LaTeX Author Guidelines for CVPR Proceedings

Anonymous CVPR submission

Paper ID *****

1. Evaluation Under Varying Evaluation Strategies

Base detection model We use the Capsule network [1] as the base deepfake detector including our MDB. The reason behind is due to its ability to achieve consistently top performance on most datasets.

Competitors We consider two training settings: (1) *Single-domain training:* The deepfake detector is trained using one dataset/domain. We repeat this practice for all six datasets with the base detector as specified in Sec 5.1 (in the main text). (2) *Multi-domain training:* We combine six datasets (FF++, CelebDFv2, UADFV, Deepfakeface, DFDB-Face, JDB-Face) with different deepfake types by simple concatenation and shuffling and train the deepfake detector. This training strategy includes the following methods: We compare the following training methods: *(a) Vanilla:* Training the base detector on the merged 6 datasets. *(b) Knowledge Distillation (KD):* [1] We replace the dynamic difficulty weighting process with a knowledge distillation loss [1] between $\bar{\theta}$ and θ. This comparison aims to evaluate the proposed MDB strategy against a knowledge distillation approach, as both require a separate network $\bar{\theta}$ to be maintained during training. *(c) Difficulty Weighing (DW) [2]:* We use difficulty weighting without momentum i.e. we directly use the in-training network θ to generate the difficulty scores, without referring to the momentum-updated network $\bar{\theta}$. This comparison is intended to evaluate the effectiveness of the proposed MDB strategy. *(d) Our proposed MDB:* We set the momentum $m = 0.97$ and the sample weight rescale factor $C = 5$. For all training strategies, we trained *from scratch* with randomly initialized weights and used the same hyper-parameters with a learning rate of 0.0001, momentum for Adam optimization of 0.9 and the alpha value of 0.99.

Cross-domain test We further evaluate the models trained as above on an unseen domain. We choose the Fake-CelebA [3] as the test dataset for the multi-domain training setting. This dataset has been generated by four diffusion models.

Results From Tables 1, we observe that:

(1) *Single-domain training* on the diffusion deepfakes improved the detector's performance on this new deepfake type. Specifically, the Capsule model's evaluation accuracy was boosted to 0.68/0.73/0.67 from 0.50/0.39/0.39 on Deepfakeface, DiffusionDB-Face, and JourneyDB-Face, when we train it from scratch on each of these datasets. However, we also noticed that this improvement comes with a sacrifice on other datasets. For example, the JourneyDB-Face trained model achieved poor accuracies on all conventional deepfake datasets, with an average value of only 0.39.

(2) Directly training a model with *vanilla* method helped improve deepfake detection performance across all datasets, but only to a limited extent. Specifically, we observe an average accuracy across the six datasets of 0.43 with the multi-domain training, a small increase compared to the individual single-domain trainings (except for FF++).

(3) In comparison with standard *knowledge distillation* (referred as *KD* in Table 1), which achieves an average accuracy of 0.70, the proposed MDB exhibits a 20% improvement. Similar observations can be made when comparing it with the naive difficulty weighting strategy without momentum updating (referred as *DW* in Table 1), which has an average accuracy of 0.59. Such observations show that the proposed MDB's improvement is non-trivial. (4) Our proposed *MDB* led to a substantial performance gain with multi-domain training.

By dynamically assigning sample weights according to their difficulties, it perfectly aligns with the diverse nature of the multi-domain training set and enables the model to focus on more difficult samples along the training. Specifically, we see an average accuracy of 0.76/0.92/0.84 for the proposed MDB approach on conventional/diffusion/all dataset respectively. The corresponding AUC and EER values show much higher ability to distinguish between real and fake images even with the unbalanced datasets. For example, our strategy has 0.94 (AUC) for FF++ (non-diffusion dataset) and 0.93(AUC) for JDB-Face (diffusion dataset). However, the results for UADFV is not up to the mark. This is due to a much smaller training set (1.3k fake images) with UADFV, in comparison to 102k for FF++, 160k for CelebDFv2, and 62k for JDB-Face.

(5) We use Fake-CelebA [3] as totally unseen data for cross-domain generalisation test. The results in Table 2 illustrate superior outcomes by our MDB when applied to a distinct or unfamiliar domain of diffusion-generated images. This validates the advantages of our proposed method compared to the other competitors. We have added more ablative analysis in *supplementary material.*

Table 1. Comparison of generalization capabilities across different datasets and training strategies using the Capsule network as the base deepfake detector. Accuracy (ACC), Equal Error Rate (EER), and Area Under the Curve (AUC) metrics are presented. The best results are in **bold**. The top part of each sub-table shows the single-domain training setting.

