Appendix A

Appendix A.1 Block-wise Fourier transform scoring We propose a scoring mechanism based on block-wise Fourier analysis. The static score is derived from features extracted with this transform; the workflow is shown in Figure 1.

First, each image is partitioned into non-overlapping 16×16 pixel blocks. The starting coordinate is set to $(64, 64)$ to avoid edge noise and ensure that most blocks cover the facial region. For an image of size $H \times W$, the stride equals the block size so that the effective area is covered without overlap.

For each block blk, a 2-D fast Fourier transform is applied to map the spatial signal to the frequency domain:

$$F(u, v) = \mathcal{F}\{\mathrm{blk}(x, y)\}$$

where (x, y) and (u, v) denote spatial and frequency coordinates, respectively. To facilitate low-frequency analysis, the spectrum is shifted to the centre. We then apply a logarithmic transform to the magnitude,

$$\log\big(1 + |F(u, v)|\big)$$

which compresses the dynamic range and stabilises the features. The log-magnitude spectrum of each block is flattened into a one-dimensional vector, yielding a local feature $\mathbf{x}_j \in \mathbb{R}^D$. For a single image, if N valid blocks are extracted, we obtain an $N \times D$ feature matrix.

Pairwise similarities between blocks are measured with the Euclidean distance:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{k=1}^{D} (x_{i,k} - x_{j,k})^2}$$

Let $M = \frac{N(N-1)}{2}$ be the number of unordered block pairs and d_i the distance for the i-th pair. We compute descriptive statistics over all distances:

$$\mathrm{mean} = \frac{1}{M} \sum_{i=1}^{M} d_i$$

$$\max = \max\{d_1, d_2, \ldots, d_M\}$$

$$\mathrm{std} = \sqrt{\frac{1}{M} \sum_{i=1}^{M} (d_i - \mathrm{mean})^2}$$

$$\mathrm{var} = \frac{1}{M} \sum_{i=1}^{M} (d_i - \mathrm{mean})^2$$

These statistics constitute the block-wise Fourier transform features used to compute the static score.

Table 1 reports variance and mean of frequency-domain correlation features, along with mean, maximum, standard deviation, and variance, for different FF++ categories. DF and NT exhibit markedly different statistics from real images ("origin"). For example, the variance mean/median of DF and NT deviate far more from Origin, revealing asymmetric frequency-energy distributions and a substantive gap from real data.

Table 1: Frequency-domain correlation statistics on FF++.

Category	fft_{mean}	fft_{max}	fft_{std}	fft_{var}
DF	18.84	43.46	6.29	41.39
F2F	19.48	45.00	6.71	47.06
FS	19.56	45.14	6.67	46.07
NT	19.20	43.91	6.53	44.85
origin	19.57	45.18	6.71	47.31

Table 2: Performance comparison using frequency-domain correlation variance (frame-level AUC). We form two test groups: F2F+FS and NT+DF. Samples are sorted by ascending fft_{var}; the 50th percentile splits them into high-variance (P50+) and low-variance (P50-) subsets.

Group	**CDFv1**	**CDFv2**	**DFDCP**	**UADFV**
FF++	0.791	0.769	0.790	0.933
F2F+FS	0.783	0.762	0.777	0.943
NT+DF	0.838	0.790	0.803	0.945
P50-	0.751	0.709	0.747	0.939
P50+	0.891	0.822	0.837	0.961

To probe distributional gaps, we compute variance features after the block-wise Fourier transform and plot their density curves in Figure 2. The fft_{var} distribution of real images is close to F2F and NT, but diverges from DF and FS. Specifically, DF shows a taller peak shifted left, with noticeably higher density in the mid-range (20–40). FS peaks about 22% higher than real images; although its mode does not shift, its mid-range density is also elevated.

Based on these observations, we hypothesise that training on data that deviate more strongly from real images enables the model to capture essential forgery characteristics more efficiently. Accordingly, we build two training sets NT+DF, FS+F2F and compare results. Table 2 shows that the NT+DF-trained model achieves higher accuracy and F1-score on the validation set, suggesting that their larger distributional shift injects more discriminative "knowledge."

Focusing further on frequency variance, we keep the real subset fixed and divide fake samples at the median into P50- and P50+. Training on the high-discrepancy subset (P50+) yields better performance: relative to using all data, the model gains 12.6% on CDF-v1—an appreciable improvement.

