Supplementary Material of ViTNT-FIQA: Training-Free Face Image Quality Assessment with Vision Transformers

Guray Ozgur[1,2], Eduarda Caldeira[1,2], Tahar Chettaoui[1,2], Jan Niklas Kolf[1,2],
Marco Huber[1,2], Naser Damer[1,2], Fadi Boutros[1]
[1]Fraunhofer IGD, Germany, [2]TU Darmstadt, Germany

This supplementary material provides comprehensive experimental results and detailed analysis of the *ViTNT-FIQA* method for face image quality assessment. The supplementary material is structured to address five fundamental research questions: (1) How do cross-block patch embedding distances correlate with image quality across different network depths? (2) Which block configurations provide optimal performance-efficiency trade-offs? (3) How does our method compare against existing state-of-the-art approaches across multiple evaluation metrics? (4) What visual patterns emerge in the ablation study EDC curves? (5) How does quality score distribution vary across different FIQA methods?

Tables: Quantitative Evidence

- **Table 1**: Block window analysis comparing consecutive 6-block segments across ViT-B architecture. This systematic evaluation identifies which transformer block windows (early: 0-5, middle: 6-17, late: 18-23) capture the most quality-discriminative information. We include this analysis to demonstrate that *early transformer blocks (0-5) achieve superior performance* across both AUC-EDC and pAUC-EDC metrics, providing empirical evidence that quality-relevant features emerge in initial processing stages rather than requiring full network depth.
- **Table 2**: Comprehensive ablation study presenting AUC-EDC performance for all four design choices (Dataset, Architecture, Block Depth, Attention-Weighting) across eight benchmark datasets using ArcFace as the face recognition model. Lower AUC-EDC values indicate better quality assessment performance. This systematic evaluation complements the pAUC-EDC analysis of the main paper, providing a complete picture of method performance across the full rejection rate spectrum (0-100%) rather than just the partial area up to 25% rejection. The AUC-EDC metric is essential to validate that our findings hold beyond the 25% rejection threshold.
- **Table 3**: State-of-the-art comparison presenting AUC-EDC values for our method against 15 competing approaches (3 IQA, 12 FIQA) across four face recognition models (ArcFace, ElasticFace, MagFace, CurricularFace). We provide this extensive comparison table in addition to the pAUC-EDC comparison, provided in the main paper, to establish the competitive advantage of our training-free approach across different evaluation metrics and operating points.

Figures: Visual Evidence and Insights

- **Figure 1**: Boxplots of mean L2 distances between consecutive ViT-S patch embeddings across 11 quality groups from 5.5M SynFIQA images. This visualization complements the SynFIQA figure in the main paper (which shows ViT-B results) by demonstrating that *our core hypothesis generalizes across different architecture scales*. The systematic decrease in cross-block distances with increasing ground-truth quality is evident in ViT-S (12 blocks) just as in ViT-B (24 blocks), confirming that patch embedding stability serves as a quality indicator regardless of network depth.
- **Figures 2, 3, 4**: Comprehensive ablation analysis via Error-versus-Discard Characteristic (EDC) curves at three security operating points (FMR=$1e-2$, $1e-3$, $1e-4$). *The visual trends provide intuitive insights that complement the quantitative tables*: In the Dataset study column, blue/green curves (WebFace4M/WebFace12M) consistently lie below brown/pink curves (CLIP/FRoundation), visually confirming FR-specific training superiority. In the Architecture study, blue (ViT-B) and orange (ViT-S) curves run nearly parallel with minimal separation, demonstrating depth-independence. In the Block Depth study, we observe progressive downward shifts from red (4 blocks) to blue (24 blocks) up to 16 blocks, after which curves plateau or slightly rise, visually identifying the optimal 12-20 block sweet spot. In the Attention-Weighting study, we see a similar trend to Block Depth study, but with slightly lower curves. In the Block Windows study, the clear downward progression of the EDC curve in the early

blocks (blocks 0-5), not seeing a similar downward progression for the others, visually confirms that quality discrimination concentrates in initial processing stages.

- **Figures 5, 6**: State-of-the-art comparison EDC curves at two additional FNMR@FMR thresholds ($1e-2$, $1e-4$) beyond the main paper's $1e-3$ threshold. These comprehensive comparisons across eight benchmark datasets and four face recognition models demonstrate that *our method's competitiveness holds across multiple security requirements*.
- **Figure 7**: Distribution of quality scores across evaluation benchmarks, comparing *ViTNT-FIQA* with SOTA methods. The normalized score distributions (range [0, 1]) reveal whether methods produce well-calibrated distributions that effectively rank sample quality or if they suffer from range compression that limits discriminative power.

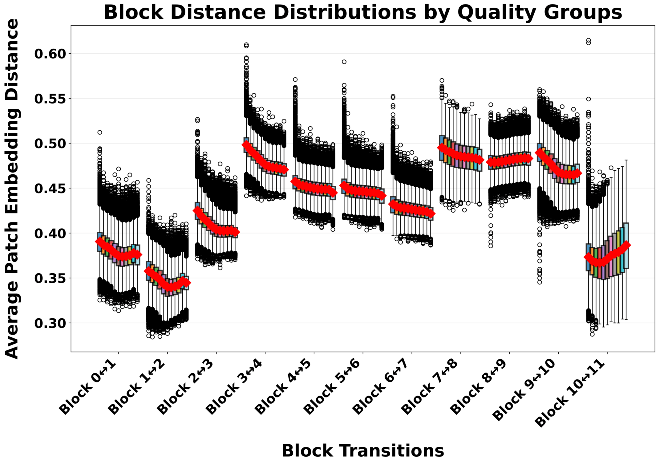

Figure 1. Boxplots of mean L2 distances between corresponding patch embeddings from consecutive ViT-S blocks computed for 11 quality groups, each having 0.5M images, from 5.5M images of SynFIQA [23]. Each box summarizes the distribution of average patch-embedding distances across images in a quality group, lower distances empirically correspond to higher ground-truth quality for most block transitions, i.e. the higher the quality, the lower the distance.

Table 1. Block-window analysis comparing quality-assessment performance across consecutive 6-block segments of ViT-B/WebFace4M. Early window (blocks 0–5) yields the strongest AUC-EDC and pAUC-EDC performance, indicating that initial feature refinements carry the most quality-discriminative signal. Mean metrics reported across seven benchmarks at FMR=1e−3 and 1e−4. Best results per metric are in bold.

