LDCformer: Incorporating Learnable Descriptive Convolution to Vision Transformer for Face Anti-Spoofing

Pei-Kai Huang, Cheng-Hsuan Chiang, Jun-Xiong Chong, Tzu-Hsien Chen, Hui-Yu Ni, Chiou-Ting Hsu
National Tsing Hua University, Taiwan

Face Anti-Spoofing (FAS)

Pros and cons of models on FAS

CNN-based methods
- Pros
  - ➔ Rich of local descriptors to extract local intrinsic features
- Cons
  - ➔ Limited receptive field of convolution operation
  - ➔ Vanilla convolution smooth intrinsic details

ViT-based methods
- Pros
  - ➔ Ability to model long-range dependency between pixels
- Cons
  - ➔ Lack of local descriptor

Goal
- Integrate the pros of both models to mitigate the cons
  - ➔ Incorporate local descriptors into ViT
- Select and enhance suitable convolutional operation for FAS
  - ➔ Introduce decoupling technique onto convolutional operation

LDC : Learnable Descriptive Convolution & Decoupled-LDC : Decoupled Learnable Descriptive Convolution

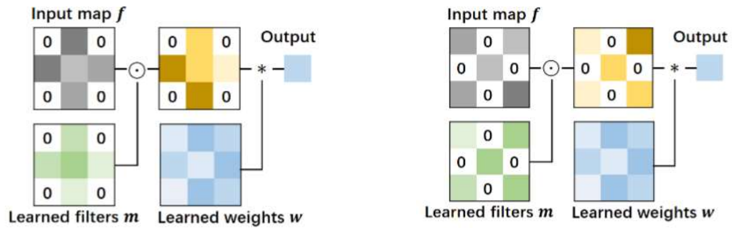
Learnable Descriptive Convolution (LDC)
- Extract detailed intrinsic features

$$g(p) = \underbrace{\sum_{p_n \in \mathcal{R}} w(p_n) \cdot f(p+p_n)}_{\text{vanilla convolution}} + \epsilon \underbrace{\sum_{p_n \in \mathcal{R}} w(p_n) \cdot (f(p+p_n) \cdot m(p_n))}_{\text{learnable descriptive convolution}}$$

LDCformer : Learnable Descriptive Convolutional Vision Transformer

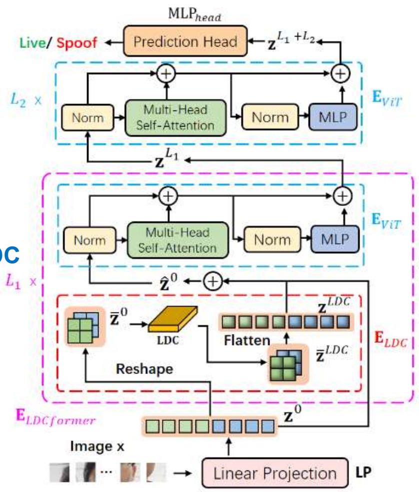
Decoupled Learnable Descriptive Convolution (Decoupled-LDC)
- Sampling sparse local regions to reduce the computational complexity

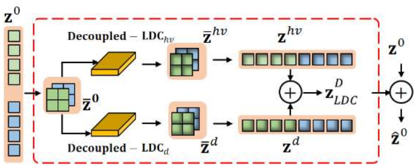
Horizontal-vertical LDC **Diagonal LDC**



LDCformer
- Incorporating LDC into ViT
- $E_{LDCformer}$ extract low-level spoofing cues & textural features

Decoupled-LDCformer
- Replacing LDC with Decoupled-LDC





Loss term

Binary cross entropy loss

$$\mathcal{L}_{ce} = -\sum_{\forall \mathbf{x}} y log(\bar{y})$$

Experiments

Datasets
- OULU(O), MSU(M), CASIA(C), Replay(I)

Evaluation metrics
- Half Total Error Rate (HTER) ↓
- Area Under Curve (AUC) ↑

Ablation studies

ViT with Different CNNs

Method	[I,C,M]→O	
	HTER(%) ↓	AUC(%) ↑
ViT [19]	15.67	88.71
ViT + CNN	14.12	89.59
ViT + CDC [9]	13.30	90.93
ViT + C-CDC [7]	12.99	90.92
ViT + LDC [6] (LDCformer)	12.21	94.36
ViT + Decoupled-LDC (LDCformer^D)	**11.17**	**95.85**

Experimental comparisons

Cross-domain testing

Method	[O,C,I]→ M		[O,M,I]→ C	
	HTER (%) ↓	AUC (%) ↑	HTER (%) ↓	AUC (%) ↑
RAEDFL [4] (*ACPR 21*)	16.67	87.93	17.78	86.11
SSAN-R [8] (*CVPR 22*)	6.67	**98.75**	10.00	**96.67**
PatchNet [5] (*CVPR 22*)	7.10	98.46	11.33	94.58
LDCN [6] (*BMVC 22*)	9.29	96.86	12.00	95.67
TransFAS ∗ [16] (*BBIS 22*)	7.08	96.69	9.81	96.13
TTN-S ∗ [17] (*TIFS 22*)	9.58	95.79	9.81	95.07
ViT ∗ [20] (*IJCB 21*)	10.95	95.05	14.33	92.10
LDCformer^D ∗	**6.43**	98.39	**8.11**	**96.67**

Method	[O,C,M]→ I		[I,C,M]→ O	
	HTER (%) ↓	AUC (%) ↑	HTER (%) ↓	AUC (%) ↑
RAEDFL [4] (*ACPR 21*)	14.64	85.64	18.06	90.04
SSAN-R [8] (*CVPR 22*)	8.88	96.79	13.72	93.63
PatchNet [5] (*CVPR 22*)	14.6	92.51	11.82	95.07
LDCN [6] (*BMVC 22*)	9.43	95.02	13.51	93.68
TransFAS ∗ [16] (*BBIS 22*)	10.12	95.53	15.52	91.10
TTN-S ∗ [17] (*TIFS 22*)	14.15	94.06	12.64	94.20
ViT ∗ [20] (*IJCB 21*)	16.64	85.07	15.67	89.59
LDCformer^D ∗	**8.57**	**97.09**	**11.17**	**95.85**