Appendix A.2 Face-image quality scoring. We construct a static scoring system based on facial image quality. Generic Face Image Quality Assessment (GFIQA) estimates perceptual fidelity and is crucial for selecting high-quality inputs in downstream tasks. We adopt the DSL-FIQA method (Chen et al. 2024) to score all FF++ images and obtain face-quality annotations. Representative examples are shown in Figure 3. Comparing high- and low-score samples reveals that high-score images are typically sharper

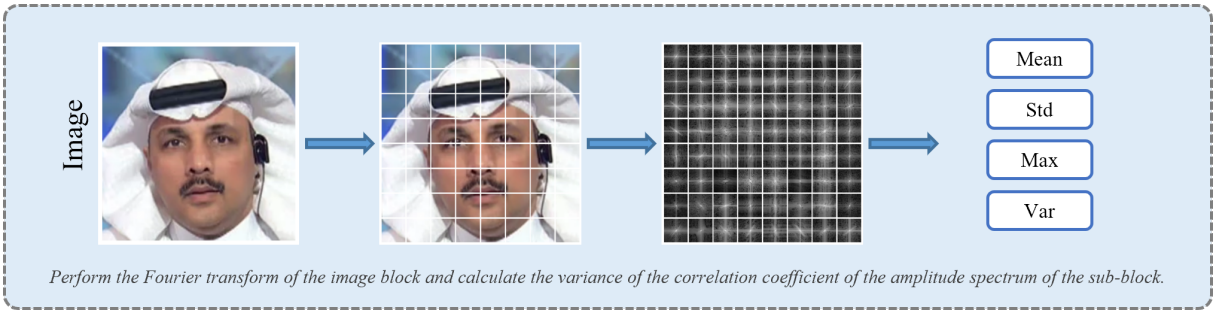

Figure 1: Pipeline for extracting image-feature statistics throught block-wise Fourier transform.

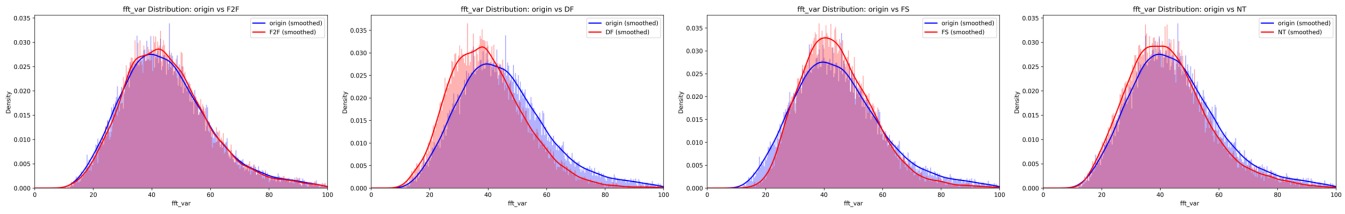

Figure 2: Density comparison of block-wise fourier transform scores across forgery types and origin images in FF++.

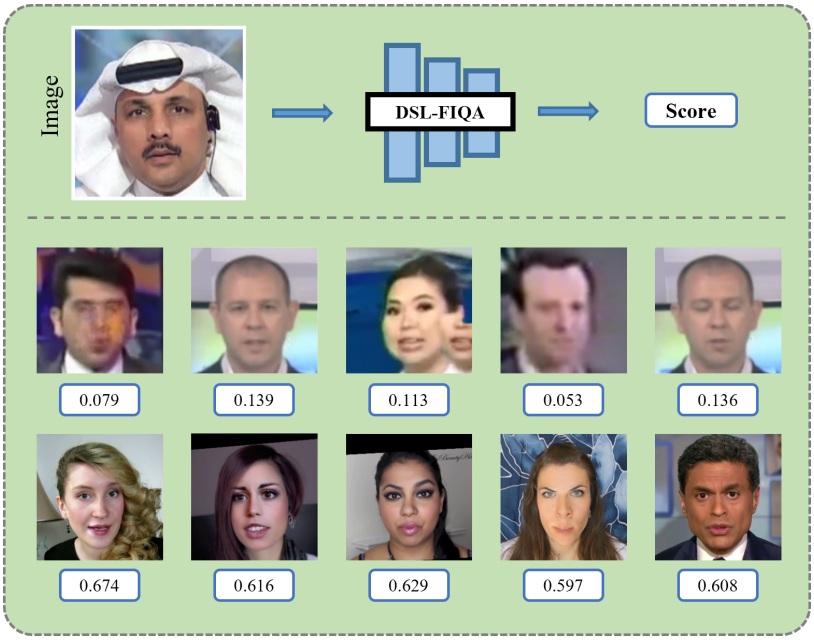

Figure 3: Static face-quality scoring and representative examples.

