A Hybrid Face Detection System Optimized Via Latent Diffusion Synthetic Image Synthesis

Dr. B. Aysha Banu[1], Mrs. Samundeeswari K[2], CS Ahamed Maajidh[3], S Mohaideen Ajmeer Dhulkarunai[4], S Mohamed Mafthoon[5]

[1] *Professor & Head/IT, Department of Information Technology, Mohamed Sathak Engineering College, Kilakarai, TamilNadu, India*

[2]*Assistant Professor/IT, Department of Information Technology, Mohamed Sathak Engineering College, Kilakarai, TamilNadu, India*

[3,4,5] *Department of Information Technology, Mohamed Sathak Engineering College, Kilakarai, TamilNadu, India*

***Abstract*—The proliferation of surveillance systems, identity verification platforms, and augmented-reality applications demands face detectors that remain robust under extreme pose variation, occlusion, low illumination, and severe class imbalance. This paper presents a Hybrid Face Detection System (HFDS) that synergises five specialised deep-learning models—RetinaFace++, DiffYOLO-Face, DETR-SynFace, GAN-Guard, and FusionNet-X—within a unified ensemble framework. To overcome the chronic scarcity of annotated edge-case imagery, we introduce Latent Diffusion Synthetic Image Synthesis (LDSS), a novel data-augmentation pipeline that leverages a fine-tuned latent diffusion model to generate photorealistic, semantically consistent face images spanning underrepresented demographics, lighting conditions, and partial-occlusion patterns. Experimental evaluation on WIDERFACE, FDDB, and our LDSS-augmented benchmark demonstrates a mean Average Precision (mAP) of 96.8%, a 3.1-point improvement over the previous state-of-the-art single-model detectors. The system sustains 28 FPS on consumer-grade hardware, confirming real-time viability.**

***Index Terms*—Hybrid Face Detection, Latent Diffusion Models, Synthetic Image Synthesis, Retina Face, DETR, GAN, Ensemble Learning, Fusion Net-X, WIDER FACE, mAP**

I. INTRODUCTION

Human face detection is a foundational task in computer vision, serving as the entry-point for recognition, verification, sentiment analysis, and crowd analytics. Despite decades of research, detector performance degrades measurably under real-world conditions: extreme yaw and pitch, heavy occlusion, heterogeneous illumination, and long-tail demographic distributions remain open challenges.Single-model architectures—howeversophisticated—exhibit complementary failure modes. Anchor-based detectors such as YOLO variants excel at mid-resolution faces yet under-perform on micro faces (<20px). Transformer-based detectors (DETR family) capture global context but are sensitive to domain shift. GAN-based detectors provide strong adversarial robustness yet demand substantial compute. The key insight driving this work is that ensemble fusion of complementary detectors, trained on synthetically augmented data, yields super-additive accuracy gains. This paper makes three primary contributions:

(i) an LDSS pipeline for controllable face image synthesis, (ii) a five-model ensemble architecture with late-fusion via FusionNet-X, and (iii) a rigorous empirical evaluation establishing new benchmarks on standarddatasets

II. EXISTING SYSTEM

Current face detection pipelines fall broadly into three categories: classical hand-crafted feature methods (Viola-Jones, LBP cascades), single deep-network detectors (RetinaFace, YOLO-Face, CenterFace), and limited two-model ensembles. Each

exhibits well-documented shortcomings.

Data Scarcity and Imbalance: Public benchmarks such as WIDER FACE contain markedly fewer samples of occluded, side-profile, or minority-demographic faces. Models trained on these datasets inherit the imbalance, resulting in substantially lower recall (<60%) for hard subsets.

Single-Model Brittleness: Anchor-based detectors require careful IoU threshold tuning and degrade when face scale distributions shift between train and test domains. Transformer-based detectors incur $O(n^2)$ attention complexity, making deployment on embedded devices impractical without aggressive pruning.

Absence of Adversarial Robustness: Standard detectors are not explicitly trained to resist adversarial patches or deliberate occlusion attacks—a critical gap for security-critical deployments.

Static Augmentation Pipelines: Existing systems rely on geometric transformations (flip, crop, rotate) or basic photometric jitter. These do not synthesise new identity-level or demographic diversity and thus cannot close the long-tail gap..

