Aggregating Diverse Cue Experts for AI-Generated Image Detection

Lei Tan[1], Shuwei Li[1], Mohan Kankanhalli[1], Robby T. Tan[1]
[1]National University of Singapore

Motivation

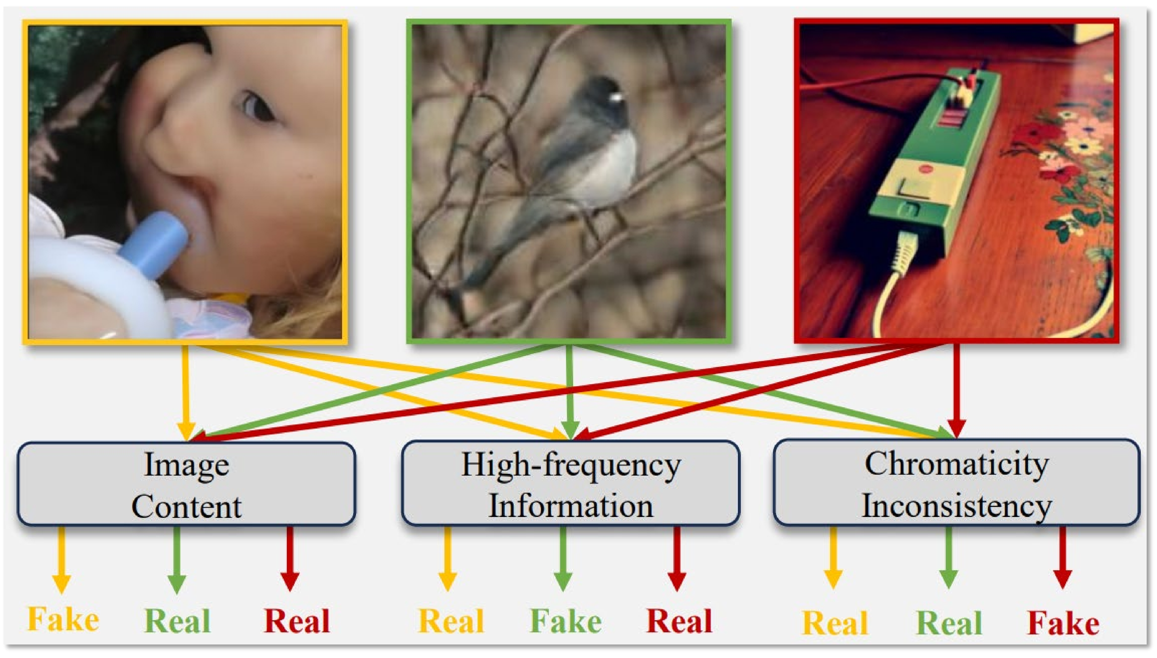

Figure 1: **Motivation for MCAN**

- Different cues exhibit complementary properties in detecting synthetic content.
- A crucial yet often overlooked factor is that real images exhibit relatively minor quality variations, whereas generated images undergo more significant shifts due to discrepancies in the training data and the structure of generative models.

Contributions

Our key contributions are as follows:

- **MCAN**: A novel framework introducing a multi-cue aggregation strategy to enhance AI-generated image detection. MCAN dynamically integrates spatial, frequency, and chromaticity-based cues, unlike existing methods.
- **Chromaticity Inconsistency (CI)**: A new representation that mitigates lighting intensity effects through chromaticity-based transformation, highlighting noise differences between real and generated images. CI complements both image content and high-frequency features in MCAN.