Table 3: Performance comparison using face-quality score (frame-level AUC). Samples are sorted in ascending order, and the 25th (P25), 50th (P50), and 75th (P75) percentiles are used to divide them into four segments.

Group	Data Size	CDFv1	CDFv2	DFDCP	UADFV
FF++	114,884	0.791	0.769	0.790	0.933
P25-	45,952	0.592	0.704	0.821	0.832
P50-	68,864	0.671	0.750	0.852	0.890
P50+	68,864	0.886	0.806	0.704	0.907
P75+	45,952	0.870	0.759	0.676	0.862

and better blended—making forgery traces less perceptible—whereas low-score images exhibit poorer overall quality and more obvious artefacts.

Figure 4 plots the density curves of face-quality scores for each forgery type versus real images. Consistent with Figure 2 (frequency variance), the distributions for DF and FS exhibit a clear shift relative to Origin, indicating more pronounced defects in facial feature alignment.

We sort samples by quality score and split them at the 25th, 50th, and 75th percentiles (P25, P50, P75), forming four segments. Cross-domain results are non-monotonic (Table 3). On CDF, the model performs best on high-score images (a 47% gain over low-score ones); on DFDCP it is markedly more sensitive to low-score images; UADFV shows no clear preference. This divergence suggests that a single, quality-based sampling policy should be domain-adaptive, tuned to the characteristics of the target set.

Appendix A.3 Block-wise Scoring in the YCbCr Space Natural images display statistically similar noise levels in the luminance and chrominance channels, whereas editing operations (e.g., compression, filtering, or face swapping) often disturb this balance. We therefore design a spatial-domain score that measures cross-channel noise consistency in the YCbCr space. The processing pipeline is depicted in Figure 5 and summarised below.

Each RGB image is first converted to the YCbCr colour space and split into one luminance channel Y and two chrominance channels Cb and Cr:

$$\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.1687 & -0.3313 & 0.5000 \\ 0.5000 & -0.4187 & -0.0813 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} + \begin{bmatrix} 0 \\ 128 \\ 128 \end{bmatrix}$$

After conversion, each channel is divided into non-overlapping 16×16 blocks. For a given block B_i, local noise variance is estimated as

$$\sigma^2_{N,i} = \max\big(0,\ \mathrm{var}(B_i) - \mathrm{var}(S_i)\big)$$

where S_i denotes the Gaussian-filtered version of B_i; subtracting $\mathrm{var}(S_i)$ removes the signal component and isolates noise. To quantify cross-channel inconsistency, the absolute variance differences between channel pairs are calculated for every block:

Table 4: Performance comparison using S_{noise} (frame-level AUC). Samples are sorted in ascending order, and the 25th (P25), 50th (P50), and 75th (P75) percentiles are used to divide them into four segments.

Group	Data Size	CDFv1	CDFv2	DFDCP	UADFV
FF++	114,884	0.791	0.769	0.790	0.933
P25-	45,952	0.677	0.744	0.738	0.882
P50-	68,864	0.765	0.779	0.751	0.942
P50+	68,864	0.778	0.687	0.725	0.939
P75+	45,952	0.729	0.615	0.705	0.893

$$\Delta\sigma^2_{Y\text{–}Cb,i} = \left|\sigma^2_{N,i}(Y) - \sigma^2_{N,i}(Cb)\right|$$
$$\Delta\sigma^2_{Y\text{–}Cr,i} = \left|\sigma^2_{N,i}(Y) - \sigma^2_{N,i}(Cr)\right|$$
$$\Delta\sigma^2_{Cr\text{–}Cb,i} = \left|\sigma^2_{N,i}(Cr) - \sigma^2_{N,i}(Cb)\right|$$

Averaging these three channel-pair differences over all blocks yields the image-level score

$$S_{\text{noise}} = \frac{\mu_{Y\text{–}Cb} + \mu_{Y\text{–}Cr} + \mu_{Cr\text{–}Cb}}{3}$$

where each μ_* represents the mean of $\Delta\sigma^2_*$ across the image. A larger S_{noise} indicates poorer cross-channel noise consistency and therefore a higher likelihood of manipulation.

Figure 6 plots kernel-density estimates of S_{noise} for real images and the four forgery types in FF++. Real images cluster in the low-score region, whereas forged images shift left, confirming that face-swap operations disrupt channel-noise balance—although the separation is less pronounced than in Sections and .