(a) Conventional deepfake datasets (FF++, CelebDFv2, UADFV)

Train Strategy	FF++			CelebDFv2			UADFV		
Metric	ACC	EER	AUC	ACC	EER	AUC	ACC	EER	AUC
FF++	0.89	0.23	0.83	0.66	0.45	0.59	0.50	0.50	0.50
CelebDFv2	0.50	0.57	0.49	0.59	0.58	0.48	0.40	0.71	0.33
UADFV	0.50	0.50	0.50	0.33	0.62	0.33	0.49	0.55	0.48
Deepfakeface	0.47	0.44	0.59	0.23	0.77	0.34	0.37	0.28	0.78
DFDB-Face	0.72	0.49	0.52	0.71	0.75	0.20	0.47	0.71	0.25
JDB-face	0.42	0.43	0.55	0.47	0.65	0.35	0.29	0.61	0.35
Vanilla	0.85	0.40	0.67	0.75	0.71	0.31	0.50	0.53	0.40
KD	0.84	0.37	0.71	0.81	0.35	0.65	0.50	0.59	0.48
DW	0.78	0.35	0.72	0.53	0.35	0.72	0.50	0.51	0.48
MDB (ours)	**0.95**	**0.10**	**0.94**	**0.82**	**0.23**	**0.81**	**0.50**	**0.48**	**0.50**

(b) Diffusion deepfake datasets (Deepfakeface, DFDB-Face, JDB-Face)

Train Strategy	Deepfakeface			DFDB-Face			JDB-Face		
Metric	ACC	EER	AUC	ACC	EER	AUC	ACC	EER	AUC
FF++	0.35	0.78	0.27	0.67	0.73	0.29	0.48	0.61	0.35
CelebDFv2	0.25	0.80	0.17	0.49	0.83	0.20	0.23	0.82	0.20
UADFV	0.42	0.48	0.57	0.26	0.77	0.24	0.49	0.71	0.27
Deepfakeface	0.68	0.33	0.57	0.23	0.44	0.58	0.51	0.52	0.51
DFDB-Face	0.73	0.65	0.33	0.73	0.41	0.58	0.57	0.55	0.48
JDB-face	0.25	0.67	0.32	0.47	0.69	0.32	0.67	0.44	0.58
Vanilla	0.38	0.76	0.32	0.43	0.55	0.37	0.51	0.64	0.38
KD	0.72	0.34	0.68	0.76	0.32	0.67	0.57	0.62	0.40
DW	0.57	0.55	0.48	0.63	0.30	0.72	0.58	0.42	0.61
MDB (ours)	**0.79**	**0.20**	**0.78**	**0.98**	**0.07**	**0.94**	**0.98**	**0.07**	**0.93**

2. Dataset Construction Workflow

In this section, we have provided the visualization of the detailed workflow, complemented by a comprehensive visual representation of both misclassified and correctly classified samples encountered throughout the process.

Table 2. Performance metrics comparison

Metric	ACC	EER	AUC
Vanilla	0.57	0.58	0.49
KD [1]	0.67	0.60	0.44
DW [2]	0.54	0.60	0.50
MDB (ours)	**0.80**	**0.21**	**0.78**

2.1. JourneyDB-Face Dataset

Figures 3 through 6 showcase visual examples of both correctly classified and misclassified samples encountered during the process. Figures 3 and 4 illustrate the results of the metadata classification using BERT and the word filtering process, focusing respectively on the Prompt and Style sections. In the Style section, particular attention was given to filtering out images with an anime style. Despite this filtering process, as depicted in these figures, the outcomes did not always align with expectations, particularly in cases where anime style was not explicitly mentioned in the prompts. Instances of such misclassifications, along with their corresponding images, are displayed in Figure 5. Figure 6 demonstrates the results post-face filtering process, successfully isolating the intended images.

2.2. DiffusionDB-Face Dataset

The creation of DiffusionDB-Face involved a bit different approach compared to JourneyDB-Face, adapted to fit the format of the source dataset. As detailed in the main paper, the initial step entailed classifying prompts likely to generate images of human faces using BERT. For instance, Figure 7 displays BERT's classification scores for several samples, providing both human face and not human face evaluations. Despite this, some metadata were inaccurately classified due to specific words or structures in the prompts, as illustrated in Figure 8. To mitigate this, face filtering was employed to exclude irrelevant images. However, as Figure 9 reveals, this method was not foolproof and occasionally included drawings or paintings of human faces. To address this, the Canny edge filter was applied to remove cartoon-styled images. In the main text (Section 3.1), we have addressed the detailed description of the Canny edge detector's threshold. This resulted in more precise outcomes, with some examples of the refined images presented in Figure 10.

3. Ablative Analysis

Sensitivity Analysis of C: (1) We examine the effect of the scale factor, C (Eq 3) in the main text. As observed from Table 3, this parameter is not sensitive with a good range of selections.

3.1. Frequency analysis:

We make visual analysis of frequency distributions across all datasets similar to Zhang *et al.* [4]. This elucidates the distinguishing characteristics between authentic and deepfake imagery. Figure 1 indicates that, the frequency distinction between authentic and synthetic images produced by diffusion models is generally more subtle and thus presents more challenges, compared to that in traditional datasets.

Table 3. Ablation of the scale factor with MDB (Accuracy).

C	FF++	CDFv2	UADFV	DFF	DFDB	JDB
1	0.87	0.78	0.48	0.69	0.73	0.74
3	0.91	0.78	0.49	0.72	0.74	0.74
5	0.95	0.82	0.50	0.79	0.98	0.98
7	0.94	0.79	0.51	0.75	0.81	0.79
9	0.92	0.79	0.49	0.75	0.81	0.81
10	0.91	0.79	0.50	0.73	0.81	0.80

3.2. Sample weight dynamics over training:

Figure 2 presents the per-dataset histogram of weights across training epochs with our MDB.

We note that DiffusionDB-Face and JourneyDB-Face datasets are assigned with highest weights, indicating more challenges presented. This difficulty aware training can benefit the performance (see Table 2).

