Chapter 1: Basics of Transformer Architectures for Biometric Analysis

Evolution from CNNs and RNNs to Transformers

Early Deep Learning (DL) models for vision and sequence data were dominated by Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), respectively. CNNs excelled at learning spatial hierarchies in images via localized receptive fields and shared filters, becoming the backbone of computer vision tasks like image classification and face recognition. RNNs (and their gated variants like LSTMs) were the go-to for sequential data (e.g., speech or text), processing inputs one timestep at a time to capture temporal dependencies. However, both approaches had limitations. RNNs struggled with long-range dependencies due to vanishing gradients and could not be easily parallelized, as they processed sequences sequentially. CNNs, while parallelizable, had fixed local receptive fields – capturing global context required deep stacks of layers, and even then, long-range interactions were indirect. Transformer architectures emerged as a solution to these issues: introduced by Vaswani et al. (2017) in *"Attention Is All You Need" [1]*, the Transformer did away with recurrence and long convolution pipelines entirely, relying solely on attention mechanisms for sequence processing [1]. This innovation enabled significantly more parallelism (processing sequences in one shot rather than step-by-step) and better capture of long-range relationships. Indeed, the original Transformer showed superior accuracy on machine translation compared to RNN-based models while training faster due to its parallelizable design. The success in NLP soon carried over to vision tasks. By 2020, researchers demonstrated that a pure Transformer model could operate on images (by treating image patches as sequence tokens) and outperform CNNs on large-scale image recognition benchmarks [2]. This was a pivotal moment: it indicated that the inductive biases of CNNs (locality and translational weight sharing) were not the only path to understanding images – given enough data and computation, self-attention could learn global and local image features effectively. In summary, the field evolved from CNNs and RNNs to Transformers to overcome the bottlenecks of locality and sequential processing. Transformers brought a paradigm shift by modeling pairwise interactions between all inputs through attention, allowing models to dynamically learn what to focus on irrespective of distance in space or time.

Key Innovations in Transformer Architecture

The Transformer architecture introduced several key innovations that enabled its success, especially relevant to biometric data analysis. We highlight three fundamental components: self-attention mechanisms, positional encoding, and the encoder–decoder architecture (particularly in sequence transduction contexts).

Self-Attention Mechanism

At the heart of every Transformer block is the self-attention mechanism, which allows the model to weigh the importance of different parts of the input relative to each other. Instead of a fixed geometric neighborhood (as in CNN filters), each element (token) in the input can potentially attend to any other element. This is implemented by computing pairwise attention weights using query, key, and value vectors derived from the inputs. For a set of input tokens (represented as vectors), the Transformer computes attention as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q, K, V are matrices collecting the *query*, *key*, and *value* vectors for all tokens, and d_k is the dimensionality of queries/keys. The softmax term produces an attention matrix whose entries indicate how strongly each token (row) attends to another token (column). This allows the model to dynamically focus on salient features: for example, in a face image, a token encoding an eye region could attend to another token encoding the other eye or surrounding skin texture if those relationships are important for recognition. Multi-head self-attention extends this idea by having multiple attention heads in parallel, each with its own learned projection (different Q,K,V matrices). Each head can learn to attend to different patterns or features (one head might focus on global structure while another on fine details). The outputs of all heads are then combined. This multi-head approach greatly increases the model's expressiveness, as different heads can capture complementary aspects of the biometric pattern. In biometric tasks, this means one attention head could learn to focus on minutiae-rich regions (like ridge endings in a fingerprint or distinguishing marks on a face), while another head captures broader context (like the overall ridge flow or face shape). The self-attention mechanism thus provides dynamic feature routing: depending on the input content, information from one part of the biometric can be routed to influence another part adaptively. This is in contrast to CNNs, which always aggregate information from a fixed local neighborhood – the Transformer can draw long-range connections (e.g., relating two distant minutiae points in a fingerprint) in a single layer. As a result, self-attention is especially powerful for biometric traits that have complex global patterns (like the network of blood vessels in a retinal scan or the full topology of fingerprint ridges) because it can integrate evidence across the entire input.

