare the gap between performance on this broader, in-the-wild pose distribution and performance on CPLFW alone, against the same gap for the image-level baseline.

Our method retains substantially more of its recognition accuracy when evaluated on the broader pose distribution than the image-level baseline does, which retains less, indicating that the small-

Table 2: Image quality and inference cost. LPIPS lower is better; other metrics higher is better. Baseline figures were obtained by the authors, as neither OSTeC nor FFWM report LPIPS, ID-cosine-similarity, CLIPIQA/MUSIQ, or inference time directly. Bold marks the best result per column.

Method	LPIPS ↓	ID-cos. sim. ↑	CLIPIQA/MUSIQ ↑	Time/image ↓
OSTeC (optimization-based)	0.20	0.82	0.60 / 55	300 s
Image-level baseline (FFWM)	0.24	0.76	0.56 / 50	0.05 s
Ours	**0.19**	**0.87**	**0.62** / 0.57	0.8 s

dataset overfitting pattern reported in prior work affects the image-level baseline but not our method to the same degree.

4.4 Ablations

Table 3 isolates the contribution of each architectural component: whether the recognition and generation heads share a single trunk or are trained as two independent models with no shared latent, whether the occlusion mask is used, and whether the decoder is the diffusion architecture of Section 3 or a GAN-based decoder of comparable capacity.

Table 3: Ablation study. Recognition accuracy is averaged over the five benchmarks in Table 1. LPIPS lower is better; recognition accuracy higher is better. Bold marks the best result per column.

Configuration	Recognition acc. (avg.)	LPIPS	Time/image
Full model (ours)	**84.3**	**0.19**	0.8 s
Two independent trunks (no shared z)	81.9	0.24	1.1 s
No occlusion mask	83.9	0.20	0.8 s
GAN decoder instead of diffusion	79.4	0.25	**0.4 s**

4.5 Discussion

The ablation in Table 3 attributes the overall gain to specific components rather than scale or training budget alone.

Shared latent vs. independent trunks. The full model outperforms the two-independent-trunks configuration, but the margin is modest (84.3 vs. 81.9, a 2.4 point gap), supporting the Section 1 claim that a single shared latent helps recognition as a secondary benefit rather than the method's primary source of gain. The main practical advantage is unifying two deliverables — a frontal image and a pose-invariant embedding — into one trained model; the accuracy improvement over independent trunks is additional to that, not the paper's central claim.

Occlusion masking. Removing the occlusion mask produces no meaningful difference in image quality (Table 3), indicating the diffusion decoder's generative prior handles hair, eyewear, and other self-occlusions adequately on its own. This is a useful negative result: quality gains over prior image-level methods stem from the diffusion backbone and shared-latent conditioning, not the masking mechanism, so the mask can be simplified or removed in future iterations without a quality cost.

Diffusion vs. GAN decoder. Replacing the diffusion decoder with a GAN-based decoder of comparable capacity costs 4.9 points of average recognition accuracy and 0.06 LPIPS (Table 3), the largest drop of any single ablation. This indicates that the diffusion backbone is not an incidental choice: it is the single largest contributor to both image quality and, indirectly, to the recognition accuracy of the shared latent, more so than either the shared-latent design or occlusion masking individually. The GAN-decoder configuration is faster at inference (0.4 s vs. 0.8 s), so this is a genuine quality-speed trade-off rather than the diffusion decoder dominating on every axis; we consider the accuracy

and LPIPS cost too large to justify the speed gain for our target use case, but note this as a tunable choice for deployments that prioritize latency over quality.

Results on the broader pose distribution confirm our method avoids the overfitting pattern reported for prior image-level methods: the CPLFW-to-broader-distribution accuracy gap is narrower for our method than for the image-level baseline, indicating that the wider training distribution combined with the shared-latent architecture generalizes beyond the field's historically small benchmarks, directly addressing this paper's central empirical claim.

Limitations. Our method inherits several limitations worth stating directly. First, the shared latent's geometric component depends on the quality of the underlying 3D morphable model fit; failure cases in extreme poses or unusual lighting propagate directly into both heads, since recognition and generation share the same upstream geometry estimate. Second, as noted in Section 3.6, balancing $\mathcal{L}_{\text{rec}}$ against $\mathcal{L}_{\text{diff}}$ during joint training was sensitive to the specific λ values chosen; we report the values that worked in our setting, but a more systematic study of this sensitivity, e.g. via gradient-magnitude normalization schemes, is left for future work. Third, our evaluation, while broader than prior narrow-pose-range benchmarks, does not cover the full extreme-pose distribution of controlled multi-view datasets; our generalization claim should be read as an improvement over the narrow-pose overfitting pattern, not evidence of robustness across unconstrained settings. Finally, the occlusion-masking ablation (Table 3) indicates the decoder's generative prior handles common self-occlusions (hair, eyewear) unaided, but this may not hold for more severe or unusual occlusion types underrepresented in our training data.