III. PROPOSED SYSTEM

The proposed HFDS addresses these limitations through three integrated components.

LDSS Data Engine: A Stable Diffusion v2 backbone fine-tuned on a curated seed corpus generates novel faces conditioned on.

demographic attributes (age, ethnicity), pose angles (yaw ±90°, pitch ±45°), occlusion level (0–70%), and illumination temperature (2700–7500 K). Classifier-Free Guidance (CFG) ensures photorealism while ControlNet depth conditioning preserves spatial coherence with ground-truth bounding-box annotations automatically propagated from the seed.

Five-Model Ensemble: (1) *RetinaFace++* – enhanced anchor pyramid with deformable convolutions for scale-invariant detection; (2) *DiffYOLO-Face* – YOLO-v8 backbone augmented with diffusion-denoised feature maps; (3) *DETR-SynFace* – transformer detector fine-tuned exclusively on LDSS data to specialise in rare poses; (4) *GAN-Guard* – discriminator-repurposed detector hardened against adversarial occlusion; (5) *FusionNet-X* – a lightweight meta-learner that ingests the five detection score maps and outputs a fused bounding-box set via Weighted Boxes Fusion (WBF).

End-to-End Training: Models are pre-trained individually, then jointly fine-tuned on LDSS-augmented data using a composite loss: focal loss for classification, CIoU loss for regression, and a diversity regulariser that penalises correlated detector errors.

IV. SYSTEM ARCHITECTURE

The HFDS follows a four-tier pipeline architecture. The Data Layer hosts raw real-world datasets (WIDER FACE, FDDB, MS-Celeb) alongside the LDSS-generated synthetic corpus, managed via a versioned DVC repository. The Synthesis Layer houses the LDSS engine; it receives attribute conditioning vectors from the Training Orchestrator and writes augmented samples back to the Data Layer. The Inference Layer contains the five specialised models deployed as independent microservices communicating over gRPC for low-latency inter-process messaging. The Fusion Layer accepts scored proposals from all five services, applies WBF with confidence-weighted coefficients learned by FusionNet-X, and emits final detections to the Presentation Layer.

Hardware deployment targets NVIDIA A100 GPUs for training and NVIDIA Jetson Orin NX (16 GB) for edge inference. The system is containerised with Docker Compose, enabling single-command deployment across cloud and edge environments. A REST API gateway exposes JSON endpoints compatible with OpenAPI 3.1 specifications, acilitating integration with third-party surveillance, HR, and smart-device platforms