Positional Encodings

One challenge with using self-attention on sequences (or image patch sequences) is that the mechanism itself is invariant to position – if we permute the tokens, the set of QK^T dot-products remains the same. Traditional sequences (text, speech) and images have meaningful ordering and spatial structure. Transformers handle this by adding a positional encoding to each token's representation. Positional encodings can be learned vectors or fixed functions (the original Transformer used a sinusoidal positional encoding). The sinusoidal encoding for position pos and feature index i is defined as:

$$PE_{(pos,2i)} = \sin\left(pos/10000^{2i/d_{\text{model}}}\right),$$
$$PE_{(pos,2i+1)} = \cos\left(pos/10000^{2i/d_{\text{model}}}\right),$$

ensuring each position has a unique, smooth embedding that the model can use to infer relative positions. In images, a 2D positional encoding (often separable into x and y components) is used when flattening patches. Positional information is crucial for biometrics: consider a fingerprint – the spatial arrangement of ridges and minutiae defines identity. If a model were oblivious to positions, it might mix up features from disparate regions. With positional encodings, a Transformer knows, for instance, that Token #5 corresponds to, say, the top-left region of an iris image, whereas Token #50 might be the bottom-right. This allows the self-attention to respect and utilize spatial layout (e.g., attending to adjacent ridge segments in order).

Encoder–Decoder Architecture

The original Transformer introduced an encoder–decoder architecture for sequence-to-sequence tasks like translation. The encoder is a stack of self-attention layers that transforms an input sequence (e.g. a biometric data sequence or image patches) into a set of contextualized embeddings. The decoder is another stack that takes the encoder's output and generates an output sequence (often using *cross-attention* to attend to encoder embeddings, and self-attention for the output sequence being generated). In many vision applications (like image classification or biometric feature extraction), a full encoder–decoder may not be needed – often just an encoder suffices to produce a rich representation of the input. For example, Vision Transformers for image recognition use an encoder that outputs a representation (with a special *[CLS]* token for classification). However, the encoder–decoder idea becomes relevant in certain biometric scenarios. For instance, consider cross-modal biometric matching (such as matching an infrared vein pattern to a visible-light image of the hand) – one could use an encoder–decoder where the encoder processes one modality and the decoder (with cross-attention) tries to reconstruct or map to the other modality. Indeed, Transformers have been applied in multi-modal biometric verification where cross-attention helps relate information from, say, a thermal face image to a visible face image of the same person. In summary, the encoder–decoder structure provides flexibility for tasks that involve transducing one representation to another, which can be useful for biometric data integration (e.g., fusing face and voice biometrics, or enhancing a partial fingerprint by "translating" it into a complete one using learned patterns).

How Self-Attention Benefits Biometric Trait Analysis

Biometric traits – such as fingerprints, iris patterns, facial features, or vein patterns – often contain complex structures and long-range dependencies. The self-attention mechanism directly addresses some of the unique challenges in biometric analysis:

- Long-Range Dependency Modeling: Biometric patterns are not entirely local. For example, the relative configuration of facial features (eyes, nose, mouth) or the pattern formed by distant minutiae points in a fingerprint can be crucial for identity. Self-attention naturally captures such long-range relationships in one layer by allowing any token to attend to any other [3]. This means a Transformer analyzing a palm vein image can correlate distant vein branches, or a Transformer analyzing a face can relate information from a scar on the forehead to the shape of the jaw, if that's pertinent to identification.

- Dynamic Focus on Relevant Features: Different biometric samples, even of the same type, may require focusing on different features. For instance, one fingerprint might have very clear minutiae in one region while another has clearer ridge texture in another region. Transformers can dynamically shift their attention to whatever features are most discriminatory in a given sample. In contrast, a CNN with fixed filters might always emphasize certain frequencies or shapes, potentially overlooking idiosyncratic features that matter in a particular case. Self-attention provides a content-dependent weighting of features – e.g., if a certain region of a fingerprint is smudged or occluded, the model can down-weight that region and concentrate on another part of the print that is clearer.