Method

❐ Multi-cue Aggregation Network

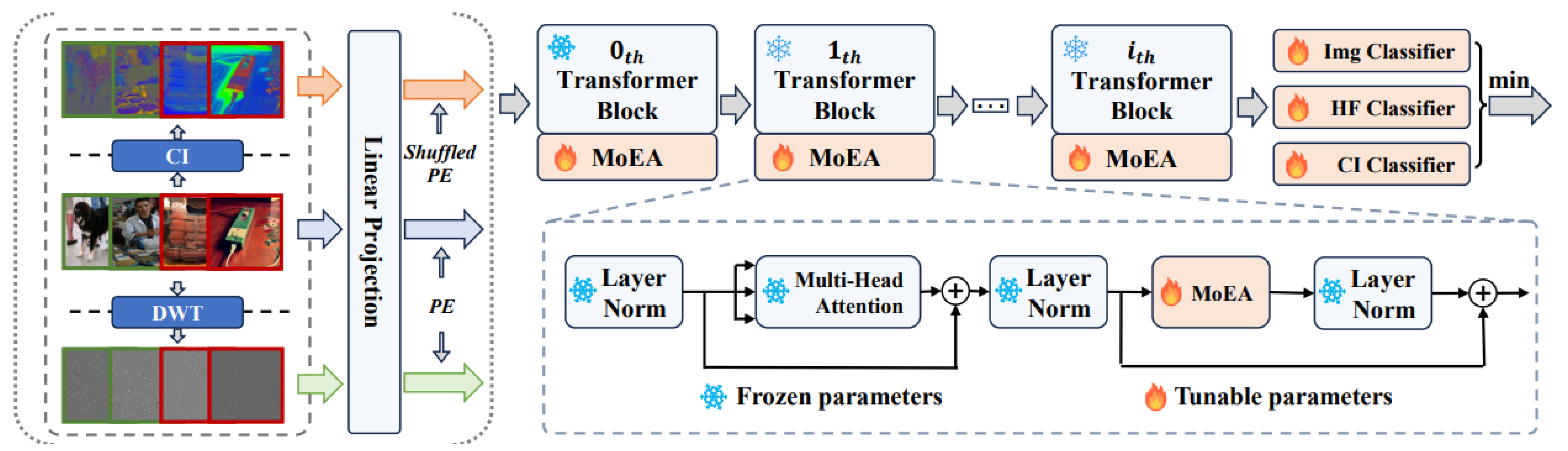

Figure 2: **Overall architecture:** MCAN combines image representation, high-frequency representation, and the novel chromaticity inconsistency as three distinct cues. To effectively integrate these cues, MCAN uses a mixture of encoder adapters that adapt efficiently to each cue's representation.