Metric	**Method**	**Blocks**	Adience [9]		AgeDB-30 [20]		CFP-FP [24]		LFW [12]		CALFW [28]		CPLFW [27]		XQLFW [14]		Mean	
			$1e{-}3$	$1e{-}4$	$1e{-}3$	$1e{-}4$	$1e{-}3$	$1e{-}4$	$1e{-}3$	$1e{-}4$	$1e{-}3$	$1e{-}4$	$1e{-}3$	$1e{-}4$	$1e{-}3$	$1e{-}4$	$1e{-}3$	$1e{-}4$
AUC-EDC	ViT-B @ 0-5	0-5	**0.0358**	**0.0808**	0.0375	0.0524	**0.0088**	**0.0131**	**0.0024**	**0.0030**	**0.0660**	**0.0724**	**0.0578**	**0.0786**	**0.2241**	**0.2714**	**0.0618**	**0.0817**
	ViT-B @ 2-7	2-7	0.0749	0.1716	0.0355	0.0496	0.0293	0.0418	0.0045	0.0051	0.0835	0.0915	0.1434	0.1785	0.5502	0.6302	0.1316	0.1669
	ViT-B @ 4-9	4-9	0.0655	0.1361	0.0295	0.0314	0.0335	0.0429	0.0023	0.0029	0.0670	0.0750	0.1287	0.1595	0.5536	0.6263	0.1257	0.1534
	ViT-B @ 6-11	6-11	0.0573	0.1318	0.0354	0.0412	0.0397	0.0489	0.0042	0.0050	0.0799	0.0865	0.1241	0.1633	0.4911	0.5467	0.1188	0.1462
	ViT-B @ 8-13	8-13	0.0480	0.1108	0.0309	0.0452	0.0261	0.0352	0.0046	0.0054	0.0695	0.0762	0.1569	0.1989	0.5570	0.6110	0.1276	0.1547
	ViT-B @ 10-15	10-15	0.0461	0.1040	0.0370	0.0516	0.0250	0.0350	0.0040	0.0050	0.0678	0.0748	0.1780	0.2235	0.5225	0.5666	0.1258	0.1515
	ViT-B @ 12-17	12-17	0.0683	0.1578	0.0342	0.0405	0.0360	0.0424	0.0028	0.0035	0.0843	0.0927	0.1790	0.2213	0.4644	0.5843	0.1241	0.1632
	ViT-B @ 14-19	14-19	0.0654	0.1458	0.0335	0.0366	0.0397	0.0513	0.0036	0.0042	0.0773	0.0813	0.1306	0.1566	0.4717	0.5622	0.1174	0.1483
	ViT-B @ 16-21	16-21	0.0574	0.1364	0.0294	0.0434	0.0264	0.0352	0.0044	0.0052	0.0802	0.0865	0.1694	0.2037	0.5539	0.6163	0.1316	0.1610
	ViT-B @ 18-23	18-23	0.0640	0.1453	**0.0258**	**0.0286**	0.0363	0.0450	0.0038	0.0044	0.0702	0.0770	0.1229	0.1589	0.5070	0.5617	0.1186	0.1458
pAUC-EDC	ViT-B @ 0-5	0-5	**0.0127**	**0.0293**	0.0090	0.0139	**0.0048**	**0.0074**	**0.0007**	**0.0008**	**0.0187**	**0.0205**	**0.0240**	**0.0364**	**0.1256**	**0.1439**	**0.0279**	**0.0360**
	ViT-B @ 2-7	2-7	0.0155	0.0355	0.0084	0.0134	0.0091	0.0138	0.0009	0.0011	0.0200	0.0223	0.0390	0.0527	0.1484	0.1665	0.0345	0.0436
	ViT-B @ 4-9	4-9	0.0151	0.0342	0.0083	**0.0096**	0.0090	0.0127	0.0009	0.0010	0.0196	0.0219	0.0376	0.0490	0.1451	0.1693	0.0337	0.0425
	ViT-B @ 6-11	6-11	0.0145	0.0335	0.0088	0.0109	0.0092	0.0136	0.0009	0.0010	0.0204	0.0231	0.0353	0.0488	0.1489	0.1667	0.0340	0.0425
	ViT-B @ 8-13	8-13	0.0147	0.0335	**0.0082**	0.0136	0.0083	0.0130	0.0009	0.0011	0.0196	0.0220	0.0430	0.0552	0.1493	0.1670	0.0349	0.0436
	ViT-B @ 10-15	10-15	0.0140	0.0330	0.0095	0.0143	0.0079	0.0123	0.0010	0.0012	0.0197	0.0223	0.0466	0.0638	0.1495	0.1676	0.0355	0.0449
	ViT-B @ 12-17	12-17	0.0154	0.0356	0.0086	0.0111	0.0095	0.0129	**0.0007**	0.0009	0.0205	0.0227	0.0491	0.0633	0.1435	0.1640	0.0353	0.0444
	ViT-B @ 14-19	14-19	0.0146	0.0338	0.0084	0.0099	0.0098	0.0142	0.0008	0.0009	0.0200	0.0228	0.0363	0.0495	0.1383	0.1672	0.0326	0.0426
	ViT-B @ 16-21	16-21	0.0152	0.0348	0.0089	0.0137	0.0085	0.0134	0.0009	0.0011	0.0199	0.0223	0.0436	0.0550	0.1485	0.1664	0.0351	0.0438
	ViT-B @ 18-23	18-23	0.0150	0.0336	0.0085	0.0105	0.0090	0.0132	0.0008	0.0010	0.0196	0.0218	0.0371	0.0483	0.1408	0.1661	0.0330	0.0421

Table 2. Ablation studies analyzing four design choices: dataset generalization (WebFace4M, WebFace12M, CLIP, FRoundation), architecture depth (ViT-S vs ViT-B), block depth trade-offs (4-24 blocks), and aggregation strategies (uniform vs attention-weighted). Results show optimal performance at 12-20 blocks with last-block attention weighting. Mean AUC-EDC computed across seven benchmarks at FMR=$1e-3$ and $1e-4$. Best per study in bold.

Study	**Method**	**Blocks**	Adience [9]		AgeDB-30 [20]		CFP-FP [24]		LFW [12]		CALFW [28]		CPLFW [27]		XQLFW [14]		Mean AUC-EDC	
			$1e{-}3$	$1e{-}4$	$1e{-}3$	$1e{-}4$	$1e{-}3$	$1e{-}4$	$1e{-}3$	$1e{-}4$	$1e{-}3$	$1e{-}4$	$1e{-}3$	$1e{-}4$	$1e{-}3$	$1e{-}4$	$1e{-}3$	$1e{-}4$
Dataset	ViT-B - WebFace4M	0-23	**0.0217**	**0.0406**	0.0255	**0.0362**	0.0133	0.0181	**0.0034**	**0.0039**	0.0707	**0.0745**	0.0538	**0.0759**	**0.2575**	**0.2978**	**0.0637**	**0.0781**
	ViT-B - WebFace12M	0-23	**0.0217**	0.0417	**0.0250**	0.0393	**0.0113**	**0.0175**	**0.0034**	0.0040	0.0907	0.0965	**0.0521**	0.0805	0.2994	0.3391	0.0719	0.0884
	CLIP	0-11	0.0634	0.1368	0.0364	0.0508	0.0351	0.0491	0.0033	0.0040	**0.0702**	0.0772	0.2049	0.2542	0.6065	0.6645	0.1457	0.1767
	FRoundation	0-11	0.0605	0.1415	0.0412	0.0595	0.0286	0.0374	0.0036	0.0043	0.0796	0.0852	0.1917	0.2527	0.5428	0.5881	0.1354	0.1670
Architecture	ViT-S	0-11	0.0231	0.0441	**0.0245**	**0.0337**	**0.0115**	**0.0164**	**0.0026**	**0.0032**	**0.0671**	**0.0718**	**0.0535**	0.0768	0.2730	0.3107	0.0650	0.0795
	ViT-B	0-23	**0.0217**	**0.0406**	0.0255	0.0362	0.0133	0.0181	0.0034	0.0039	0.0707	0.0745	0.0538	**0.0759**	**0.2575**	**0.2978**	**0.0637**	**0.0781**
Block Depth	ViT-B @ 4	0-3	0.0509	0.1109	0.0390	0.0533	0.0145	0.0200	0.0037	0.0043	0.0675	0.0734	0.0935	0.1160	0.2499	0.3023	0.0741	0.0972
	ViT-B @ 8	0-7	0.0304	0.0677	0.0340	0.0479	0.0072	0.0111	**0.0020**	0.0026	0.0622	0.0678	0.0492	0.0689	0.2129	0.2565	0.0568	0.0746
	ViT-B @ 12	0-11	0.0266	0.0577	0.0304	0.0429	**0.0065**	**0.0102**	0.0021	0.0026	**0.0598**	**0.0640**	0.0446	0.0640	**0.2043**	**0.2423**	0.0535	0.0691
	ViT-B @ 16	0-15	0.0240	0.0505	0.0272	0.0385	0.0069	0.0110	**0.0020**	**0.0025**	0.0603	0.0641	**0.0442**	**0.0634**	0.2071	0.2440	**0.0531**	**0.0677**
	ViT-B @ 20	0-19	**0.0213**	0.0412	0.0268	0.0377	0.0075	0.0124	0.0024	0.0029	0.0639	0.0670	0.0450	0.0638	0.2169	0.2527	0.0548	0.0682
	ViT-B @ 24	0-23	0.0217	**0.0406**	**0.0255**	**0.0362**	0.0133	0.0181	0.0034	0.0039	0.0707	0.0745	0.0538	0.0759	0.2575	0.2978	0.0637	0.0781
Attention-Weighting	Last Block Attention @ 4	0-3	0.0475	0.1046	0.0362	0.0513	0.0148	0.0203	0.0031	0.0038	0.0681	0.0737	0.0837	0.1073	0.2372	0.2885	0.0701	0.0928
	Last Block Attention @ 8	0-7	0.0286	0.0608	0.0335	0.0474	0.0072	0.0116	0.0023	0.0029	0.0626	0.0672	0.0474	0.0668	0.2072	0.2484	0.0555	0.0722
	Last Block Attention @ 12	0-11	0.0254	0.0539	0.0302	0.0419	**0.0060**	**0.0099**	**0.0021**	0.0027	0.0609	0.0651	0.0428	0.0617	**0.1993**	**0.2363**	**0.0524**	0.0674
	Last Block Attention @ 16	0-15	0.0240	0.0514	0.0263	0.0370	0.0064	0.0103	**0.0021**	**0.0026**	**0.0605**	**0.0640**	**0.0424**	**0.0613**	0.2054	0.2395	**0.0524**	0.0666
	Last Block Attention @ 20	0-19	**0.0203**	**0.0392**	0.0249	0.0349	0.0063	0.0101	0.0033	0.0038	0.0614	0.0650	0.0428	0.0619	0.2078	0.2473	**0.0524**	**0.0660**
	Last Block Attention @ 24	0-23	0.0228	0.0433	0.0282	0.0402	0.0135	0.0182	0.0039	0.0044	0.0668	0.0719	0.0507	0.0768	0.2619	0.3005	0.0640	0.0793
	Attention (All Blocks) @ 24	0-23	0.0213	0.0397	**0.0244**	**0.0346**	0.0093	0.0137	0.0036	0.0041	0.0704	0.0741	0.0464	0.0693	0.2434	0.2808	0.0598	0.0738