- Multi-Modal and Cross-Spectral Integration: Biometric systems increasingly fuse information from multiple sources (e.g., face + voice, or combining infrared and visible images). Transformers' attention mechanism can operate across mixed token sequences or through cross-attention in a multi-stream architecture, enabling integration of heterogeneous biometric features. The ability to capture inter-dependencies between different modalities has been demonstrated in cross-spectral transformers for biometrics [4]. For example, a cross-spectral biometric Transformer can learn the relationships between patterns in an infrared vein image and a visible-light periocular (eye region) image, effectively learning a joint representation that links the two. This is important because different biometric modalities may compensate for each other's weaknesses (vein patterns are robust to spoofing and occlusions like masks, while periocular images carry complementary identity info).

- Robustness to Partially Missing Data: Biometric data can be incomplete – a partial fingerprint, a face with occlusion (e.g., mask or sunglasses), etc. Self-attention offers a graceful way to handle missing or corrupt regions: tokens corresponding to missing parts simply won't attract much attention (they carry little useful information), and the model can still connect the remaining informative parts. There is no strict requirement for a contiguous receptive field to cover the gap, unlike a CNN that might propagate corrupted information through convolution. In essence, attention can skip over missing data by reweighting focus to available data. Later in Chapter 2, we will see how this property is exploited for partial fingerprint recognition.

In summary, self-attention confers Transformers the ability to globally reason over biometric inputs, focusing on salient patterns and ignoring irrelevant noise or missing parts. This aligns

well with the needs of biometric analysis, where the spatial arrangement and global coherence of features often determine identity.

Vision Transformers (ViT) and Their Application to Biometrics

The extension of Transformers to image analysis gave rise to the Vision Transformer (ViT), which has significantly influenced biometric recognition research. ViT was first demonstrated by Dosovitskiy et al. (2020) as a *pure transformer* applied to image patches [2]. The model splits an image into fixed-size patches (e.g., 16×16 pixels), flattens each patch into a vector, linearly projects it to an embedding, and feeds the sequence of embeddings into Transformer encoder layers (with an added classification token). This design proved that, given sufficient training data, a ViT can outperform CNNs on image classification benchmarks like ImageNet [5]. Notably, ViTs achieved state-of-the-art results when pre-trained on large datasets (e.g., ImageNet-21k) and then fine-tuned to a specific task. This was a critical insight for biometrics because biometric datasets are often relatively small or specialized. By leveraging a ViT pre-trained on generic imagery and fine-tuning it, researchers found they could attain excellent performance on various biometric tasks despite limited biometric data.

In the context of biometric applications, Vision Transformers have been successfully applied across several modalities:

- Facial Recognition: Transformers have been explored to supplement or replace CNN backbones in face recognition systems. While CNNs like ResNet+ArcFace have dominated this area, recent studies use ViT architectures to learn face embeddings, sometimes in a hybrid form (CNN for low-level features, Transformer for high-level aggregation). The benefit is again the global reasoning: a Transformer-based face model can capture interactions between distant facial landmarks or skin patches that a CNN might not link until very deep layers.

- Iris Recognition: An eye's iris pattern is a textured pattern – some works use Transformers (or local-attention variants) to encode the ring-shaped iris texture, taking advantage of the long-range pattern comparisons that attention enables (for instance, relating two distant iris crypts or furrows).

- Fingerprint and Palmprint Recognition: (We will cover fingerprint in depth in Chapter 2.) Transformers have been used to encode whole fingerprint images into fixed-length representations or to enhance minutiae-based matching. Similarly, palmprints (which have features like principal lines and texture) have been analyzed with ViT-based models.

- Vein Pattern Recognition: ViTs have been very effective in vascular biometrics, such as finger vein, palm vein, or wrist vein recognition. These tasks involve patterns of blood vessels beneath the skin captured via near-infrared imaging. A CNN might pick up local vein segments, but a ViT can better integrate the overall vein network structure. In fact,

the first application of pure pre-trained ViTs in vascular biometrics (Garcia-Martin et al., 2023) achieved exceptionally high identification rates across finger, palm, dorsal hand, and wrist vein datasets, after fine-tuning from ImageNet. For example, on a finger vein dataset (HKPU), the ViT-based approach reached ~99.5% True Positive Identification Rate, and similarly high performance (~99%) was observed on a variety of vein datasets. These results surpassed prior CNN-based methods, illustrating the viability of Transformers in even small-sample biometric domains when transfer learning is used.