❐ Chromaticity Inconsistency

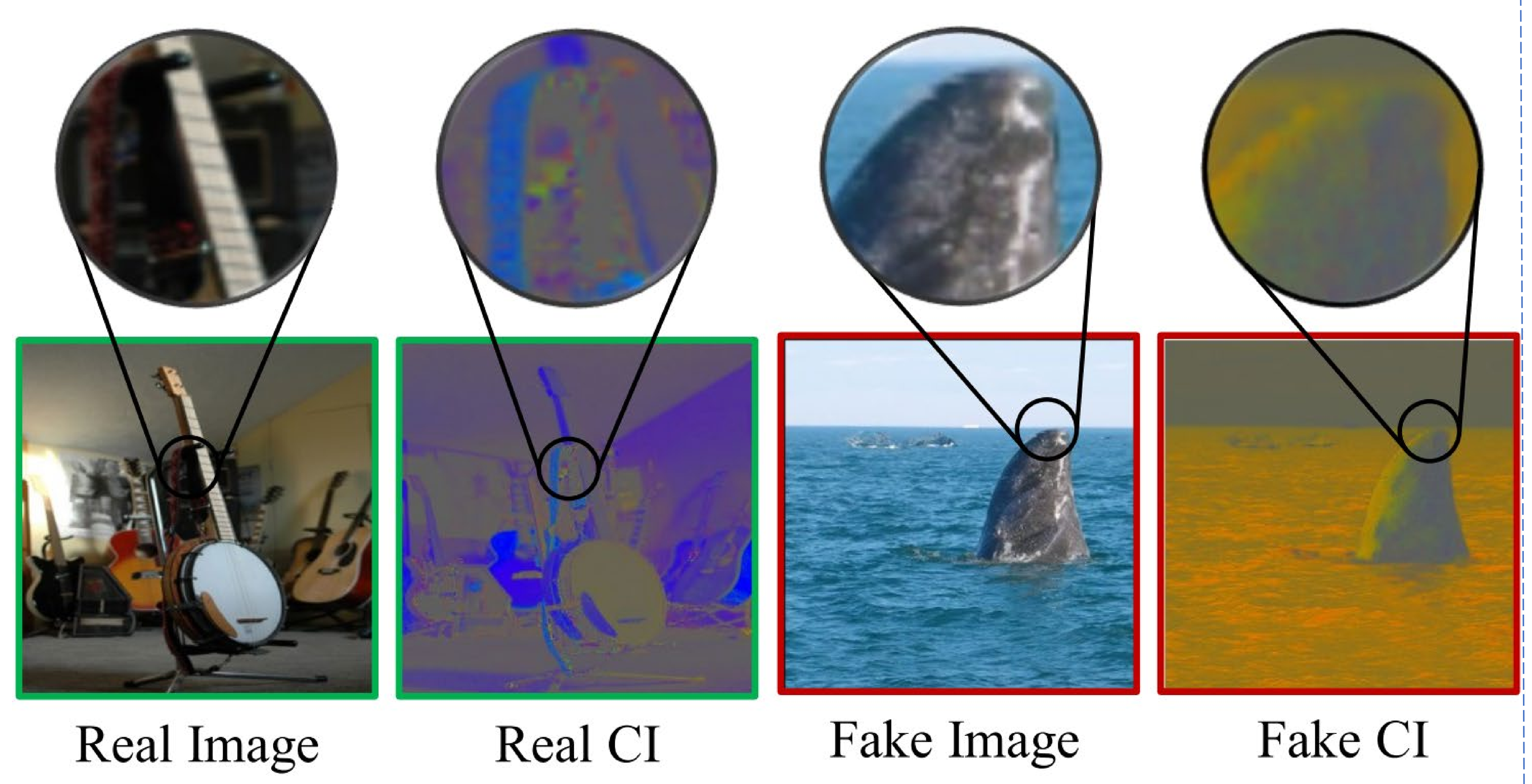

Figure 3: **Visualization of Chromaticity Inconsistency (CI) transformation.**

- CI employs chromaticity transformation to minimize the influence of illumination, highlighting noise differences between real and generated images

❐ Mixture-of-encoder Adapter

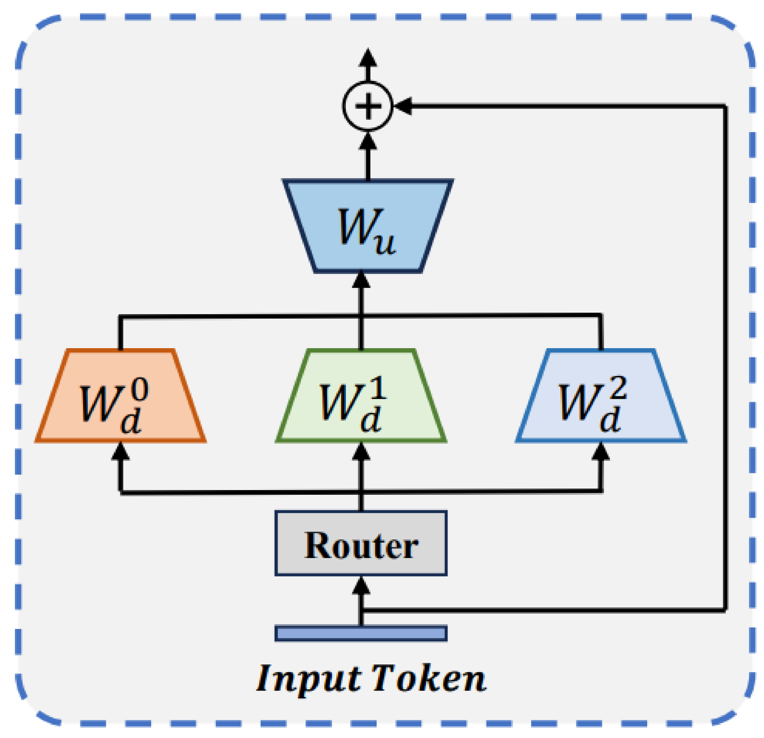

Figure 4: **Overall structure of MoEA.**

- To effectively integrate these cues, MCAN uses a mixture of encoder adapters that adapt efficiently to each cue's representation.