Table 3. The AUCs of EDC achieved by our method and the SOTA methods under different experimental settings. The notions of $1e-3$ and $1e-4$ indicate the value of the fixed FMR at which the EDC curves (FNMR vs. reject) were calculated. The results are compared to three IQA and twelve FIQA approaches. The XQLFW dataset uses SER-FIQ (marked with *) as the FIQ labeling method.

FR		Method	Adience [9]		AgeDB-30 [20]		CFP-FP [24]		LFW [12]		CALFW [28]		CPLFW [27]		XQLFW [14]		IJB-C [17]	
			$1e-3$	$1e-4$	$1e-3$	$1e-4$	$1e-3$	$1e-4$	$1e-3$	$1e-4$	$1e-3$	$1e-4$	$1e^{-3}$	$1e-4$	$1e-3$	$1e-4$	$1e-3$	$1e-4$
ArcFace[8]	IQA	BRISQUE[19]	0.0565	0.1285	0.0400	0.0585	0.0343	0.0433	0.0043	0.0049	0.0755	0.0813	0.2558	0.3037	0.6680	0.7122	0.0381	0.0656
		RankIQA[16]	0.0400	0.0933	0.0372	0.0523	0.0301	0.0384	0.0039	0.0045	0.0846	0.0915	0.2437	0.2969	0.6584	0.7039	0.0385	0.0640
		DeepIQA[4]	0.0568	0.1372	0.0403	0.0523	0.0238	0.0292	0.0049	0.0056	0.0793	0.0850	0.2309	0.2856	0.5958	0.6458	0.0383	0.0640
	FIQA	RankIQ[7]	0.0353	0.0873	0.0322	0.0420	0.0152	0.0260	0.0018	0.0024	0.0608	0.0672	0.0633	0.0848	0.2789	0.3332	0.0227	0.0342
		PFE[25]	0.0212	0.0428	0.0172	0.0226	0.0092	0.0129	0.0023	0.0028	0.0647	0.0681	0.0450	0.0638	0.2302	0.2710	0.0176	0.0248
		SER-FIQ[26]	0.0223	0.0434	0.0167	0.0223	0.0065	0.0103	0.0023	0.0028	0.0595	0.0627	0.0389	0.0584	0.1812*	0.2295*	0.0161	0.0241
		FaceQnet[10, 11]	0.0346	0.0734	0.0197	0.0245	0.0240	0.0273	0.0022	0.0027	0.0774	0.0822	0.1504	0.1751	0.5829	0.6136	0.0270	0.0376
		MagFace[18]	0.0207	0.0425	0.0156	0.0198	0.0073	0.0105	0.0016	0.0021	0.0568	0.0602	0.0492	0.0642	0.4022	0.4636	0.0171	0.0254
		SDD-FIQA[21]	0.0248	0.0562	0.0186	0.0206	0.0122	0.0193	0.0021	0.0027	0.0641	0.0698	0.0517	0.0670	0.3090	0.3561	0.0186	0.0270
		CR-FIQA(L) [6]	0.0204	0.0353	0.0159	0.0189	0.0050	0.0082	0.0023	0.0029	0.0616	0.0632	0.0360	0.0515	0.2084	0.2441	0.0138	0.0207
		DifFIQA(R) [2]	0.0232	0.0581	0.0199	0.0265	0.0054	0.0095	0.0025	0.0029	0.0599	0.065	0.0356	0.0522	0.1864	0.2339	0.0135	0.0200
		eDifFIQA(L) [3]	0.0208	0.0402	0.0147	0.0174	0.0045	0.0078	0.0018	0.0022	0.0573	0.0621	0.0342	0.0502	0.1968	0.2459	0.0136	0.0199
		GraFIQs (L) [15]	0.0225	0.0403	0.0176	0.0219	0.0070	0.0111	0.0032	0.0038	0.0644	0.0692	0.0415	0.0612	0.2058	0.2447	0.0162	0.0237
		CLIB-FIQA [22]	0.0217	0.0429	0.0151	0.0178	0.0053	0.0088	0.0016	0.0020	0.0569	0.0615	0.0357	0.0517	0.1881	0.2277	0.0143	0.0209
		ViT-FIQA(T)[1]	0.0197	0.0395	0.0177	0.0207	0.0057	0.0084	0.0023	0.0027	0.0593	0.0627	0.0366	0.0519	0.1864	0.2274	0.0147	0.0216
		ViTNT-FIQA (Ours)	0.0203	0.0392	0.0249	0.0349	0.0063	0.0101	0.0033	0.0038	0.0614	0.0650	0.0428	0.0619	0.2078	0.2473	0.0169	0.0245
ElasticFace[5]	IQA	BRISQUE[19]	0.0644	0.1184	0.0375	0.0403	0.0281	0.0372	0.0034	0.0047	0.0726	0.0747	0.2641	0.4688	0.6343	0.6964	0.0357	0.0622
		RankIQA[16]	0.0433	0.0862	0.0374	0.0436	0.0269	0.0318	0.0033	0.0045	0.0810	0.0835	0.2325	0.4306	0.6189	0.6856	0.0366	0.0590
		DeepIQA[4]	0.0645	0.1203	0.0384	0.0411	0.0191	0.0256	0.0043	0.0056	0.0756	0.0772	0.2401	0.4541	0.5400	0.5832	0.038	0.0599
	FIQA	RankIQ[7]	0.0400	0.0777	0.0309	0.0337	0.0149	0.0180	0.0013	0.0020	0.0598	0.0614	0.0581	0.0727	0.2468	0.2776	0.0226	0.0334
		PFE[25]	0.0222	0.0381	0.0163	0.0172	0.0088	0.0113	0.0018	0.0025	0.0628	0.0643	0.0419	0.0895	0.2112	0.2436	0.0171	0.0247
		SER-FIQ[26]	0.0240	0.0417	0.0163	0.0179	0.0061	0.0085	0.0021	0.0028	0.0574	0.0590	0.0387	0.0513	0.1576*	0.1868*	0.0156	0.0235
		FaceQnet[10, 11]	0.0369	0.0667	0.0194	0.0207	0.0227	0.0247	0.0021	0.0026	0.0763	0.0777	0.1420	0.2880	0.5549	0.5844	0.0263	0.0370
		MagFace[18]	0.0225	0.0385	0.0150	0.0158	0.0069	0.0095	0.0014	0.0021	0.0553	0.0563	0.0474	0.0597	0.3973	0.4282	0.0166	0.0243
		SDD-FIQA[21]	0.0277	0.0512	0.0187	0.0200	0.0098	0.0118	0.0019	0.0027	0.0624	0.0638	0.0493	0.0634	0.3052	0.3562	0.0183	0.0266
		CR-FIQA(L) [6]	0.0214	0.0357	0.0149	0.0159	0.0045	0.0065	0.0018	0.0025	0.0594	0.0608	0.0350	0.0462	0.1798	0.2060	0.0135	0.0203
		DifFIQA(R) [2]	0.0255	0.0499	0.0193	0.0205	0.0049	0.0071	0.0024	0.0029	0.0575	0.0593	0.0323	0.0438	0.1629	0.1944	0.0132	0.0198
		eDifFIQA(L) [3]	0.0219	0.0373	0.0143	0.0152	0.0040	0.0061	0.0017	0.0022	0.0558	0.0574	0.0325	0.0436	0.1731	0.2160	0.0132	0.0197
		GraFIQs (L) [15]	0.0233	0.0394	0.0182	0.0200	0.0070	0.0091	0.0029	0.0037	0.0614	0.0632	0.0393	0.0633	0.1930	0.2319	0.0158	0.0235
		CLIB-FIQA [22]	0.0229	0.0401	0.0152	0.0159	0.0045	0.0069	0.0014	0.0019	0.0562	0.0574	0.0343	0.0454	0.1660	0.2016	0.0139	0.0206