4. Limitations

Despite the extensive filtering processes applied to the two substantial datasets, JourneyDB and DiffusionDB, there might remain a handful of instances where the images are either overly cartoonized or lack sufficient realism. These anomalies may be overlooked in subsequent stages, such as the further Face Filtering and the custom model designed for animated or human facial images. As highlighted in the primary paper concerning dataset statistics, there is a significant gender distribution disparity, originating from the source databases (likely due to the processes of both prompting and training the generative models).

References

[1] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *IJCV*, 129: 1789–1819, 2021. 1, 2

[2] Tomoumi Takase. Difficulty-weighted learning: A novel curriculum-like approach based on difficult examples for neural network training. *Expert Systems with Applications*, 135: 83–89, 2019. 1, 2

[3] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. *arXiv preprint arXiv:2303.09295*, 2023. 1, 2

[4] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *2019 IEEE international workshop on information forensics and security (WIFS)*, pages 1–6, 2019. 2

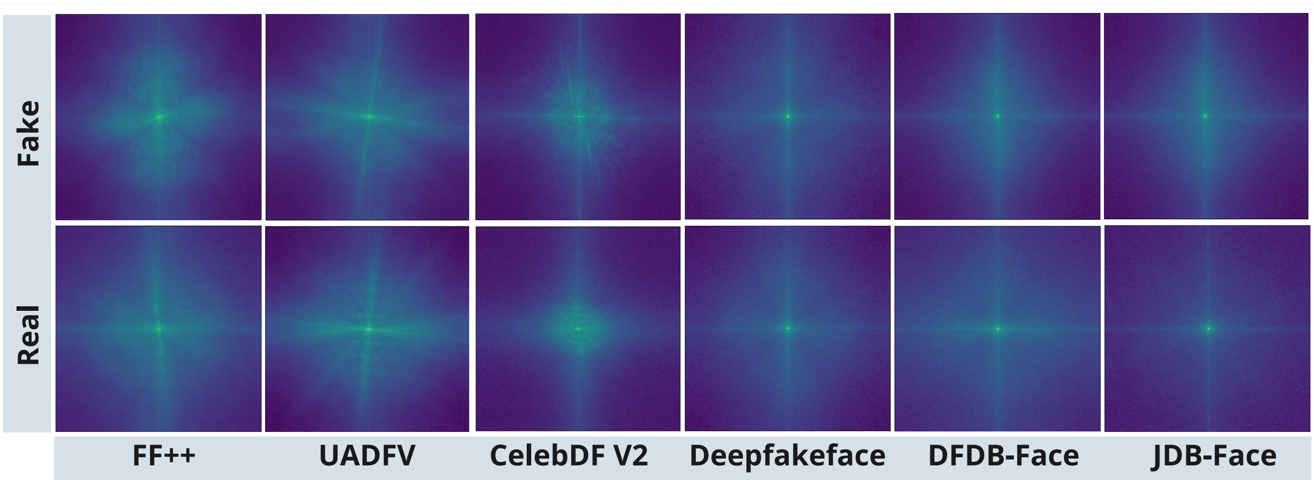


Figure 1. Frequency analysis: the average spectra of each high-pass filtered image

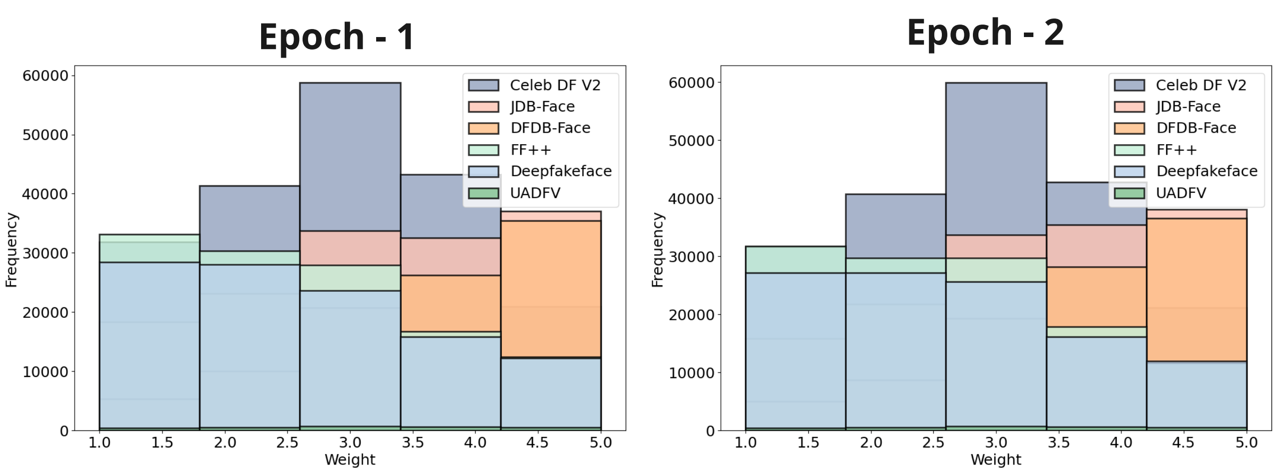


Figure 2. Ablative Analysis: (a) Frequency analysis and (b) weight distribution and dynamics .