Experiments

Method	Venue	Testing Subset								Avg
		Midjourney	SDV1.4	SDV1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	Accuracy(%)
CNNSpot[44]	CVPR'20	58.2	70.3	70.2	57.0	57.1	67.7	56.7	56.6	61.7
Spec[53]	WIFS'19	56.7	72.4	72.3	57.9	65.4	70.3	61.7	64.3	65.1
F3Net[32]	ECCV'20	55.1	73.1	73.1	66.5	57.8	72.3	62.1	56.5	64.6
GramNet[20]	CVPR'20	58.1	72.8	72.7	58.7	65.3	71.3	57.8	61.2	64.7
DIRE[46]	ICCV'23	65.0	73.7	73.7	61.9	69.1	74.3	63.4	56.7	67.2
LaRE^2[21]	CVPR'24	66.4	87.3	87.1	66.7	81.3	85.5	84.4	74.0	79.1
LaRE^2-ViT†[21]	CVPR'24	72.5	76.1	76.0	67.9	80.6	73.7	69.0	70.5	73.3
FatFormer†[18]	CVPR'24	76.8	78.3	78.6	78.0	93.1	76.1	79.0	68.7	78.5
MCAN	-	**83.3**	**88.0**	**88.2**	**85.9**	**96.9**	**87.0**	**90.2**	**97.7**	**89.7**

Table 1: **Performance on GenImage datasets.**

Setting	Testing Subset								Avg
	Midjourney	SDV1.4	SDV1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	ACC.(%)
Img only	88.1	**99.6**	**99.5**	75.6	93.2	98.7	89.8	51.4	87.0
HF only	90.2	96.5	96.5	83.5	95.6	96.2	93.6	96.7	93.6
CI only	87.5	96.9	96.7	81.7	85.2	96.7	94.1	51.7	86.3
CI-Shuffled Only	89.1	97.2	97.0	86.7	86.5	96.9	95.0	70.0	89.8
Naïve Combination(Img, HF, CI-Shuffled)	92.8	97.6	97.4	85.6	96.9	97.2	94.5	97.5	95.9
MCAN-(Img, HF)	93.7	97.4	96.9	88.6	96.4	97.0	94.8	98.2	95.4
MCAN(Img, CI-Shuffled)	93.3	98.6	98.5	88.8	97.5	98.4	96.5	69.5	92.6
MCAN(HF, CI-Shuffled)	94.1	98.2	98.3	89.8	95.5	97.3	96.3	98.8	96.0
MCAN-(Img, HF, CI-Shuffled)	**94.3**	98.8	98.5	**90.2**	**98.6**	**98.8**	**96.8**	**98.8**	**96.9**

Table 2: **Ablation study of the different components in MCAN.**

Methods	GAN						Deep fakes	Low level		Perceptual loss		Guided	LDM			Glide			Dalle	mAcc
	Pro-GAN	Cycle-GAN	Big-GAN	Style-GAN	Gau-GAN	Star-GAN		SITD	SAN	CRN	IMLE		200 steps	200 w/cfg	100 steps	100 27	50 27	100 10		
CNNSpot [36]	100.0	85.2	70.2	85.7	79.0	91.7	53.5	66.7	48.7	86.3	86.3	60.1	54.0	55.0	54.1	60.8	63.8	65.7	55.6	69.6
Patchfor [7]	75.0	69.0	68.5	79.2	64.2	63.9	75.5	75.1	75.3	72.3	55.3	67.4	76.5	76.1	75.8	74.8	73.3	68.5	67.9	71.2
Co-occurence [23]	97.7	63.2	53.8	92.5	51.1	54.7	57.1	63.1	55.9	65.7	65.8	60.5	70.7	70.6	71.0	70.3	69.6	69.9	67.6	66.9
Freq-spec [40]	49.9	99.9	50.5	49.9	50.3	99.7	50.1	50.0	48.0	50.6	50.1	50.9	50.4	50.4	50.3	51.7	51.4	50.4	50.0	55.5
F3Net [27]	99.4	76.4	65.3	92.6	58.1	100.0	63.5	54.2	47.3	51.5	51.5	69.2	68.2	75.4	68.8	81.7	83.3	83.1	66.3	71.3
UniFD [25]	100.0	98.5	94.5	82.0	99.5	97.0	66.6	63.0	57.5	59.5	72.0	70.0	94.2	73.8	94.4	79.1	79.9	78.1	86.8	81.4
LGrad [33]	99.8	85.4	82.9	94.8	72.5	99.6	58.0	62.5	50.0	50.7	50.8	77.5	94.2	95.9	94.8	87.4	90.7	89.6	88.4	80.3
FreqNet [34]	97.9	95.8	90.5	97.6	90.2	93.4	97.4	88.9	59.0	71.9	67.4	86.7	84.6	99.6	65.6	85.7	97.4	88.2	59.1	85.1
NPR [35]	99.8	95.0	87.6	96.2	86.6	99.8	76.9	66.9	98.6	50.0	50.0	84.6	97.7	98.0	98.2	96.3	97.2	97.4	87.2	87.6
FatFormer [21]	99.9	99.3	99.5	97.2	99.4	99.8	93.2	81.1	68.0	69.5	69.5	76.0	98.6	94.9	98.7	94.4	94.7	94.2	98.8	90.9
MCAN	100.0	99.6	98.8	97.0	99.3	100.0	94.0	86.7	68.9	87.3	87.3	70.9	98.8	94.2	98.5	97.4	97.2	97.1	98.8	93.3