		ViT-FIQA(T)[1]	0.0214	0.0362	0.0169	0.0179	0.0052	0.0073	0.0017	0.0023	0.0575	0.0592	0.0354	0.0461	0.1698	0.2174	0.0142	0.0212
		ViTNT-FIQA (Ours)	0.0218	0.0379	0.0234	0.0249	0.0068	0.0092	0.0030	0.0037	0.0590	0.0607	0.0404	0.0524	0.1814	0.2350	0.0163	0.0242
MagFace[18]	IQA	BRISQUE[19]	0.0594	0.1308	0.0442	0.0799	0.0422	0.0589	0.0043	0.0058	0.0758	0.0788	0.4649	0.6809	0.6911	0.7229	0.0462	0.0787
		RankIQA[16]	0.0407	0.0889	0.0370	0.0681	0.0369	0.0543	0.0041	0.0056	0.0829	0.0857	0.3251	0.6475	0.6706	0.7046	0.0462	0.0750
		DeepIQA[4]	0.0571	0.1302	0.0417	0.0721	0.0322	0.0545	0.0048	0.0059	0.0787	0.0809	0.3672	0.6632	0.6162	0.6519	0.0474	0.0765
	FIQA	RankIQ[7]	0.0359	0.0837	0.0361	0.0531	0.0213	0.0332	0.0019	0.0027	0.0602	0.0629	0.0659	0.1642	0.3076	0.3475	0.0270	0.0383
		PFE[25]	0.0215	0.0423	0.0192	0.0317	0.0107	0.0138	0.0023	0.0029	0.0640	0.0652	0.0449	0.1435	0.2615	0.2926	0.0200	0.0283
		SER-FIQ[26]	0.0233	0.0451	0.0185	0.0293	0.0080	0.0139	0.0025	0.0033	0.0590	0.0607	0.0397	0.0821	0.2139*	0.2562*	0.0189	0.0270
		FaceQnet[10, 11]	0.0365	0.0720	0.0217	0.0314	0.0271	0.0351	0.0022	0.0027	0.0763	0.0773	0.2988	0.5218	0.6016	0.6210	0.0305	0.0423
		MagFace[18]	0.0212	0.0417	0.0159	0.0247	0.0085	0.0129	0.0017	0.0022	0.0562	0.0578	0.0506	0.0887	0.4478	0.4900	0.0195	0.0279
		SDD-FIQA[21]	0.0253	0.0562	0.0216	0.0305	0.0146	0.0201	0.0021	0.0027	0.0643	0.0657	0.0525	0.1188	0.3404	0.3928	0.0215	0.0307
		CR-FIQA(L) [6]	0.0211	0.0372	0.0174	0.0235	0.0062	0.0080	0.0023	0.0028	0.0614	0.0628	0.0374	0.0679	0.2369	0.2839	0.0163	0.0236
		DifFIQA(R) [2]	0.0237	0.0560	0.0218	0.0367	0.0071	0.0158	0.0025	0.0030	0.0600	0.0622	0.0362	0.0838	0.2242	0.2729	0.0161	0.0230
		eDifFIQA(L) [3]	0.0215	0.0412	0.0169	0.0243	0.0058	0.0126	0.0018	0.0023	0.0574	0.0586	0.0358	0.0813	0.2384	0.2800	0.0161	0.0228
		GraFIQs (L) [15]	0.0233	0.0419	0.0182	0.0253	0.0087	0.0186	0.0033	0.0041	0.0640	0.0652	0.0428	0.0987	0.2524	0.3018	0.0191	0.0273
		CLIB-FIQA [22]	0.0225	0.0442	0.0172	0.0255	0.0068	0.0138	0.0016	0.0021	0.0572	0.0582	0.0380	0.0839	0.2234	0.2708	0.0169	0.0239
		ViT-FIQA(C)[1]	0.0197	0.0381	0.0186	0.0295	0.0064	0.0114	0.0024	0.0028	0.0634	0.0648	0.0375	0.0684	0.2187	0.2706	0.0170	0.0244
		ViTNT-FIQA (Ours)	0.0206	0.0398	0.0259	0.0512	0.0092	0.0127	0.0033	0.0038	0.0608	0.0617	0.0428	0.0775	0.2386	0.3156	0.0195	0.0275
CurricularFace[13]	IQA	BRISQUE[19]	0.0502	0.1095	0.0433	0.0491	0.0323	0.0357	0.0041	0.0047	0.0755	0.0784	0.2709	0.5057	0.6146	0.6336	0.0363	0.0589
		RankIQA[16]	0.0359	0.0752	0.0394	0.0510	0.0298	0.0356	0.0039	0.0045	0.0806	0.0865	0.2346	0.4654	0.5900	0.6212	0.0361	0.0556
		DeepIQA[4]	0.0492	0.1070	0.0407	0.0476	0.0227	0.0278	0.0050	0.0056	0.0764	0.0786	0.2488	0.4961	0.5165	0.5526	0.0376	0.0571
	FIQA	RankIQ[7]	0.0314	0.0715	0.0365	0.0417	0.0186	0.0249	0.0018	0.0024	0.0590	0.0640	0.0541	0.0730	0.2449	0.2880	0.0220	0.0320
		PFE[25]	0.0198	0.0365	0.0197	0.0227	0.0100	0.0134	0.0024	0.0028	0.0630	0.0657	0.0402	0.0983	0.1982	0.2220	0.0170	0.0238
		SER-FIQ[26]	0.0211	0.0381	0.0167	0.0193	0.0074	0.0111	0.0025	0.0030	0.0587	0.0610	0.0356	0.0520	0.1558*	0.1866*	0.0153	0.0228
		FaceQNet[10, 11]	0.0326	0.0626	0.0221	0.0267	0.0226	0.0274	0.0022	0.0027	0.0767	0.0799	0.1384	0.3229	0.5035	0.5411	0.0259	0.0354
		MagFace[18]	0.0200	0.0364	0.0167	0.0195	0.0078	0.0111	0.0016	0.0021	0.0563	0.0590	0.0449	0.0607	0.3758	0.4178	0.0163	0.0232
		SDD-FIQA[21]	0.0230	0.0462	0.0219	0.0254	0.0138	0.0185	0.0021	0.0027	0.0637	0.0675	0.0465	0.0671	0.2649	0.3053	0.0178	0.0255
		CR-FIQA(L) [6]	0.0198	0.0336	0.0162	0.0200	0.0054	0.0080	0.0023	0.0029	0.0605	0.0618	0.0324	0.0462	0.1716	0.2318	0.0134	0.0194
		DifFIQA(R) [2]	0.0215	0.0444	0.0223	0.0252	0.0053	0.0089	0.0025	0.0029	0.0575	0.0615	0.0300	0.0436	0.1585	0.1884	0.0130	0.0190
		eDifFIQA(L) [3]	0.0197	0.0337	0.0170	0.0193	0.0044	0.0077	0.0018	0.0022	0.0568	0.0596	0.0303	0.0439	0.1756	0.2084	0.0130	0.0188
		GraFIQs (L) [15]	0.0220	0.0365	0.0167	0.0200	0.0068	0.0099	0.0033	0.0038	0.0610	0.0641	0.0369	0.0663	0.1713	0.1959	0.0156	0.0223
		CLIB-FIQA [22]	0.0207	0.0357	0.0162	0.0191	0.0049	0.0083	0.0016	0.0020	0.0574	0.0596	0.0315	0.0451	0.1600	0.1826	0.0137	0.0198
		ViT-FIQA(C)[1]	0.0187	0.0334	0.0195	0.0227	0.0058	0.0082	0.0023	0.0028	0.0629	0.0650	0.0326	0.0458	0.1641	0.1962	0.0141	0.0205
		ViTNT-FIQA (Ours)	0.0192	0.0348	0.0262	0.0302	0.0078	0.0114	0.0035	0.0040	0.0599	0.0624	0.0372	0.0520	0.1738	0.2147	0.0164	0.0234