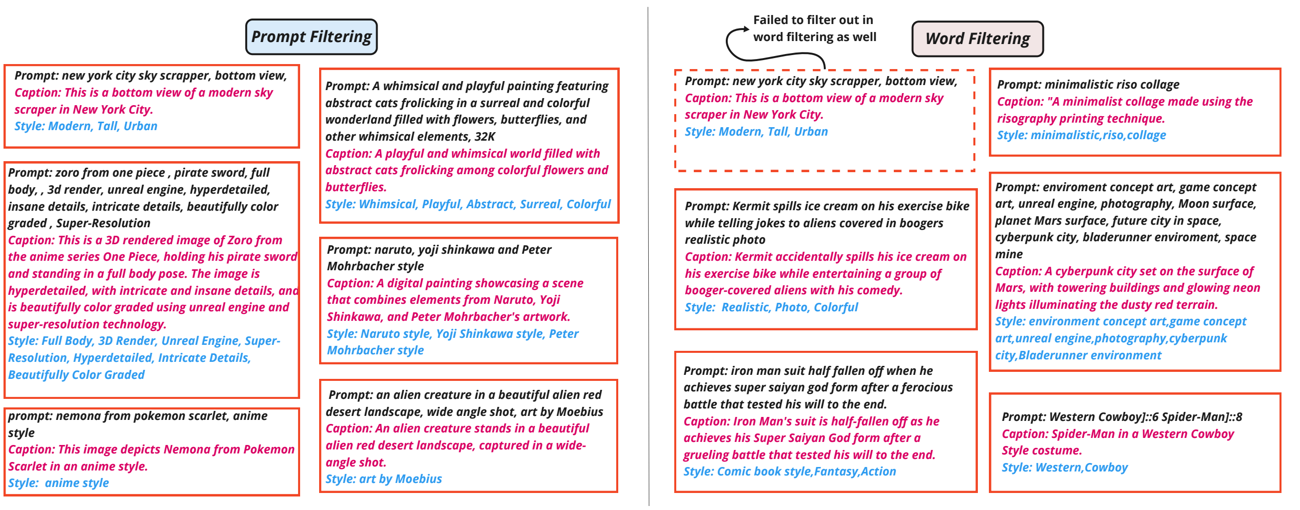


Figure 3. JourneyDB-Face: Examples of misclassified metadata by BERT.

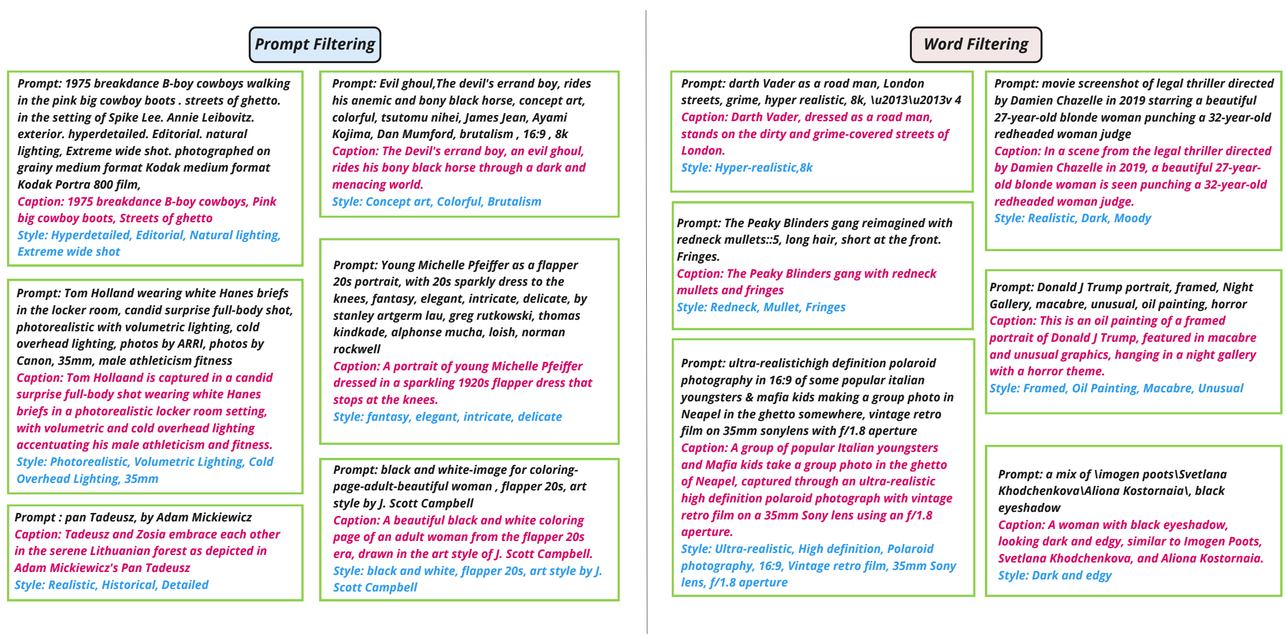


Figure 4. JourneyDB-Face: Examples of correctly classified metadata by BERT.