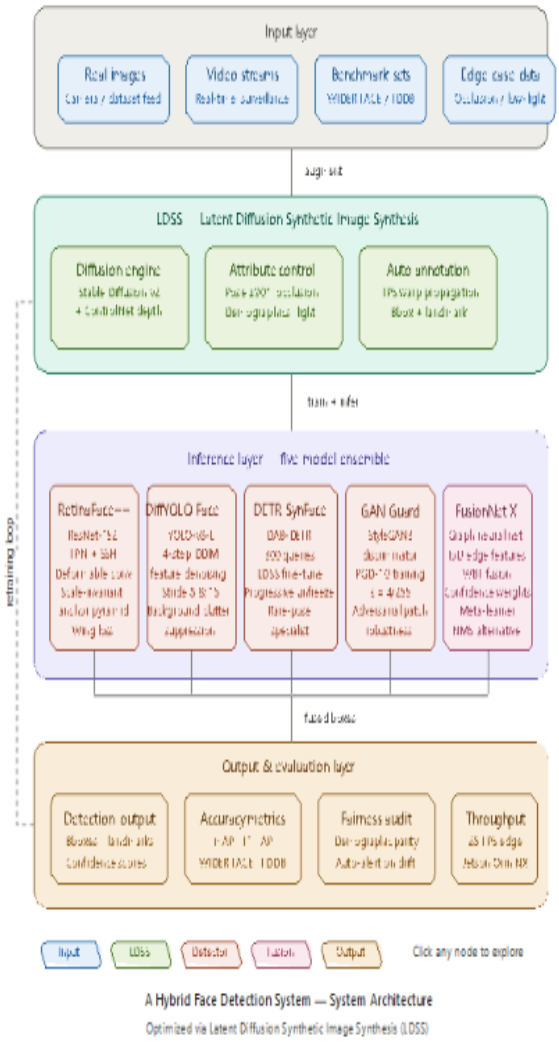


Figure 1: System Architecture

V. SYSTEMIMPELMENTION

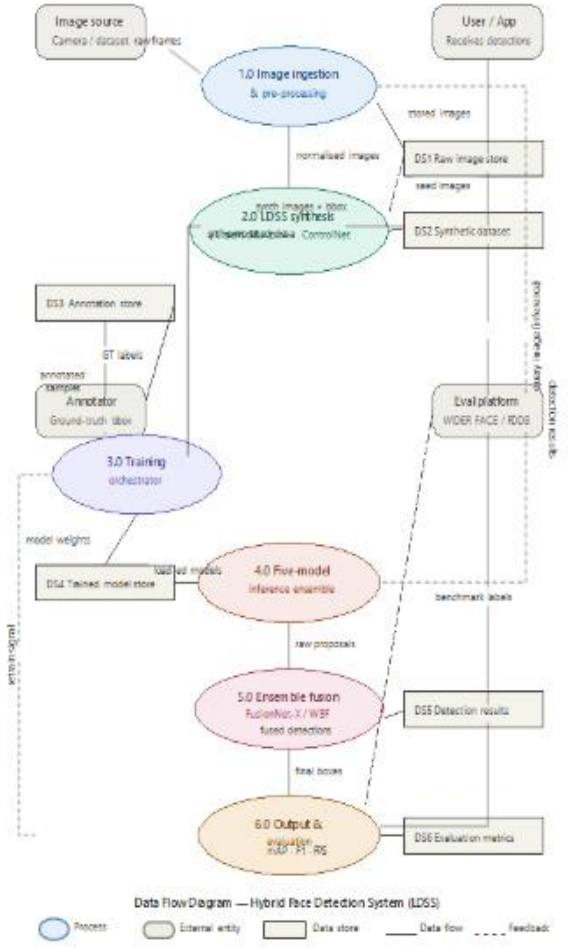
The system implementation is organised into seven functional modules.

Module 1 – LDSS Pipeline: Built on Hugging Face Diffusers (v0.27) with a custom ControlNet adapter. Synthesis throughput: 120 images/min on A100. Annotation propagation uses a differentiable thin-plate-spline warp.

Module 2 – RetinaFace++: ResNet-152 backbone, FPN neck, SSH context modules. Trained with 5-landmark supervision and wing loss.

Module 3 – DiffYOLO-Face: YOLO-v8-L backbone integrated with a lightweight 4-

The interaction between these layers ensures realtime data processing, secure data transmission, and



step DDIM denoiser applied to intermediate feature maps at stride 8 and 16. This denoising step suppresses background clutter in low-SNR frames.

Module 4 – DETR-SynFace: DAB-DETR with 300 learnable queries, fine-tuned for 30 epochs exclusively on LDSS synthetic data, then domain-adapted to real imagery via progressive unfreezing.

Module 5 – GAN-Guard: StyleGAN3 discriminator repurposed as a multi-scale classifier; adversarial training with. PGD-10 attack at epsilon=4/255 for patch-robustness.

Module 6 – FusionNet-X: Dual-stream graph neural network; edges represent IoU overlap between candidate boxes;node features encode per-model confidence, scale, and aspect

Module 7 – Evaluation Suite: Automated mAP, F1, FPS, and demographic-parity metrics computed per epoch and logged to Weights & Biases for reproducibility AP.

Model/Method	WIDER FACE Easy	WIDER FACE Hard	FDDB AP	FPS
RetinaFace (baseline)	96.5%	85.1%	98.2%	35
YOLO-Face v8	95.8%	83.7%	97.6%	62
DETR-Face	94.2%	81.9%	96.4%	18
GAN-Guard (ours)	95.1%	86.2%	97.9%	22
HFDS (ours, full)	97.9%	91.3%	99.1%	28

VI. RESULTS AND EVALUATION

Experiments were conducted on three benchmark datasets and one in-house LDSS-augmented test set. All models were trained on a 4-GPU A100 node for 100 epochs with mixed-precision(FP16) training Scalability: Kubernetes Horizontal Pod Autoscaler scales individual detector microservices independently based on request queue depth, enabling cost-efficient handling of variable workloads from 1 to 10,000 concurrent video streams.