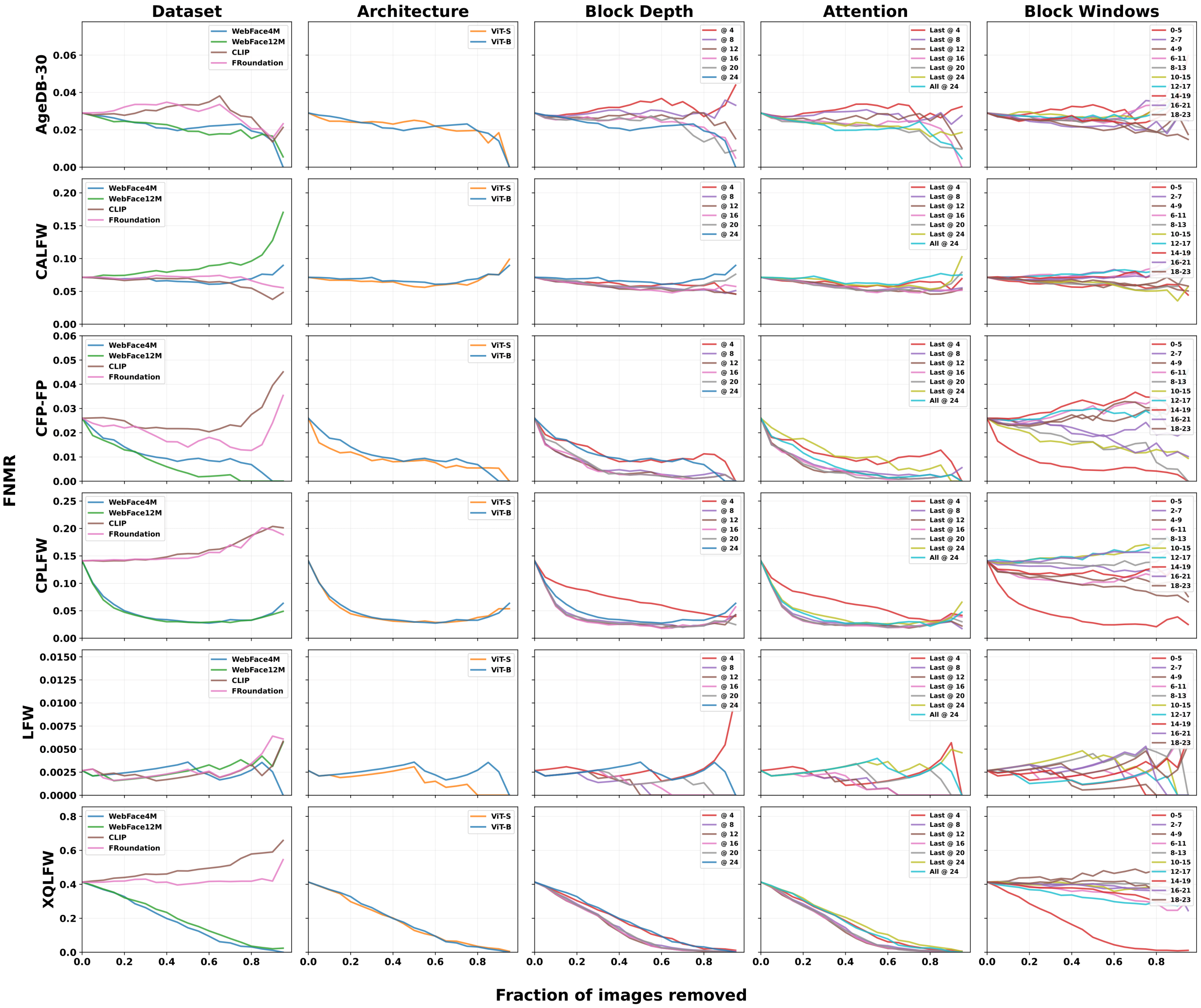

Figure 2. Comprehensive ablation analysis via Error-versus-Discard Characteristic (EDC) curves at FMR=$1e-2$. Each column represents one of five ablation studies: **Dataset** (generalization across WebFace4M, WebFace12M, CLIP, FRoundation), **Architecture** (ViT-S vs ViT-B depth comparison), **Block Depth** (computational trade-offs from 4 to 24 blocks), **Attention** (last-block vs all-blocks aggregation at varying depths), and **Block Windows** (consecutive 6-block segments from early to late network stages). Each row shows results on a different benchmark dataset (AgeDB-30, CALFW, CFP-FP, CPLFW, LFW, XQLFW). The Dataset study confirms cross-model generalization with FR-trained models (WebFace4M, WebFace12M) outperforming foundation models (CLIP, FRoundation). The Architecture study reveals minimal performance gap between ViT-S and ViT-B, validating depth-independence. The Block Depth study demonstrates that 12-20 blocks provide optimal efficiency-performance balance, with diminishing returns beyond 16 blocks. The Attention study shows consistent improvements from attention-weighting, particularly at 12-20 block depths. The Block Windows study reveals that early transformer blocks (0-5) capture the strongest quality signals. All curves use ArcFace for cross-model evaluation. Across all studies, FNMR decreases steadily as low-quality samples are discarded, validating *ViTNT-FIQA*'s effectiveness in identifying quality-degraded images. The consistent color coding highlights method performance: WebFace4M-based configurations (blue) serve as the primary baseline across multiple studies.