Sorting FF++ samples by S_{noise} and splitting at the 25th, 50th, and 75th percentiles yields four quality tiers. As shown in Table 4, cross-domain performance drops in almost every tier, unlike the monotonic trends observed for frequency- and face-quality scores. This suggests that noise-consistency alone contributes limited discriminative power and may even introduce noise when other salient forgery cues are ignored; a purely spatial-noise sampling policy therefore requires domain-specific tuning.

Appendix A.4 Discussion We examine three scoring mechanisms for forgery detection—frequency-domain features, spatial-domain features, and deep-learning–based measures. The block-wise Fourier score effectively captures spectral discrepancies between forged and real images and delivers clear gains on manipulations such as NeuralTextures and DeepFakes. The face-image quality score targets perceptual fidelity and performs well on high-quality inputs, but its behaviour is inconsistent under cross-domain evaluation.

Our training strategy echoes the empirical observation in CDFA that "using only FF++ at the initial stage leads to faster convergence" (Song, Lin, and Li 2024). CDFA did not provide a mechanistic explanation, and our analysis fills this gap. In FF++, certain forgeries (for example, NT and DF) deviate markedly from the real-image distribution

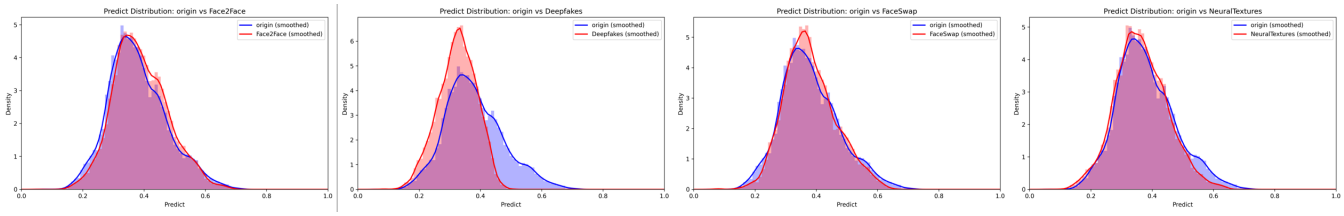

Figure 4: Density comparison of face-quality scores across forgery types and origin images in FF++.

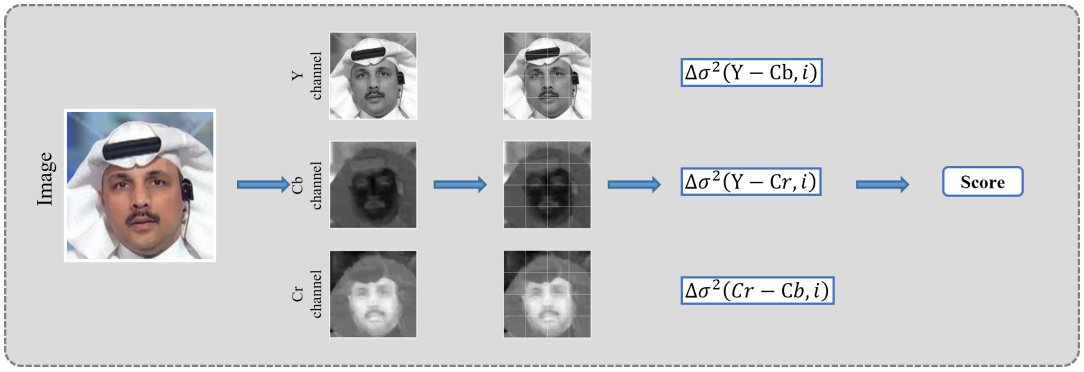

Figure 5: Static scoring based on noise-consistency across colour channels.

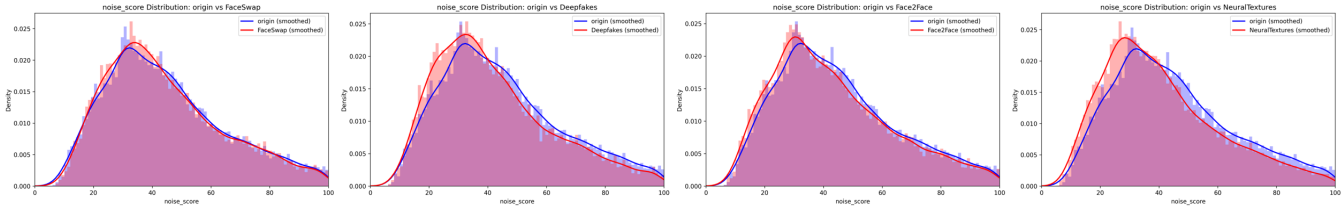

Figure 6: Density comparison of block-wise fourier transform scores across forgery types and origin images in FF++.

and present a clear, structured gap in feature space. Low-variance NT and DF samples, in particular, carry strong discriminative signals. When we sort training samples by variance, the curriculum lets the model stabilise on the most salient cues first and then adapt to subtler artefacts. This ordering reduces optimisation oscillation, shortens convergence time, and enhances generalisation, thereby offering a feature-distribution-level and strategy-level rationale for the faster convergence observed in CDFA.

