Leveraging Representations from Intermediate Encoder-Blocks for Synthetic Image Detection

Christos Koutlis and Symeon Papadopoulos

{ckoutlis,papadop}@iti.gr

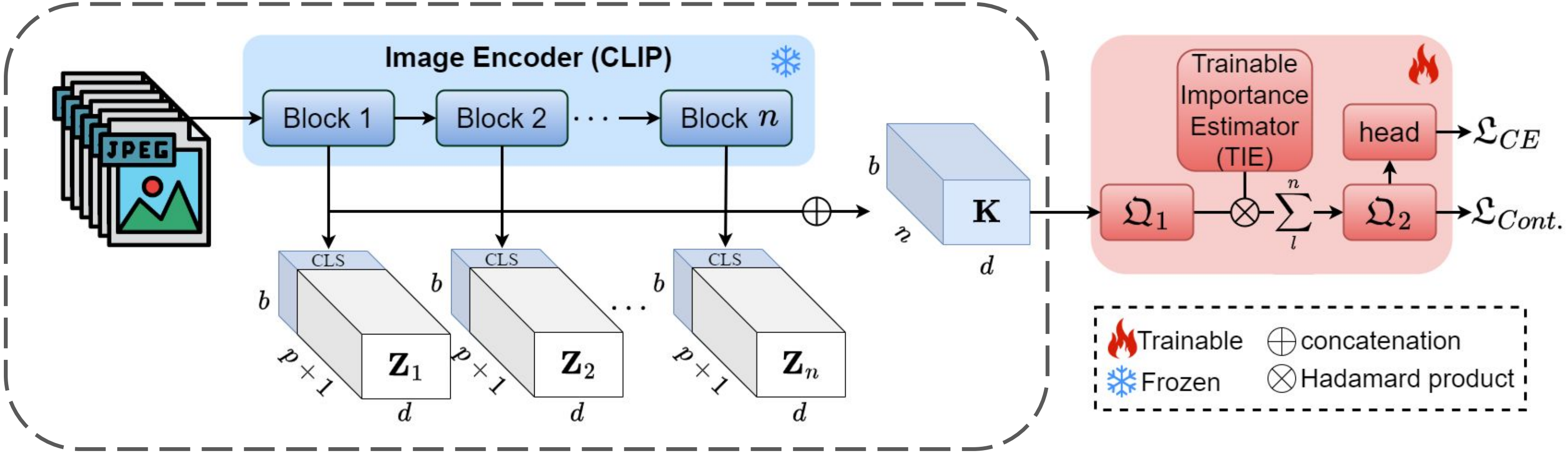
Information Technologies Institute @ CERTH, Thessaloniki, Greece

Motivation

- Feature extraction by foundation models yields high SID performance with minimal training requirements (Ojha et al. 2023), although these last layer features capture high-level semantics.
- Low-level features from intermediate layers, known to be relevant for SID (Bayar & Stamm 2018, Corvi et al. 2023), have not been explored to date, despite their performance-boosting potential.

Ojha et al. (2023). Towards universal fake image detectors that generalize across generative models. CVPR (pp. 24480-24489).
Bayar & Stamm (2018). Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection. IEEE Trans. Inf. Forensics Secur., 13(11), 2691-2706.
Corvi et al. (2023). On the detection of synthetic images generated by diffusion models. ICASSP (pp. 1-5).

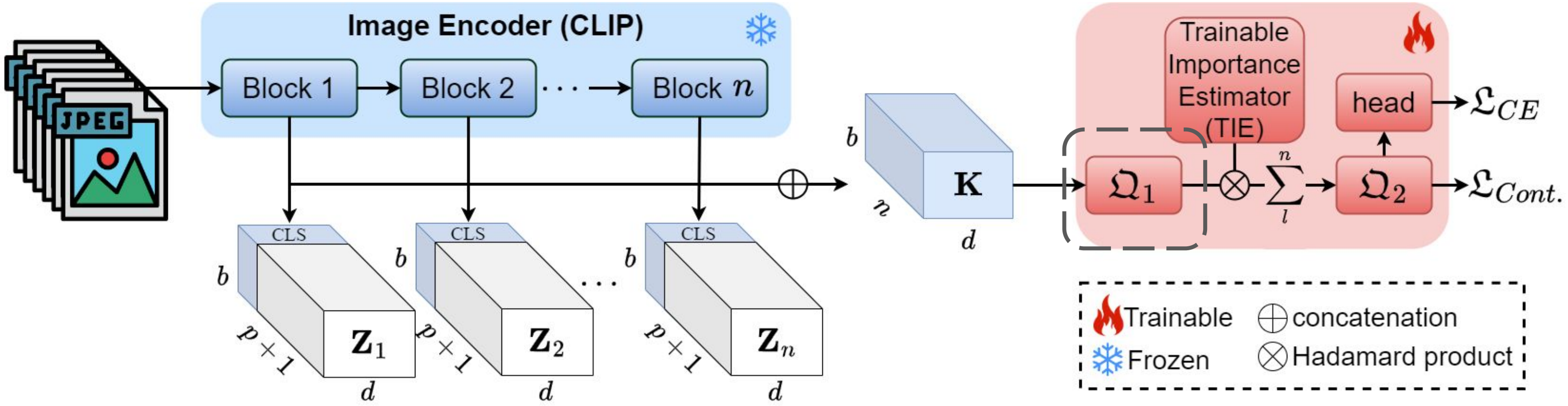
Methodology



We define the Representations from Intermediate Encoder-blocks $\mathbf{K}$, as the concatenation of CLS tokens from the corresponding n blocks of CLIP's image encoder:

$$\mathbf{K} = \oplus \{ \mathbf{Z}_l^{[0]} \}_{l=1}^{n} \in \mathbb{R}^{b \times n \times d}$$