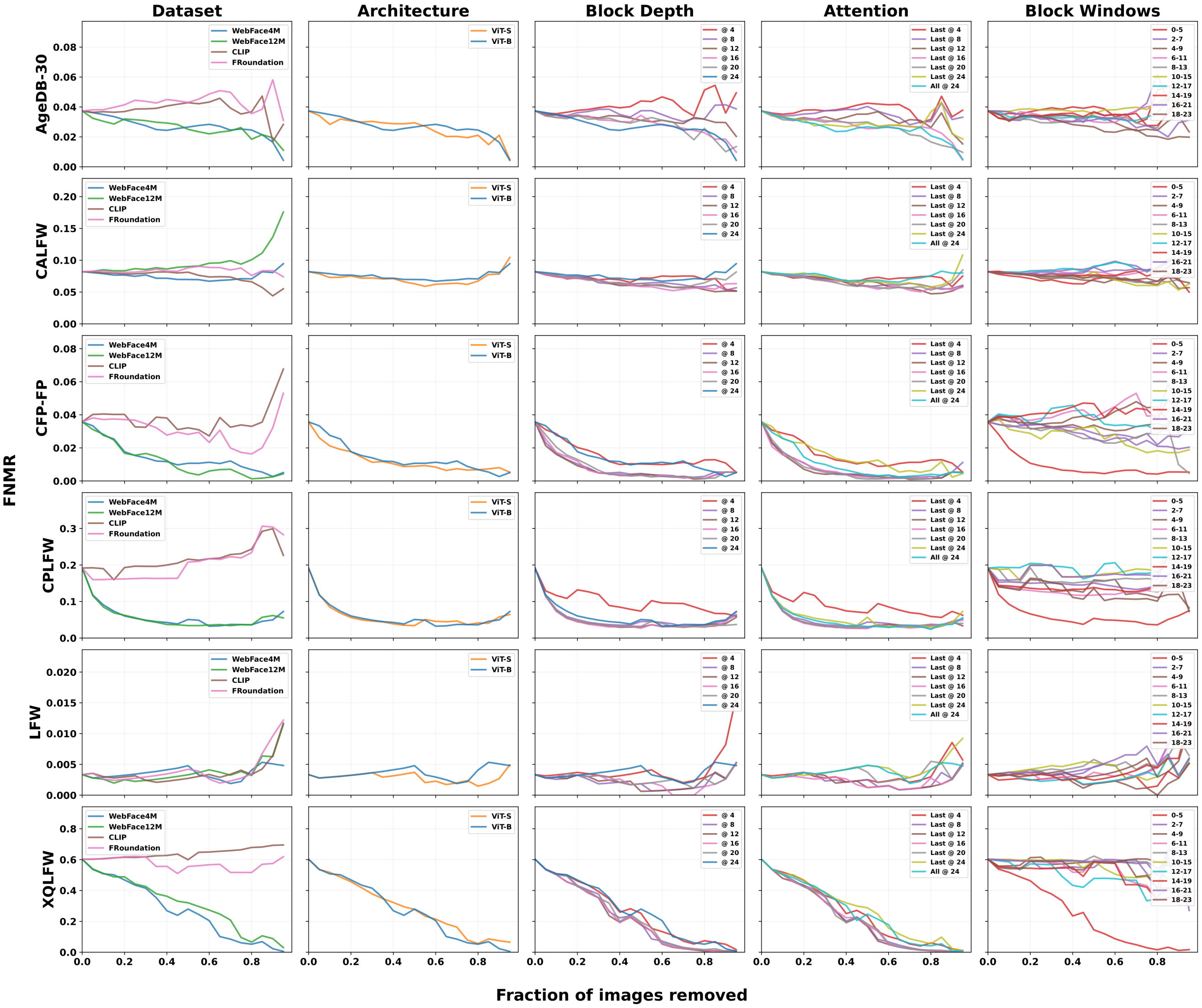

Figure 3. Comprehensive ablation analysis via Error-versus-Discard Characteristic (EDC) curves at FMR=$1e-3$ (Frontex-recommended threshold for border control applications). Layout identical to Figure 2, similar conclusions are also drawn.

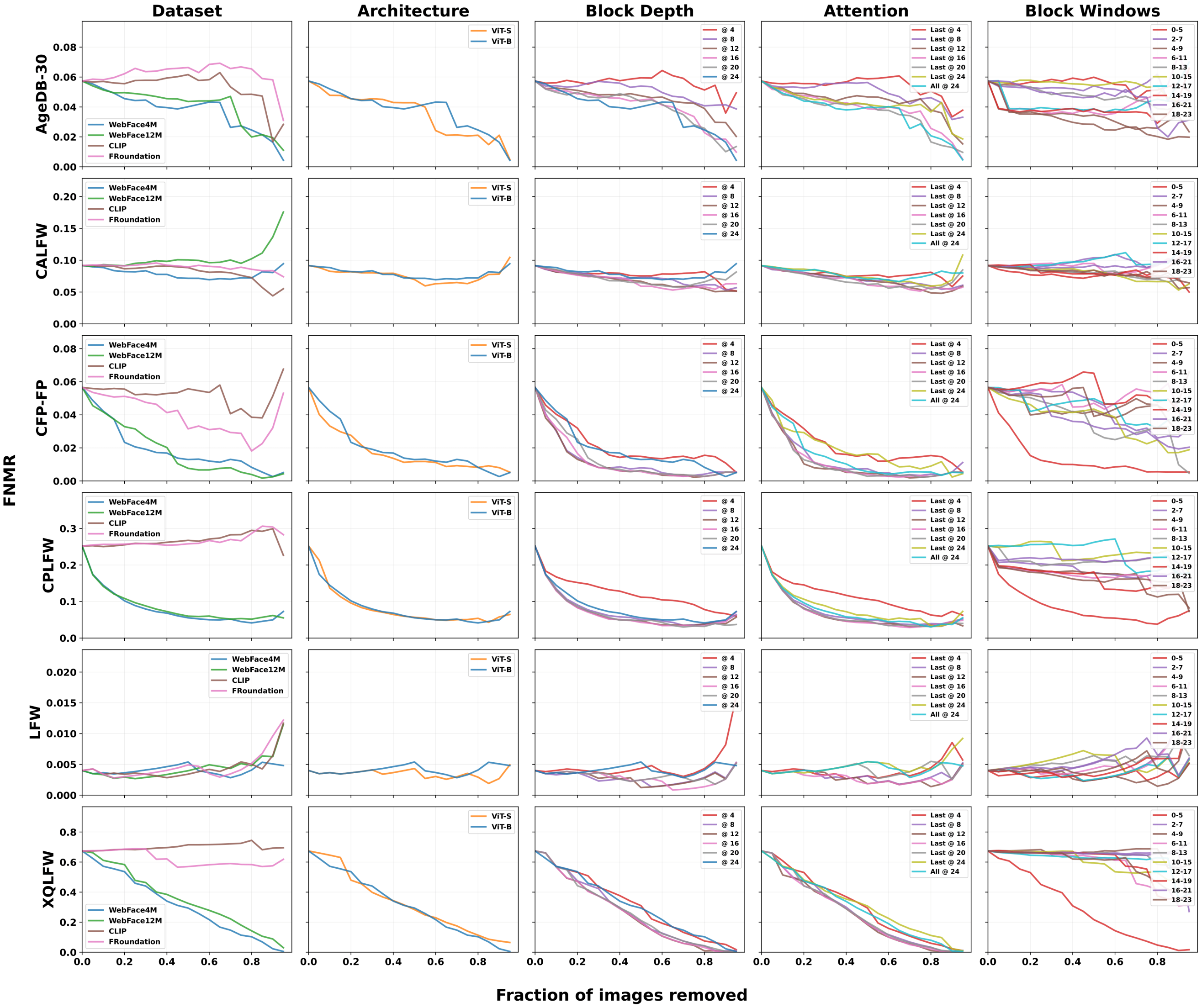

Figure 4. Comprehensive ablation analysis via Error-versus-Discard Characteristic (EDC) curves at FMR=$1e-4$ (high-security operating point). Layout identical to Figures 2 and 3, similar conclusions are also drawn.