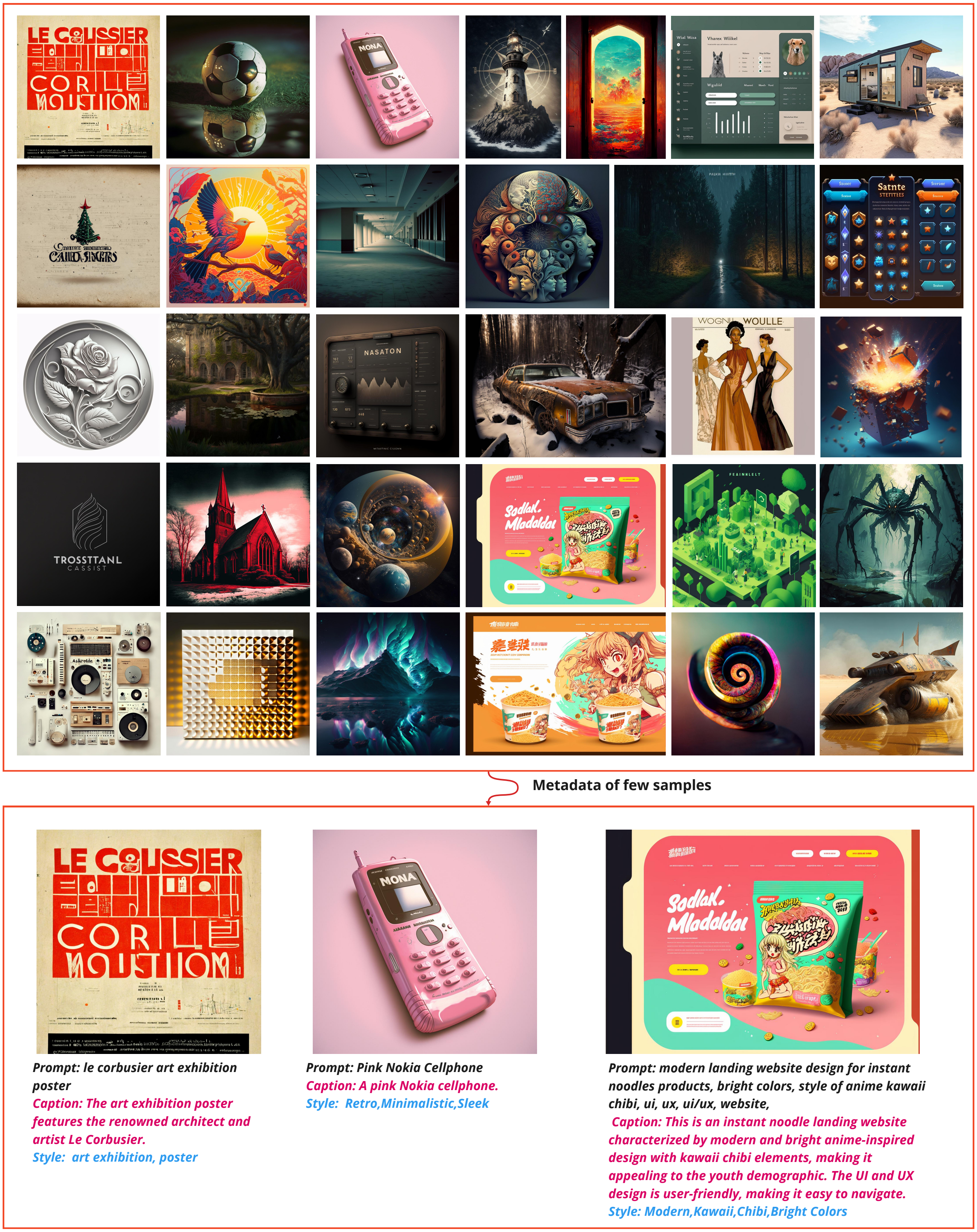
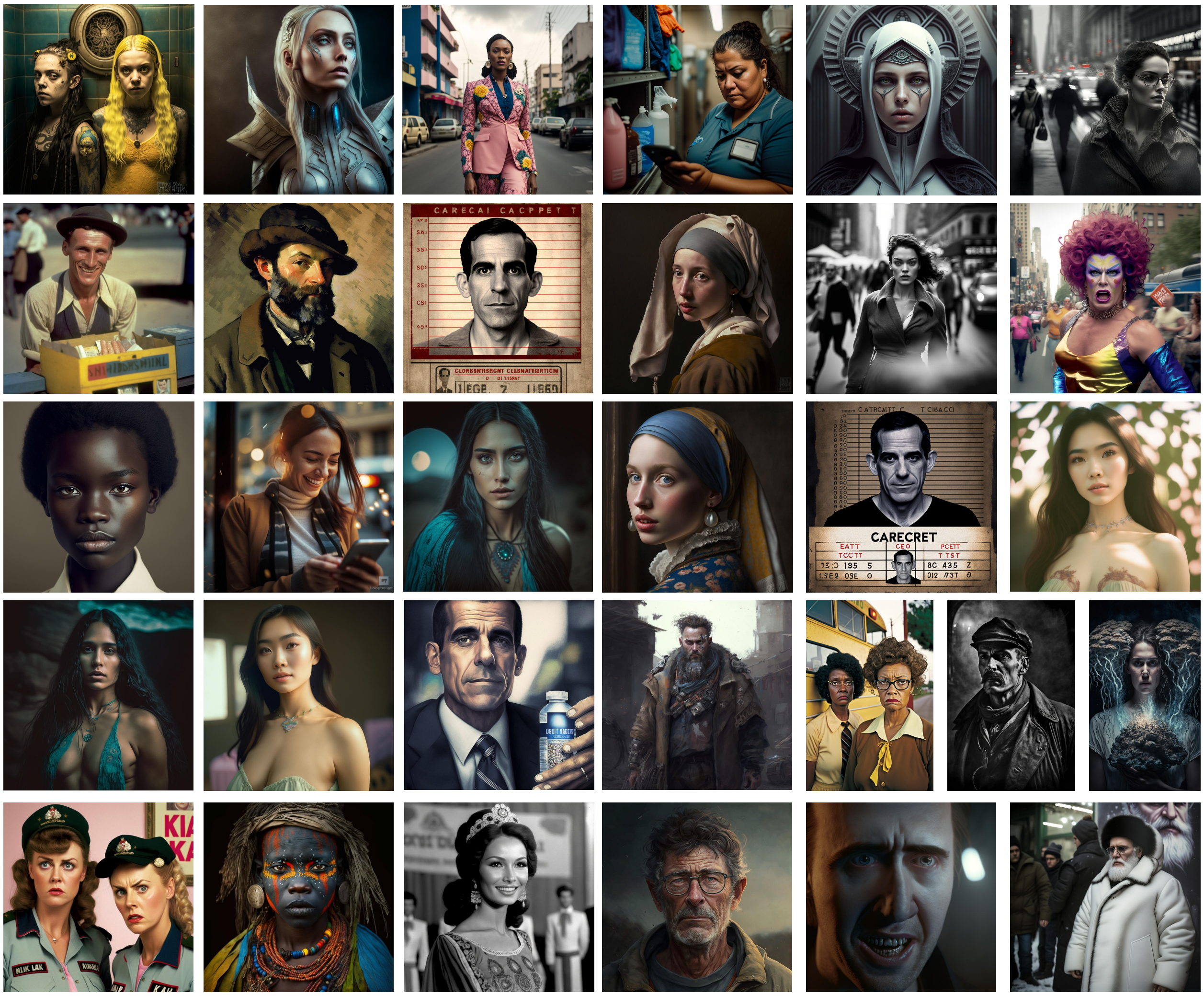
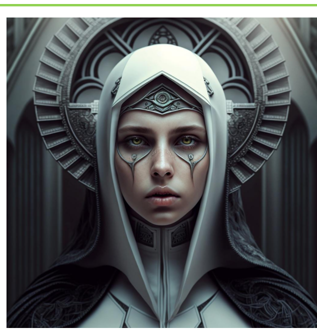