One reason ViTs excel in biometrics is their capacity to model high-level feature interactions without inductive bias. CNNs have a built-in bias for local translational features, which is useful but can also miss global patterns (like the overall topology of veins or ridges). Transformers, by contrast, treat the image more like a graph of patches where any patch can influence any other. This can uncover subtle global patterns important for identity. Additionally, ViTs can be more parameter-efficient in some cases – for a given level of accuracy, a ViT may require fewer multiply-add operations than an equivalently performing CNN (though ViTs usually need more data to train from scratch). The ability to train ViTs on generic data and fine-tune means practitioners can deploy Transformers in biometrics without always requiring massive biometric datasets – a significant practical advantage.

Case Study: Vein Pattern Recognition with Transformers

Vein pattern recognition is an emerging biometric modality that benefits greatly from Transformer architectures. Vein biometrics (e.g., scanning the blood vessel patterns in one's finger, hand, or forehead) are attractive because they are hidden (internal) traits – difficult to counterfeit and unaffected by external obscuration like gloves or cosmetics. However, vein images often have complex patterns of branching vessels that span the entire image. Traditional CNN approaches can detect local vessel segments but may miss the forest for the trees if the vessel network is discontinuous or partially visible. Transformers, with global self-attention, are well-suited to capture the connectivity and overall layout of vein networks.

A notable example is the work by Garcia-Martin and Sanchez-Reillo (2023), who applied Vision Transformers to Vascular Biometric Recognition (VBR). They fine-tuned pre-trained ViT models on fourteen vein datasets encompassing finger vein, palm vein, dorsal hand vein, and wrist vein images. The ViTs achieved extremely high identification accuracies (in many cases >99% TPIR) across these datasets. These results not only outperformed many CNN-based benchmarks but also demonstrated the versatility of a single Transformer architecture across different vein modalities. The key advantage observed was the transfer learning capability: by leveraging features learned from general computer vision data, the ViT could recognize vein patterns even with the limited samples typical in each vein dataset.

Beyond accuracy, explainability in vein transformers has been explored. In vein recognition, it's important to ensure the model is focusing on actual vein structures rather than, say, incidental artifacts or lighting differences. Researchers have used attention map visualization (discussed

later) to confirm that ViT models focus on the regions with significant vein patterns. For instance, an explainability study reported that recognition performance improves when the ViT's attention is concentrated on true vein regions, and they visualized attention to show the model highlighting the vein lines on wrist images [6]. This not only builds trust in the system (the model is looking at the "right" features) but also provides a measure of security: if the attention map were diffuse or focused on irrelevant regions, it could indicate the model is picking up spurious cues that might fail under presentation attacks or varied conditions.

Another case study is the Forehead Vein and Periocular Transformer by Sharma et al. (2025). In the wake of COVID-19, face recognition faltered due to masks, and fingerprint use declined due to hygiene concerns [7]. This prompted exploration of contactless biometrics like forehead subcutaneous veins (captured via IR) combined with periocular (eye region) images. Sharma et al. proposed a *Cross-Spectral Vision Transformer (CS-ViT)* for this task. Their model uses a dual-channel Transformer: one stream for vein images and one for periocular images, with specialized cross-spectral attention modules that connect the two. Each channel's Transformer uses a Phase-Only correlation-based self-attention to remain robust to illumination and intensity differences between spectra. The CS-ViT achieved 98.8% classification accuracy, combining vein and periocular data – a remarkable performance demonstrating that even lightweight Transformer models can effectively fuse heterogeneous biometric traits. This case highlights how Transformers facilitate multi-biometric fusion: the attention mechanism can learn the optimal way to weight and combine features from two different sources (IR vein vs. visible eye image) for each individual, something that would be harder to achieve with separate models or simple score fusion. Moreover, by being *cross-spectral*, the model handles different imaging conditions within one architecture, showcasing Transformers' flexibility.