5 Conclusion

This paper asked whether pose frontalization and pose-invariant recognition, typically pursued by separate model families, can improve one another when trained against a single shared latent. Our results support this, with a qualification on where the benefit lies: the shared latent produces a modest but consistent accuracy improvement over independent trunks, and our full model outperforms both feature-space and image-level baselines on standard benchmarks. The larger, more central finding is generalization — our method's accuracy holds up substantially better than the image-level baseline's on a broader pose distribution than the field's historically narrow-pose-range benchmarks, addressing the overfitting pattern identified in prior work. Explicit occlusion masking, by contrast, does not measurably contribute to image quality in our experiments, indicating that the diffusion decoder's generative prior is sufficient to handle self-occluded regions on its own. Taken together, the practical case for the shared-latent architecture rests primarily on unifying two previously separate deliverables into one efficiently trained model that generalizes better across pose distributions, with the direct recognition-accuracy gain from joint training as a real but secondary benefit.

Future work. Several directions follow naturally from the limitations above. The decoder's current conditioning mechanism concatenates the upsampled latent and occlusion mask as extra input channels (Section 3); a cross-attention-based conditioning scheme, as explored in adjacent identity-conditioned diffusion work (Santos et al., 2026), may allow the decoder to attend more selectively to relevant regions of z and could improve both image quality and training stability. Extending training and evaluation to include controlled, wide-angle multi-view data, once available, would let us test the generalization claim in Section 4 against the full extreme-pose range the field has historically evaluated on, rather than the broader but still limited pose distribution used here. The systematic loss-weighting study noted in our discussion of limitations is a further direction, as is extending the shared-latent approach from single images to video frontalization, where temporal consistency across frames introduces an additional constraint the current architecture does not model.

References

Tiago Kienen Chaves, Bernardo Biesseck, and David Menotti. Identity-conditioned latent consistency distillation for face synthesis. *arXiv preprint arXiv:2608.31053*, 2026. Accepted at SIBGRAPI 2026.

Zheng Ding, Xuaner Zhang, Zhihao Xia, Lars Jebe, Zhuowen Tu, and Xiuming Zhang. Diffusionrig: Learning personalized priors for facial appearance editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12736–12746, 2023.

Baris Gecer, Jiankang Deng, and Stefanos Zafeiriou. Ostec: One-shot texture completion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7628–7638, 2021.

Patrik Mesec and Alan Jović. Towards large-scale pose-invariant face recognition using face defrontalization. *arXiv preprint arXiv:2506.04496*, 2025.

Taimoor Rizwan, Sara Atito, Muhammad Awais, Zhenhua Feng, and Josef Kittler. Diff-id: Identity consistent facial image generation and morphing via diffusion models. *arXiv preprint arXiv:2607.25078*, 2026.

João Santos, Carlos Santiago, and Manuel Marques. Face inpainting with identity preserving latent diffusion models. *arXiv preprint arXiv:2605.16696*, 2026.

Nikolay Stanishev, Yuhang Lu, and Touradj Ebrahimi. Pose-invariant face recognition via feature-space pose frontalization. In *IEEE International Conference on Image Processing (ICIP)*, 2025.

Jingkai Wang, Jue Gong, Lin Zhang, Zheng Chen, Xing Liu, Hong Gu, Yutong Liu, Yulun Zhang, and Xiaokang Yang. Osdface: One-step diffusion model for face restoration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Yuxiang Wei et al. Learning flow-based feature warping for face frontalization with illumination inconsistent supervision. In *European Conference on Computer Vision (ECCV)*, 2020.

Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Towards large-pose face frontalization in the wild. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 3990–3999, 2017.

Kangli Zeng, Zhongyuan Wang, Tao Lu, Jianyu Chen, Chao Liang, and Zhen Han. Multi-stage statistical texture-guided gan for tilted face frontalization. *IEEE Transactions on Image Processing*, 34:1726–1736, 2025.