Figure 5. JourneyDB-Face: Unfiltered samples in word filtering due to the absence of Anime Style mention.

CVPR
#*****



Metadata of few samples



Prompt: vampire goth e-girl dressed as a sad nun and a look designed by H.R. Giger and ghostmane, award winning image, 50mm, perfect social media image
Caption: A sad nun, dressed in a style inspired by vampire goth e-girl fashion, is depicted in this award-winning image with a touch of H.R. Giger and Ghostmane influence.
Style: vampire, goth, e-girl, H.R. Giger, Ghostmane

Prompt: Rick from Curse of Oak Island
Caption: Rick from Curse of Oak Island.
Style: Realistic, Mysterious

Prompt: supreme leader ajatollah khomeini in a long white feather nylon puffer coat, nylon, photorealistic, press photograph, taken outside talking to public, detailed, depth
Caption: Supreme Leader Ajatollah Khomeini ...the scene.
Style: photorealistic, press photograph

Figure 6. JourneyDB-Face: Examples of correctly filtered samples after word filtering process.

CVPR
#*****

Prompt: doom eternal, game concept art, veins and worms, muscular, crustacean exoskeleton, chiroptera head, chiroptera ears, mecha, ferocious, fierce, hyperrealism, fine details, artstation, cgsociety, zbrush, no background [B : 0.57]

Prompt: a beautiful photorealistic painting of cemetery urbex unfinished building building industrial architecture nature abandoned by thomas cole, nature extraterrestial tron forest darkacademia thermal vision futuristic tokyo, archdaily, wallpaper, highly detailed, trending on artstation [B : 0.59]

Prompt: beautiful garden at twilight by nicholas roerich and jean delville and maxfield parrish, glowing paper lanterns, strong dramatic cinematic lighting, ornate tiled architecture, lost civilizations, smooth, sharp focus, extremely detailed [B : 0.61]

Prompt: symmetry!! a tiny cute chinese spring festival oriental tale mascot cat - lion toys, magic, intricate, smooth line, light dust, mysterious dark background, warm top light, hd, 8 k, smooth \uff0c sharp high quality artwork in style of greg rutkowski, concept art, blizzard warcraft artwork, bright colors [B : 0.55]

Prompt: film still cinematic photo by 3 4 3 industries, matte painting [B : 0.55]

Prompt: a beautiful very detailed rendering of urbex unfinished building industrial architecture kingdom architecture nature by georges seurat, tundra retrowave sunset myst landscape hyperrealism tokyo rainforest bladerunner 2 0 4 9 lightpaint uv light infrared flowers morning sun nature at dawn, archdaily, wallpaper, highly detailed, trending on artstation. [B : 0.51]

Prompt: a man with an shrivelled up walnut brain inside his skull [A : 0.59]

Prompt: a sinister walnut man [A : 0.64]

Prompt: studio ghibli anime, adorable woman sitting at a cat cafe with a drink, romantic magical, fairytale, fantasy [A : 0.60]

Prompt: painted closeup portrait of intense woman, fierce, charming, fantasy, intricate, elegant, extremely detailed by by chuck close, charcoal on canvas [A : 0.60]

Prompt: painted closeup portrait of fierce, elegant woman. extremely detailed by chuck close, charcoal on canvas [A : 0.60]

Prompt: a boy holding on to a dying old dog connecting him to his childhood [A : 0.60]

Prompt: victo ngai girl succubi sticker decal design, highly detailed, high quality, digital painting, by ross tran and studio ghibli and alphonse mucha, artgerm [A : 0.60]

Figure 7. DiffusionDB-Face : Examples of BERT classified metadata with the corresponding scores. A: ”human face” ; B: ”not human face”.

