Chapter 4: Vision Transformers for Face Biometrics

Face recognition has evolved from classic statistical methods (PCA, LDA) and handcrafted descriptors (e.g., LBP) to deep learning pipelines. Eigenfaces (PCA) and Fisherfaces (LDA) treat the whole face image linearly: PCA finds global principal components of face images, while LDA finds directions that best separate identities. These linear methods have limited capacity and struggle with variations (pose, expression). Local Binary Patterns (LBP) and related texture features added some robustness by capturing local texture differences, but they, too, are shallow and hand-designed.

In contrast, CNNs learn hierarchical features from data. Modern face CNNs (e.g., using ResNet backbones) achieve >99% accuracy on benchmarks, thanks to millions of parameters and discriminative loss functions (ArcFace, etc.). However, CNNs have fixed local receptive fields and rely on many layers for global context. They can still fail under large occlusions or have redundant training for invariance. Interpretability is also limited: it's hard to see why a CNN recognizes a face in terms of specific facial features.

Vision Transformers (ViTs) offer a different paradigm: an image is split into patches (e.g., 16×16 pixels), each patch is linearly embedded and endowed with positional information, and then multi-head self-attention layers model relationships among all patches [r]. In face biometrics, this has several advantages:

- Global Context: A ViT's attention heads can directly relate distant patches (e.g., eyes and mouth regions) without many convolution layers. This provides a *global receptive field* from the first layer. In practice, ViTs have been shown to be robust to occlusion (if an attacker covers part of the face, other parts can still inform the identity) and to modeling variations in pose or illumination [r].

- Interpretability: Because each attention head produces an attention map over image patches, we can visualize which regions contribute most to identity classification. For example, attention might highlight the nose tip, eyes, or facial hair, revealing the model's "focus". This interpretability can help diagnose failure cases (e.g., if the model attends to background edges instead of the face) and analyze demographic effects (by comparing attention maps across different groups).

- Cancelable Templates: Embeddings from ViTs can serve as cancelable biometric templates. Unlike raw facial images, a transformer embedding can be hashed or transformed (e.g., via random projections) to produce a non-invertible template. Recent work on cancelable face templates shows that applying random projections to deep face features achieves unlinkability and revocability while retaining accuracy [r]. With ViTs, the

same principle applies: one could randomly project the ViT's final face embedding to store a secure template. The slides highlight that ViT embeddings naturally enable template protection, since they live in a vector space where invertibility is hard without knowing the projections.

Comparing ViT-based face recognition with older methods:

- Vs. PCA/LDA: ViTs are non-linear and data-driven. They do not require extracting eigenvectors or class-specific discriminants manually; instead, they learn complex feature hierarchies. As a result, ViTs have achieved far higher accuracy on modern face benchmarks than any linear method. PCA/LDA methods might still be used for quick indexing or as a baseline, but they lack the representational power of attention networks.

- Vs. LBP (Texture): LBP is a simple local descriptor that is invariant to monotonic lighting changes. While computationally cheap, LBP-based face recognition is orders of magnitude less accurate than CNNs/ViTs on large datasets, due to its limited modeling. ViTs, on the other hand, implicitly learn complex texture features (eyes, skin patterns, edges) but can also combine them with holistic cues (face shape).

- Vs. CNNs: Transformers have shown comparable or slightly higher accuracy than CNNs on large-scale face tasks, given enough data. For example, the LVFace model (a ViT with >100M parameters) topped a major face recognition challenge, outperforming all CNN-based competitors [r]. One key difference is that CNNs embed strong inductive bias (locality, translation equivariance), which is helpful with limited data, whereas ViTs rely on large-scale training to learn similar biases. In practice, a pre-trained ViT fine-tuned on millions of face images (as in LVFace) yields SOTA performance. Transformer features also tend to be more uniform and robust: ViTs are found to have higher "shape bias" (focusing on overall structure) and more consistent outputs under adversarial perturbations than CNNs. However, CNNs are still faster to train from scratch on small datasets and benefit from decades of engineering (feature fusion, skip connections, etc.).

Attention Visualization: A unique advantage of ViTs is that we can visualize the attention maps. For a given face image, each self-attention head produces a heatmap over the 2D patch grid (and each transformer layer has multiple heads). Studies have shown that on real vs spoofed faces, these attention patterns differ: a model trained for deepfake detection will attend to different micro-texture regions than one trained on live. By inspecting these maps, developers can see which facial regions the model deems discriminative. For instance, a ViT might always attend to the eye corners and mouth in one head, and to hairline in another. If attention shifts unexpectedly (e.g. focusing on jewelry or background), it may indicate a domain shift or bias.

Demographic Fairness: A known challenge in face biometrics is bias: many large face datasets (e.g. WebFace, MS-Celeb) contain mostly young white males, so networks often perform worse

on underrepresented groups (women, elderly, certain ethnicities) [r]. Vision transformers do not automatically fix this bias, but they offer tools to study it. Because attention is explicit, one can compare whether a ViT trained on biased data attends differently to different groups. Some recent work suggests incorporating fairness constraints into attention models (e.g., by balancing training data or adding demographic-aware loss), though this area is still emerging. Meanwhile, as noted earlier, one can mitigate bias post-hoc by demographic-aware score normalization. In summary, ViTs promise greater transparency in the analysis of bias, but face recognition fairness must still be addressed by dataset curation and/or algorithmic techniques.

Spoof and Deepfake Detection: ViTs have also been applied to face anti-spoofing (detecting printed or replay attacks) and deepfake detection. The idea is that real faces and spoofed images differ in subtle texture patterns (e.g., moiré patterns, printing artifacts) or temporal dynamics. Transformers, with their fine-grained patch attention, can pick up these cues. Indeed, recent works report strong anti-spoofing performance using ViT backbones. For example, Vision Transformers fine-tuned on face liveness datasets have achieved state-of-the-art generalization across attack types by capturing global cues (like inconsistent lighting) that CNNs might miss. Similarly, *Deepfake detection* models have begun to incorporate self-attention (e.g., FakeFormer, ConViT) to highlight inconsistencies in eyes or skin that GANs often produce. The Clarkson lab slides mention that “attention patterns differ: real vs spoof”, and there is evidence that multi-head attention can reveal where a face synthesis fails (e.g., unusual patches around the mouth). Thus, a practical face biometric system might use a ViT not only for identity but also to flag anomalies via its attention distribution.

Summary: Vision Transformers bring a powerful, end-to-end learned model to face biometrics. Compared to traditional PCA/LDA/LBP, they offer dramatically higher accuracy and robustness at the cost of needing more data and computing. Compared to CNNs, ViTs match or exceed accuracy on large-scale benchmarks, while offering built-in interpretability through attention. They enable modern features like cancelable templates and better spoof resilience. As the literature shows (Face Transformer, LVFace, etc.), with proper training, they can set new performance benchmarks. However, they should be used thoughtfully: practitioners must still pay attention to data biases, security (template protection), and evaluation on real-world criteria (e.g., ISO-compliant FAR/FRR levels). In the end, ViTs complement rather than entirely replace earlier methods—hybrid approaches (CNN+ViT, graph networks, etc.) may further advance face biometrics.