In both vein-focused cases, a theme emerges: Transformers can capture complex relational patterns inherent in biometric traits (like branching vein structures) better than purely local methods. They also easily support multi-input setups (two images, multi-view data, etc.) via multi-head and cross-attention. This makes them an excellent choice for advanced biometric systems that aim to be robust (working under occlusion/mask), multi-modal, and secure.

Case Study: Cross-Spectral Biometrics (CS-ViT)

(Covered partially in the previous section as the forehead vein + periocular example, we provide additional context here.)

Cross-spectral biometrics refers to matching or fusing data captured in different spectra or sensing modalities – for example, verifying if an infrared face image matches a visible-light face image, or using thermal hand vein images in conjunction with an RGB photo of a hand. This is challenging because different spectra reveal different aspects of the biometric trait. Traditional systems might require separate feature extractors and a carefully designed fusion algorithm. Transformers simplify this by offering a unified framework where different spectral inputs can be processed in parallel and then integrated through attention.

The CS-ViT by Sharma et al. [7] is a prime example. It employs a dual transformer encoder: one for each spectral channel. Crucially, it introduces a specialized attention mechanism called *Phase-Only Correlation Cross-Spectral Attention (POC-CSA)* that links the two channels. Phase-Only Correlation is a technique known in image matching to align images using the phase of their Fourier transform – by incorporating it into the attention, the CS-ViT effectively aligns features from the vein image with those from the periocular image, making the attention robust to resolution or illumination differences between IR and visible images. In practice, one can imagine that the model learns relationships like: “this pattern in the vein image (e.g., a certain bifurcation in the forehead veins) corresponds to that pattern in the periocular region (e.g., a particular wrinkle or brow shape)” which might correlate across people. The cross-spectral attention ensures the combined representation leverages these correlations. The result is a highly discriminative fused representation, as evidenced by near-perfect accuracy on a test database.

From a security perspective, cross-spectral transformers like CS-ViT add resilience against spoofing or environmental challenges. A face mask might block many visible facial features, but an IR vein camera can still capture the unique vein pattern. A combined transformer will recognize the person from the vein pattern even if the rest of the face is covered. Similarly, if lighting is poor for a normal camera, an IR-based trait could fill in. The attention mechanism’s flexibility means the model can lean more on whatever modality is more informative for a given subject or condition. This adaptive multi-modal feature usage is another strength of Transformers that static fusion methods lack.

In summary, cross-spectral biometric Transformers represent a cutting-edge application where Transformer architectures unify heterogeneous biometric sources. They exemplify the broader trend in biometrics to move beyond single-modality recognition toward integrated systems that use whatever biometric evidence is available, exactly the kind of problem Transformers are inherently designed to tackle (integrating information from multiple signals through learned attention).

Visualizing Attention Maps for Interpretability and Security

Transformers offer not only powerful performance but also new ways to interpret model decisions. Each Transformer layer produces attention weight matrices that show how the model is attending to different parts of the input. By visualizing these attention maps, we can gain insight into what the model “thinks” is important. This is particularly valuable in biometrics for both interpretability (to verify the model uses genuine biometric features) and security (to ensure the model isn’t focusing on an artifact that could be exploited).

One common approach is to visualize the attention from the classification token (in a ViT) to the image patch tokens. For an image input, one can take the final layer’s attention matrix for the class token and reshape it to the spatial layout of patches. The result is essentially a heatmap over the image, highlighting regions that contributed most to the model’s decision. There are also techniques like Attention Rollout, which accumulates attention weights across all layers to

get a more holistic importance map. The idea from Abnar & Zuidema (2020) is to treat the attention matrices as directed connectivity and multiply them through the layers (with appropriate normalization and identity skip addition) to obtain an overall influence map. This rolled-out attention can then be visualized. In practice, this means we can produce an image where, for example, the forehead and eyes light up for a face recognition transformer (meaning those areas were most influential), or the core and delta regions light up for a fingerprint transformer (indicating the model focused on those global landmarks and minutiae around them).