CVPR #*****

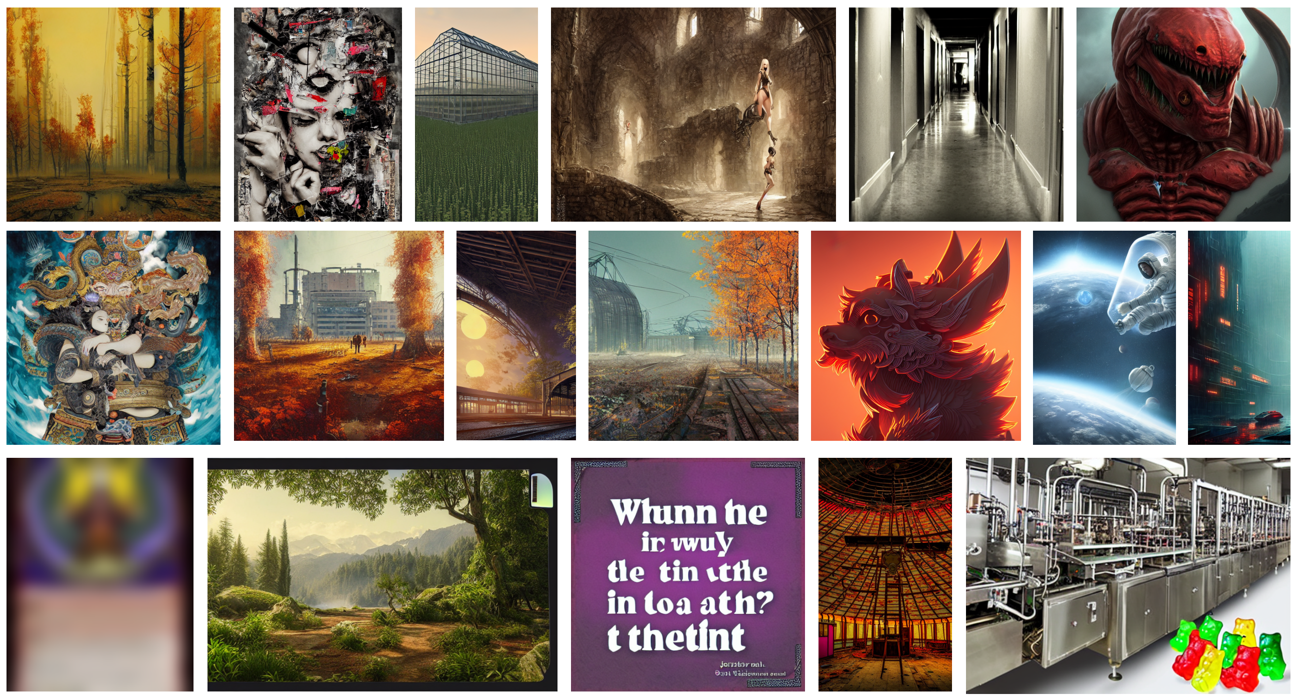


Figure 8. Misclassified samples by BERT's metadata classification round.