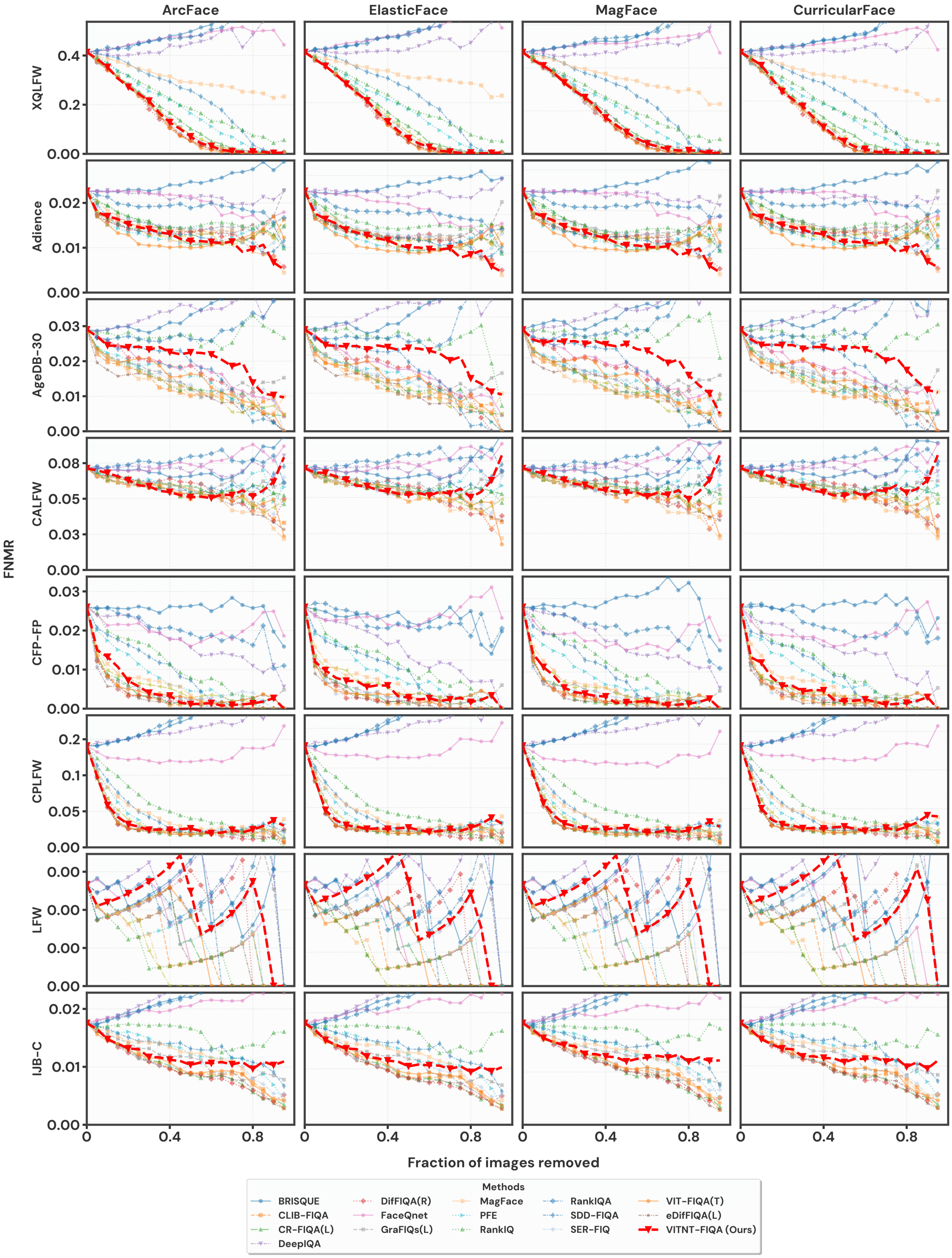

Figure 5. Error-versus-Discard Characteristic (EDC) curves for FNMR@FMR=$1e-2$ of our proposed method in comparison to SOTA. Results shown on eight benchmark datasets: LFW [12], AgeDB-30 [20], CFP-FP [24], CALFW [28], Adience [9], CPLFW [27], XQLFW [14], and IJB-C [17], using ArcFace [8], ElasticFace [5], MagFace [18], and CurricularFace [13] FR models. Our method *ViTNT-FIQA* is marked with the red line.

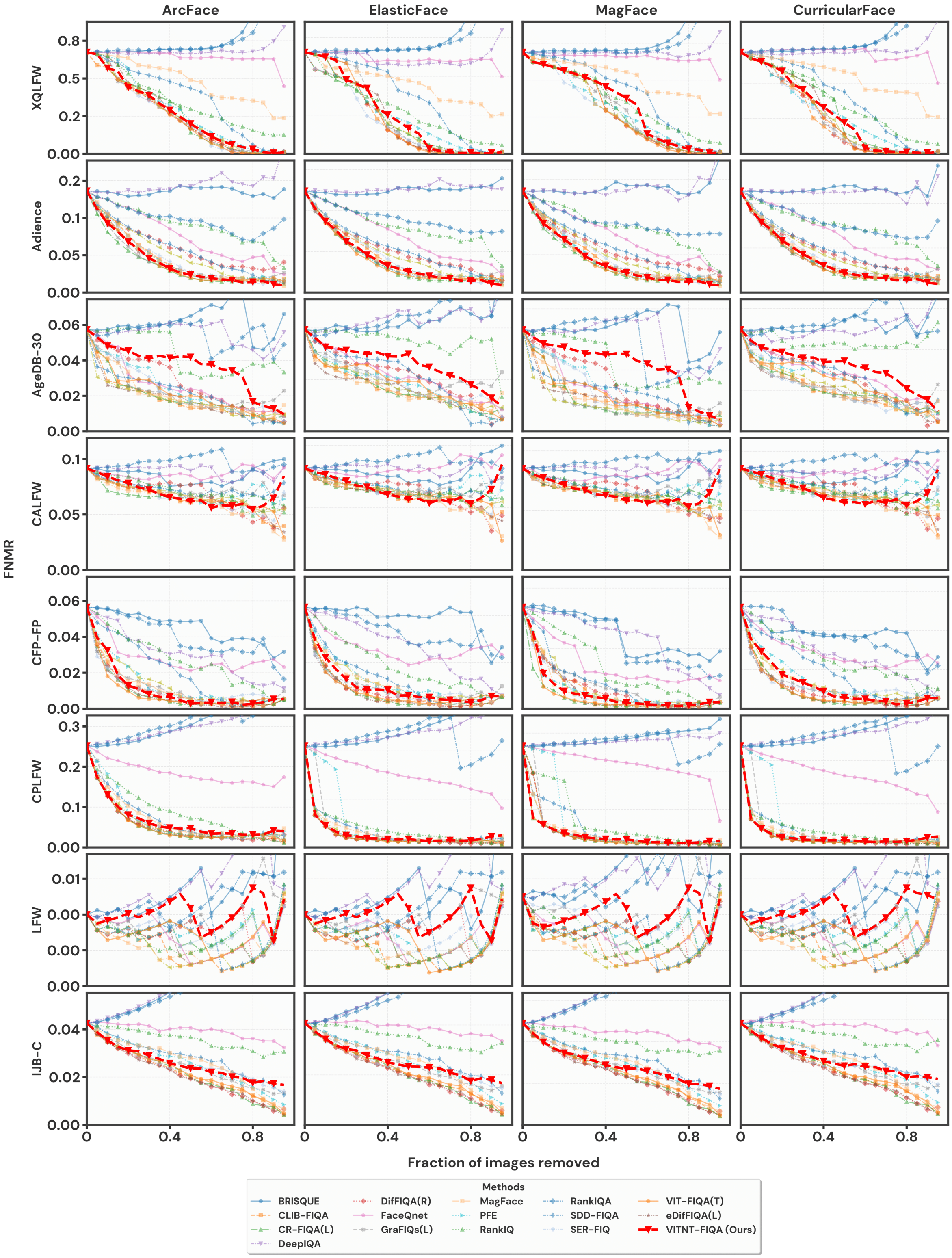

Figure 6. Error-versus-Discard Characteristic (EDC) curves for FNMR@FMR=$1e-4$ of our proposed method in comparison to SOTA. Results shown on eight benchmark datasets: LFW [12], AgeDB-30 [20], CFP-FP [24], CALFW [28], Adience [9], CPLFW [27], XQLFW [14], and IJB-C [17], using ArcFace [8], ElasticFace [5], MagFace [18], and CurricularFace [13] FR models. Our method *ViTNT-FIQA* is marked with the red line.