Table 3: **Performance on UniversalFakeDetect dataset.**

Training Setting	Methods										
	CNNSpot [37]	FreDect [13]	Fusing [15]	GramNet [22]	LNP [20]	UniFD [25]	DIRE [38]	Patchcraft [41]	NPR [35]	AIDE [39]	MCAN
ProGAN	56.94	55.62	56.98	58.94	57.11	57.22	58.19	53.76	57.29	58.37	**60.81**
SDV1.4	60.11	56.86	57.07	60.95	55.63	55.62	59.71	56.32	58.13	62.60	**69.61**

Table 4: **Performance on Chamelone datasets.**

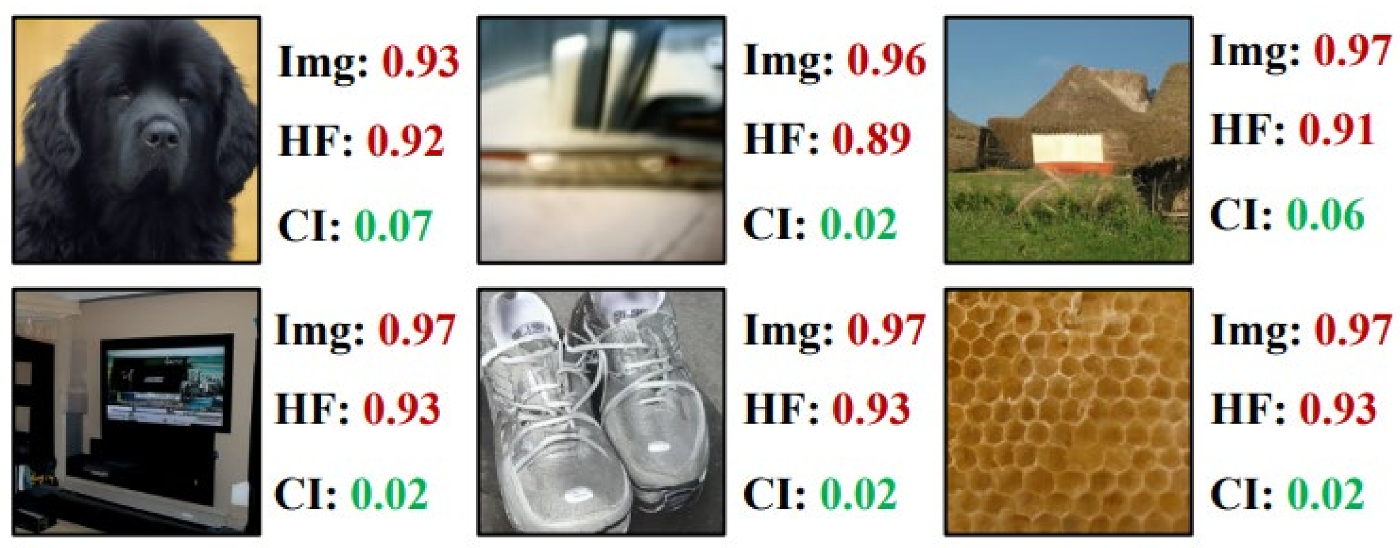

Figure 5: **Visualization of classification results for generated images under different cues.**