Appendix B

Table 5: Prompt templates for augmented captions.

Fake-image Prompt

`Generate one English sentence with the structure: [Declaration] [Scene] [Defects].`
`[Declaration]:` `In a word, you need to specify whether this is a generated image or a forged image, such as: "A computer-generated image" and similar words.`
`[Scene]:` `Be consistent with this sentence, {{Scene}}.`
`[Defects]:` `In a word, it is necessary to give the basic flaws of deepfake, such as: "facial contours have artifacts", "facial expressions are unnatural", "eyes are unnatural", etc.`
`For example: A modified image, a man with a suit and tie and a beard, with facial expressions that are unnatural.`

Real-image Prompt

`Generate one sentence in English following this structure: [Declaration] [Scene].`
`[Declaration]:` `In a word, you need to point out that this is a real/natural image, such as: "A real image" and similar words.`
`[Scene]:` `Be consistent with this sentence, {{Scene}}.`
`For example: A naturel image, a man with a suit and tie and a beard.`

We first employ CLIPCAP to generate an initial caption for each FF++ image and then refine that caption with a large language model to obtain the final label (Table 5).

For *fake* images, the label contains three slots, including **[Declaration]**, **[Scene]**, and **[Defects]**.

The declaration explicitly states that the image is generated or forged; the scene string preserves the original visual description; and the defects slot lists common deepfake artefacts such as contour anomalies or unnatural expressions.

Table 6: Effect of LoRA rank r and scaling factor α on parameter count and detection accuracy (video-level AUC).

r	α	#param	CDFv2	DFDCP	Avg.
1	2	0.09M	0.947	0.923	0.935
1	4	0.09M	0.953	**0.928**	**0.941**
1	8	0.09M	0.942	0.920	0.931
4	4	0.38M	0.950	0.923	0.937
4	8	0.38M	0.952	0.923	0.938
4	16	0.38M	**0.956**	0.921	0.939

Table 7: Comparison of PEFT methods for cross-dataset deepfake detection (video-level AUC).

PEFT	CDFv1	CDFv2	DFDCP
LoRA	**0.973**	0.942	**0.920**
AdaLoRA	0.962	0.942	0.912
BitFit	**0.973**	**0.948**	0.888
LNTuning	0.944	0.922	0.863
LoHA	0.961	0.941	0.905

For *real* images, only the declaration and scene slots are retained.

Appendix C

Ablation study and analysis

LoRA ablation study. We evaluate how the LoRA rank r and scaling factor α affect detection accuracy (Table 6). With $r = 1$, the model introduces only 0.09M additional parameters, and $\alpha = 4$ provides the strongest performance. Raising the rank to $r = 4$ increases the best AUC on CDF-v2 by just 0.003, while slightly reducing the score on DFDCP. Balancing accuracy against parameter efficiency, we fix $r = 1$ and $\alpha = 4$ in all subsequent experiments, retaining state-of-the-art accuracy while keeping the parameter budget minimal.

PEFT ablation study. Table 7 compares parameter-efficient fine-tuning (PEFT) strategies—LoRA, AdaLoRA, BitFit, LN-Tuning, and LoHA—reporting video-level AUC on CDF-v1, CDF-v2, and DFDCP. LoRA delivers the best average performance with strong cross-dataset generalisation. BitFit, while using the fewest trainable parameters, attains reasonable accuracy on the CDF series but degrades on the more challenging DFDCP set, indicating limited capacity to capture fine-grained, out-of-distribution artefacts.

References

Chen, W.-T.; Krishnan, G.; Gao, Q.; Kuo, S.-Y.; Ma, S.; and Wang, J. 2024. DSL-FIQA: Assessing Facial Image Quality via Dual-Set Degradation Learning and Landmark-Guided Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2931–2941.

Song, W.; Lin, Y.; and Li, B. 2024. Towards General Deep-

fake Detection with Dynamic Curriculum. *arXiv preprint arXiv:2410.11162*.