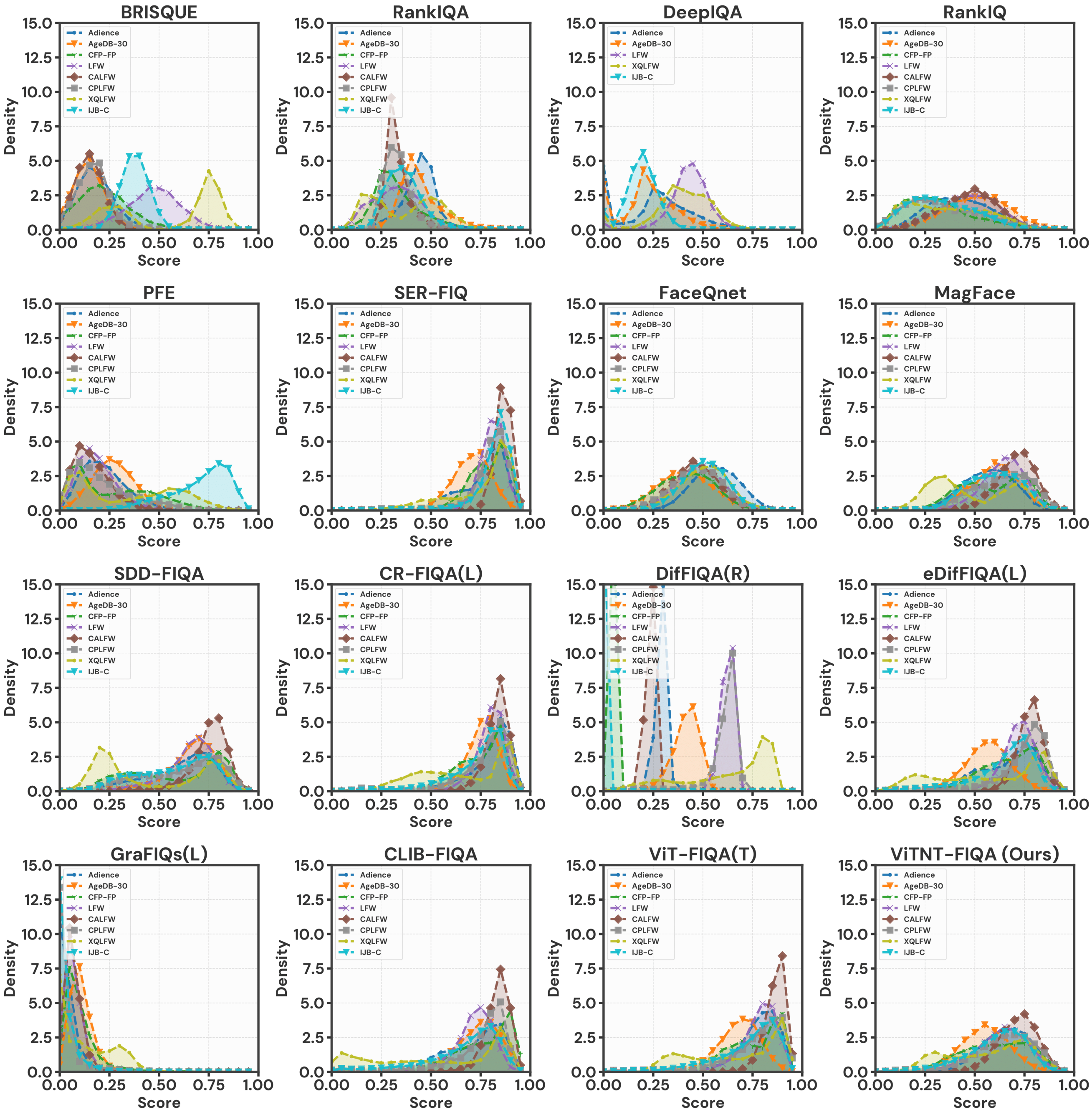

Figure 7. Distribution of quality scores across the evaluation benchmarks, comparing our proposed method (*ViTNT-FIQA*) with SOTA methods. All scores are normalized to the range [0, 1].

References

[1] Andrea Atzori, Fadi Boutros, and Naser Damer. Vit-fiqa: Assessing face image quality using vision transformers. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2025. 4

[2] Žiga Babnik, Peter Peer, and Vitomir Štruc. Diffiqa: Face image quality assessment using denoising diffusion probabilistic models. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10, 2023. 4

[3] Žiga Babnik, Peter Peer, and Vitomir Štruc. eDifFIQA: Towards Efficient Face Image Quality Assessment based on Denoising Diffusion Probabilistic Models. *IEEE Transactions on Biometrics, Behavior, and Identity Science (TBIOM)*, 2024. 4

[4] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Trans. Image Process.*, 27(1):206–219, 2018. 4

[5] Fadi Boutros, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Elasticface: Elastic margin loss for deep face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022*, pages 1577–1586. IEEE, 2022. 4, 8, 9

[6] Fadi Boutros, Meiling Fang, Marcel Klemt, Biying Fu, and Naser Damer. CR-FIQA: face image quality assessment by learning sample relative classifiability. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 5836–5845. IEEE, 2023. 4

[7] Jiansheng Chen, Yu Deng, Gaocheng Bai, and Guangda Su. Face image quality assessment based on learning to rank. *IEEE Signal Process. Lett.*, 22(1):90–94, 2015. 4

[8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4690–4699. Computer Vision Foundation / IEEE, 2019. 4, 8, 9

[9] Eran Eidinger, Roee Enbar, and Tal Hassner. Age and gender estimation of unfiltered faces. *IEEE Trans. Inf. Forensics Secur.*, 9(12):2170–2179, 2014. 3, 4, 8, 9

[10] Javier Hernandez-Ortega, Javier Galbally, Julian Fiérrez, Rudolf Haraksim, and Laurent Beslay. Faceqnet: Quality assessment for face recognition based on deep learning. In *2019 International Conference on Biometrics, ICB 2019, Crete, Greece, June 4-7, 2019*, pages 1–8. IEEE, 2019. 4

[11] Javier Hernandez-Ortega, Javier Galbally, Julian Fiérrez, and Laurent Beslay. Biometric quality: Review and application to face recognition with faceqnet. *CoRR*, abs/2006.03298, 2020. 4

[12] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007. 3, 4, 8, 9

[13] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: Adaptive curriculum learning loss for deep face recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 5900–5909. Computer Vision Foundation / IEEE, 2020. 4, 8, 9

[14] Martin Knoche, Stefan Hörmann, and Gerhard Rigoll. Cross-quality LFW: A database for analyzing cross- resolution image face recognition in unconstrained environments. In *16th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2021, Jodhpur, India, December 15-18, 2021*, pages 1–5. IEEE, 2021. 3, 4, 8, 9

[15] Jan Niklas Kolf, Naser Damer, and Fadi Boutros. Grafiqs: Face image quality assessment using gradient magnitudes. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1490–1499, 2024. 4

[16] Xialei Liu, Joost van de Weijer, and Andrew D. Bagdanov. Rankiqa: Learning from rankings for no-reference image quality assessment. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1040–1049. IEEE Computer Society, 2017. 4

[17] Brianna Maze, Jocelyn C. Adams, James A. Duncan, Nathan D. Kalka, Tim Miller, Charles Otto, Anil K. Jain, W. Tyler Niggel, Janet Anderson, Jordan Cheney, and Patrick Grother. IARPA janus benchmark - C: face dataset and protocol. In *2018 International Conference on Biometrics, ICB 2018, Gold Coast, Australia, February 20-23, 2018*, pages 158–165. IEEE, 2018. 4, 8, 9

[18] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 14225–14234. Computer Vision Foundation / IEEE, 2021. 4, 8, 9

[19] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.*, 21(12):4695–4708, 2012. 4

[20] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: The first manually collected, in-the-wild age database. In *2017 IEEE CVPRW, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1997–2005. IEEE Computer Society, 2017. 3, 4, 8, 9

[21] Fu-Zhao Ou, Xingyu Chen, Ruixin Zhang, Yuge Huang, Shaoxin Li, Jilin Li, Yong Li, Liujuan Cao, and Yuan-Gen Wang. SDD-FIQA: unsupervised face image quality assessment with similarity distribution distance. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 7670–7679. Computer Vision Foundation / IEEE, 2021. 4

[22] Fu-Zhao Ou, Chongyi Li, Shiqi Wang, and Sam Kwong. Clib-fiqa: Face image quality assessment with confidence calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1694–1704, 2024. 4

[23] Fu-Zhao Ou, Chongyi Li, Shiqi Wang, and Sam Kwong. Mr-fiqa: Face image quality assessment with multi-reference representations from synthetic data generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12915–12925, 2025. 2

[24] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Domingo Castillo, Vishal M. Patel, Rama Chellappa, and David W. Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016, Lake Placid, NY, USA, March 7-10, 2016*, pages 1–9. IEEE Computer Society, 2016. 3, 4, 8, 9

[25] Yichun Shi and Anil K. Jain. Probabilistic face embeddings. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6901–6910. IEEE, 2019. 4

[26] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. SER-FIQ: unsupervised estimation of face image quality based on stochastic embedding robustness. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 5650–5659. Computer Vision Foundation / IEEE, 2020. 4

[27] T. Zheng and W. Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. Technical Report 18-01, Beijing University of Posts and Telecommunications, 2018. 3, 4, 8, 9

[28] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments. *CoRR*, abs/1708.08197, 2017. 3, 4, 8, 9