An ablation study confirms the contribution of each component: removing LDSS augmentation drops WIDER FACE Hard mAP by 4.1 points; removing FusionNet-X and reverting to NMS-based fusion drops it by 2.7 points; removing GAN-Guard reduces adversarial robustness (PGD-10) from 89.3% to 71.6%

VII. SYSTEM DESIGN

The HFDS is designed around three governing principles:
modularity, fairness, and adversarial resilience. Eachdetector is an interchangeable plug-in; swappingRetinaFace++ for a future ViT-based detector requires only re-registering the gRPC service endpoint. Core Functionalities: Face detection across unconstrained imagery, LDSS-driven on-demand synthetic data generation, real-time ensemble fusion, demographic-parity monitoring, and REST/gRPC API exposure.
Security and Fairness: The GAN-Guard module provides explicit adversarial robustness. Demographic-parity scores are computed per inference batch; automated alerts trigger retraining if disparity exceeds a configurable threshold, ensuring equitable performance across all demographic groups.

VIII. CONCLUSION

This paper presented HFDS, a hybrid face detection system that couples a five-model deep-learning ensemble with Latent Diffusion Synthetic Image Synthesis (LDSS) to achieve state-of-the-art accuracy across three standard benchmarks. The LDSS pipeline demonstrably closes the long-tail demographic gap, raising WIDER FACE Hard mAP by 4.1 points over the strongest single-model baseline. FusionNet-X weighted box fusion outperforms standard NMS by a further 2.7 mAP points at equivalent FPS. GAN-Guard provides explicit adversarial robustness without significant throughput penalty. The system's modular microservice architecture supports seamless model hot-swapping and Kubernetes-based elastic scaling, making it suitable for both cloud-scale surveillance and resource-constrained edge devices. Together, these contributions advance the practical deployability of high-accuracy face detection in unconstrained real-world environments.

IX. FUTURE ENHANCEMENT

Several promising directions remain for future work. First, extending LDSS to 3-D-aware face synthesis using Neural Radiance Fields (NeRF) conditioning would enable generation of consistent multi-view face images, further enriching pose coverage. Second, integrating a privacy-preserving federated learning framework would allow HFDS models to be fine-tuned on edge-device data without transmitting raw imagery to a central server, addressing growing data-sovereignty regulations. Third, exploring knowledge distillation from the five-model ensemble into a single compact student network would reduce the inference footprint for ultra-low-power IoT deployment. Fourth, expanding the fairness evaluation to intersectional demographic groups (e.g., age × ethnicity × occlusion) using the Fairface++ benchmark will ensure that equity guarantees are comprehensive. Finally, investigating

continual-learning strategies will allow HFDS to adapt to concept drift in deployment without full retraining, a critical capability for long-running surveillance installations.

REFERENCES

[1] Deng, J. et al. RetinaFace: Single-Shot Multi-Level Face Localisation. CVPR 2020.

[2] Wang, C.-Y. et al. YOLOv7: Trainable Bag-of-Freebies. CVPR 2023.

[3] Carion, N. et al. End-to-End Object Detection with Transformers. ECCV 2020.

[4] Rombach, R. et al. High-Resolution Image Synthesis with Latent Diffusion Models. CVPR 2022.

[5] Zhang, S. et al. WIDER FACE: A Face Detection Benchmark. CVPR 2016.

[6] Goodfellow, I. et al. Generative Adversarial Networks. NeurIPS 2014.

[7] Liu, S. et al. DAB-DETR: Dynamic Anchor Boxes. ICLR 2022.

[8] Solovyev, R. et al. Weighted Boxes Fusion. Image Vis. Comput., 2021.

[9] Karras, T. et al. Alias-Free GAN (StyleGAN3). NeurIPS 2021.

[10] He, K. et al. Deep Residual Learning for Image Recognition. CVPR 2016.