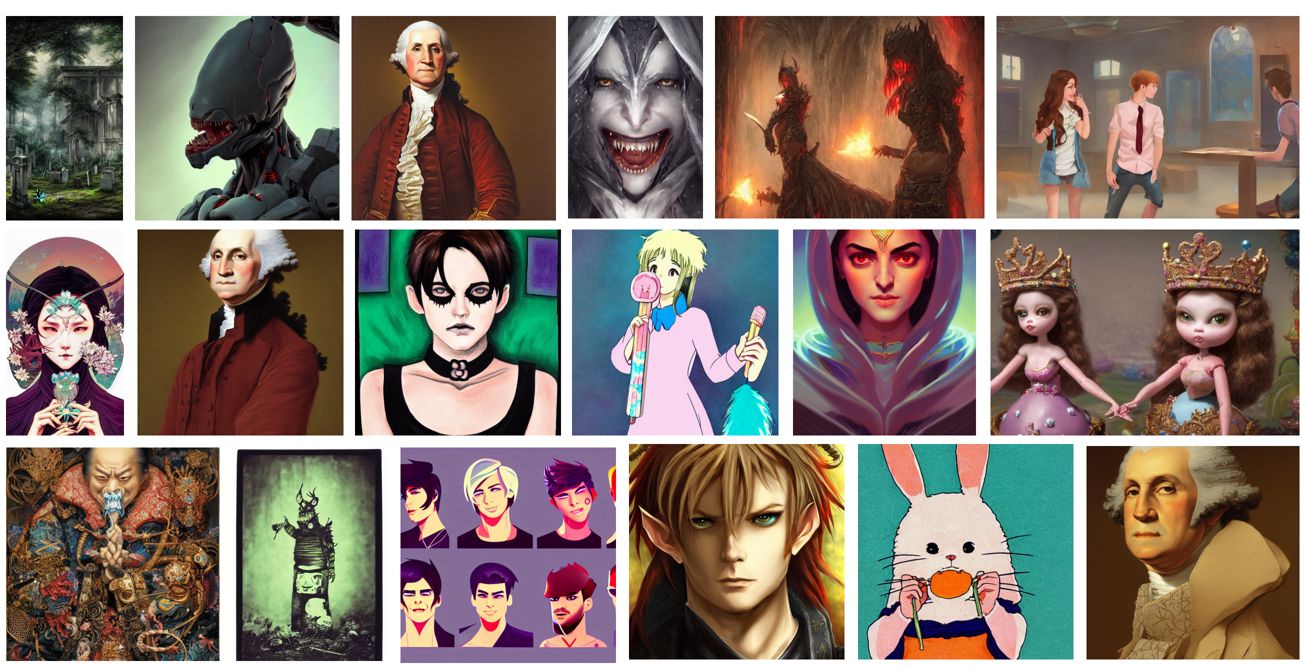


Figure 9. DiffusionDB-Face: Animated face image samples after face filtering.

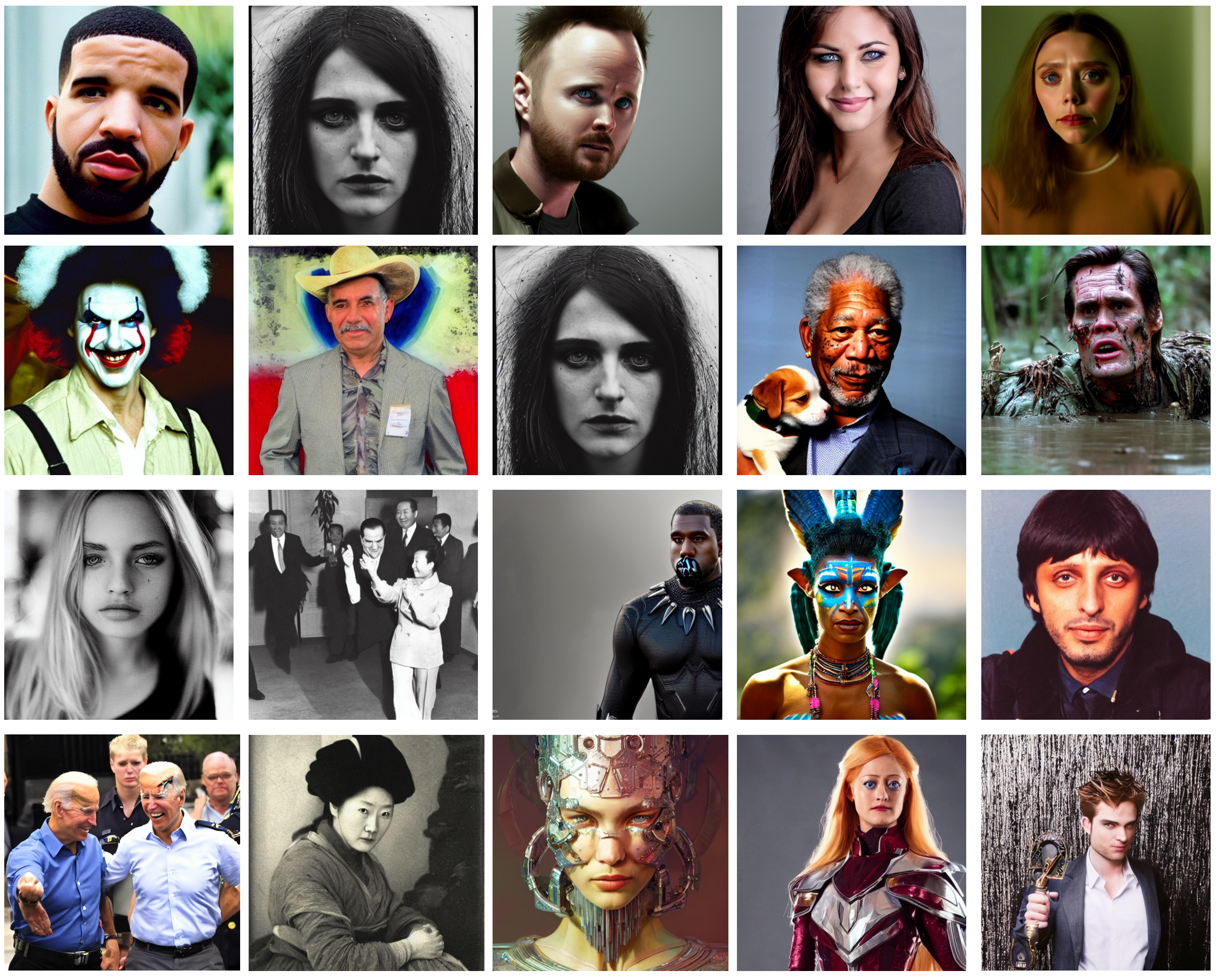


Figure 10. DiffusionDB-Face: Few samples after applying Canny edge detector.