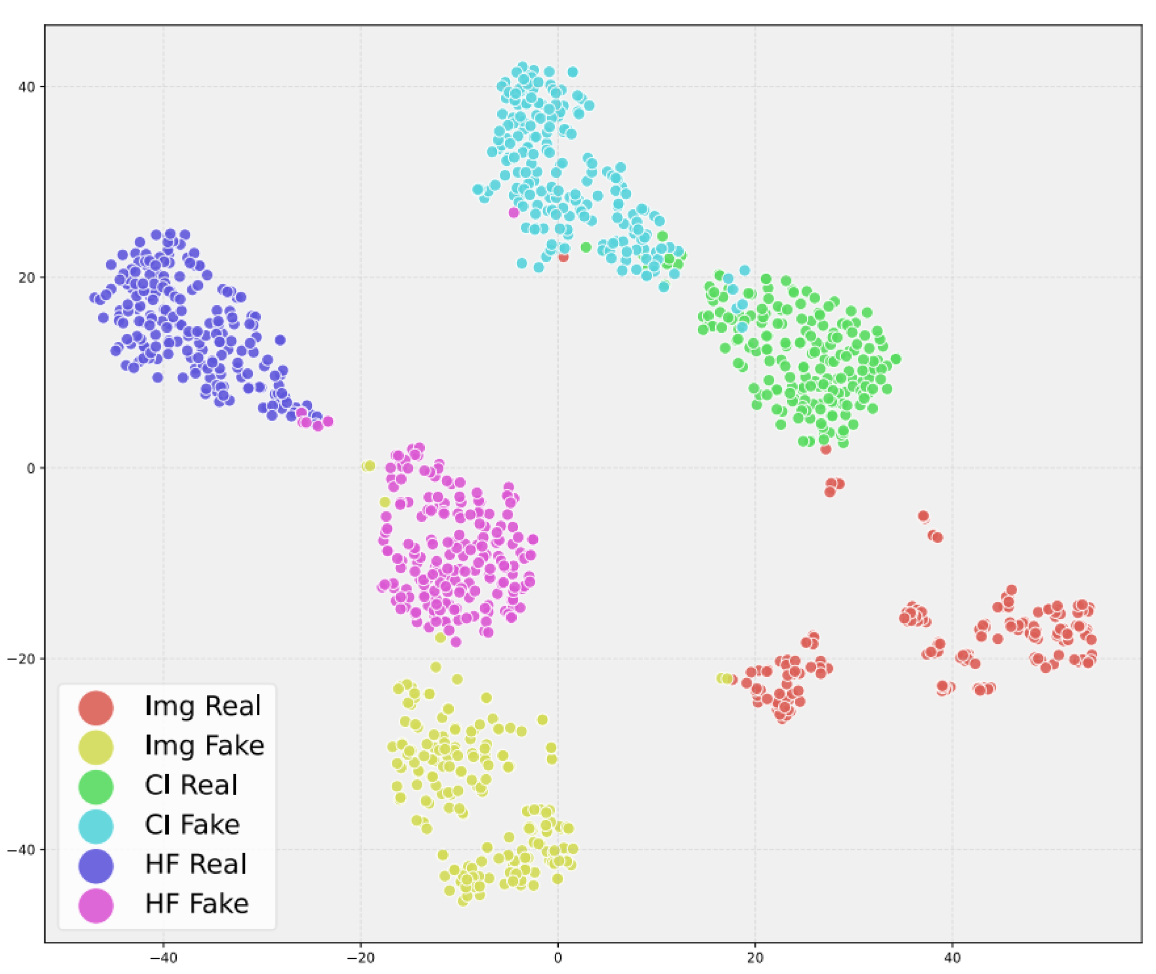

Figure 6: **Visualization of the learned feature space of MCAN when tested on the training model.**