Figure 1 shows an example of attention visualization on incomplete fingerprint images. In this example, the model (Finger Recovery Transformer, see Chapter 2) is dealing with partial or noisy fingerprint inputs. The attention maps (overlaid in red intensity) indicate which parts of the image the transformer is focusing on to reconstruct and recognize the fingerprint. We can see that even when large portions of the fingerprint are missing or smudged, the model concentrates on the remaining clear ridge flows and minutiae-dense areas, effectively ignoring the gaps or noise. This aligns with how a forensic examiner might focus on any available minutiae in a partial print. Such visualizations confirm that the transformer isn't guessing blindly; it's leveraging the structured information that remains.

Fig. 1: Examples of partial fingerprint inputs (left: rolled print with missing area; middle: a "snapped" finger with only tip captured; right: a latent fingerprint with noisy background) and the attention map learned by a Transformer-based model. The red overlay highlights regions with high attention – the model focuses on the available ridge structures and minutiae while disregarding areas of missing or irrelevant data. This interpretability check ensures the model's decisions are based on legitimate fingerprint features, enhancing trust and security.

Visualizing attention is not limited to class tokens. We can also inspect intermediate attention between patches. For instance, one could pick a particular patch (say a region of a vein image) and see which other patches it attends to strongly – this might reveal that a certain vein bifurcation patch attends strongly to another patch further along the same vein, indicating the model has linked those two parts of the vein as part of one structure. This kind of interpretability is valuable for debugging models (are they looking at the finger's pattern or the background? are they focusing on a scar that might not be stable as an identifier?) and for user confidence (presenting a user with a heatmap on their face showing the system focused on their eyes and not, say, an irrelevant backdrop).

From a security standpoint, attention maps can help detect if a model is focusing on potentially spoofable cues. Suppose a face transformer consistently paid high attention to the border of the image; that might indicate it's picking up on some sensor-specific artifact rather than facial features, which could be a vulnerability. By identifying such issues via attention visualization, developers can retrain the model or adjust the data to refocus attention on proper biometric traits.

It's important to note that attention visualization is an explanatory tool, but not a complete explanation of a model's decision. There is debate in the literature whether attention weights

directly correlate with feature importance. Nonetheless, in practice, they often provide a useful approximation. Techniques like Grad-CAM (from CNNs) have analogues in transformers as well, and hybrid methods are being researched [8]. Still, the direct availability of attention matrices in Transformers makes them one of the more transparent deep learning models for analysis.

In conclusion, attention map visualization serves as a bridge between the complex computations of a transformer and human understanding. For biometric systems that operate in high-stakes domains (security checkpoints, forensic analysis), having the ability to explain why the model matched a fingerprint or flagged a face is crucial. Transformers offer this to a greater extent than previous biometric models, thereby not only improving performance but also enabling interpretable and trustworthy biometric AI systems.

Summary

Transformers have revolutionized the landscape of deep learning, and their impact is increasingly felt in biometric analysis. In this chapter, we reviewed how the field progressed from CNNs and RNNs towards transformers to overcome limitations in modeling long-range dependencies and parallelizing computation. We examined the core innovations of the transformer architecture – particularly self-attention and positional encodings – and why they are a natural fit for biometric data, which often contains globally dependent features (like patterns spread across a fingerprint or face). Vision Transformers have matched or surpassed CNN performance in many vision tasks, and biometrics is no exception: we discussed case studies in vein pattern recognition and cross-spectral biometrics where transformers achieved state-of-the-art results by effectively fusing information and capturing complex feature relationships. A key advantage of transformers in these applications is their flexibility – a single architecture can integrate multiple modalities and focus on the most salient traits of a biometric identifier, yielding robust performance even under occlusions or spectrum changes. Finally, we highlighted how attention maps from transformers can be visualized to interpret model decisions, giving insight into what parts of a biometric trait the model relies on. This interpretability not only builds confidence in automated systems but can enhance security by ensuring models use genuine biometric features rather than spurious cues. In summary, transformers provide a powerful, unifying framework for biometric analysis, offering accuracy, adaptability, and transparency. The next chapter will delve deeper into a specific application: using transformers for fingerprint feature extraction, where we will see these concepts applied to one of the oldest and most studied biometric modalities.