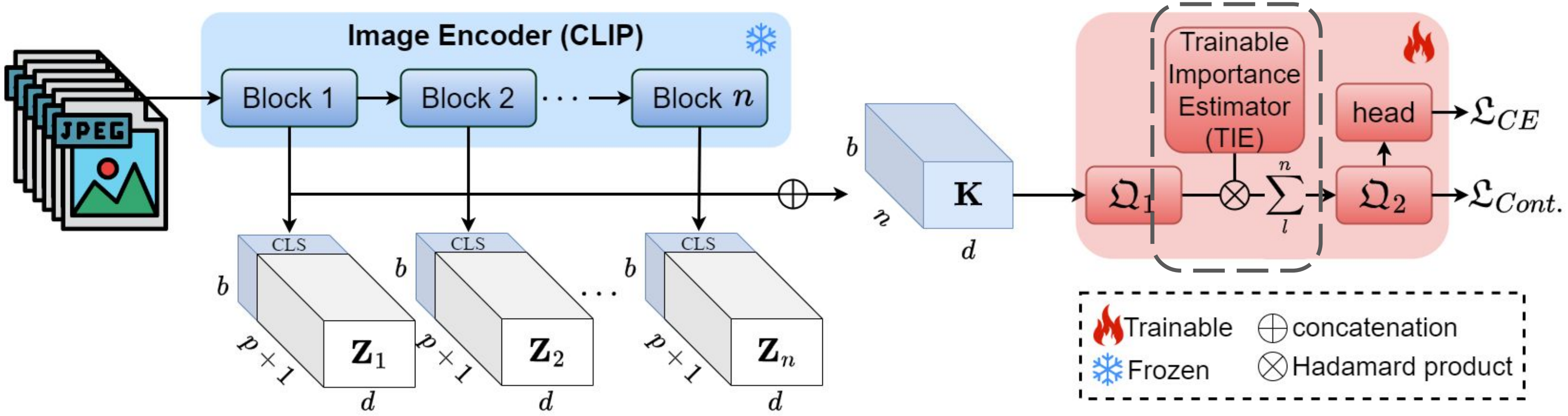
Methodology



Feed-forward projection networks process $\mathbf{K}$, appling ReLU and dropout after each layer:

$$\mathbf{K}_m = \mathrm{ReLU}(\mathbf{K}_{m-1}\mathbf{W}_m + \mathbf{b}_m) \in \mathbb{R}^{b\times n\times d'}$$

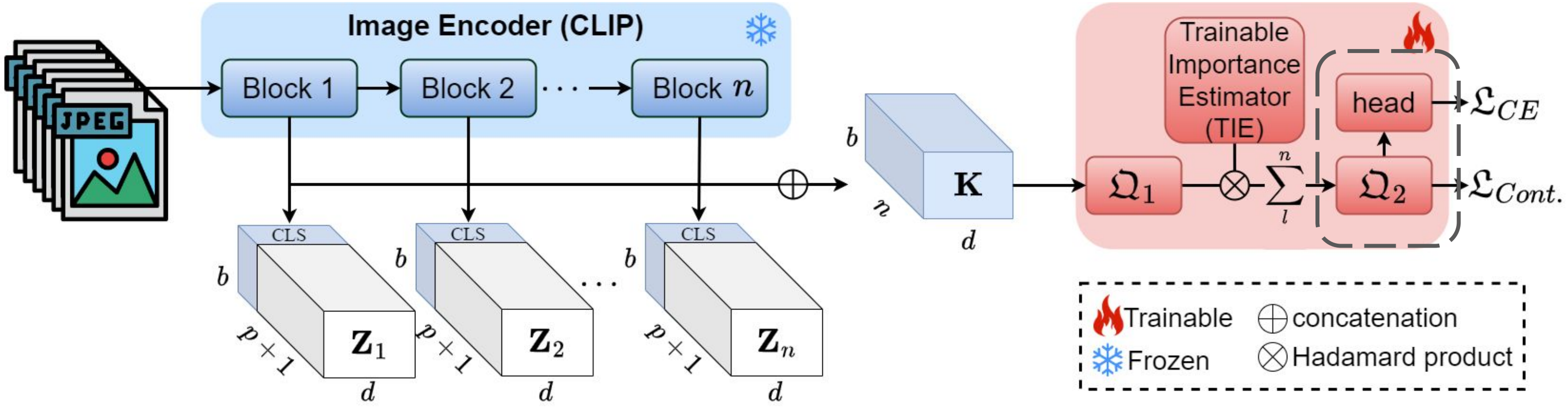
Methodology



The Trainable Importance Estimator (TIE) module $\mathbf{A}$, estimates the importance of feature k at processing stage l and applies the weights accordingly:

$$\tilde{\mathbf{K}}^{(ik)} = \sum_{l}^{n} \mathcal{S}(\mathbf{A})^{(lk)} \cdot \mathbf{K}_q^{(ilk)}$$

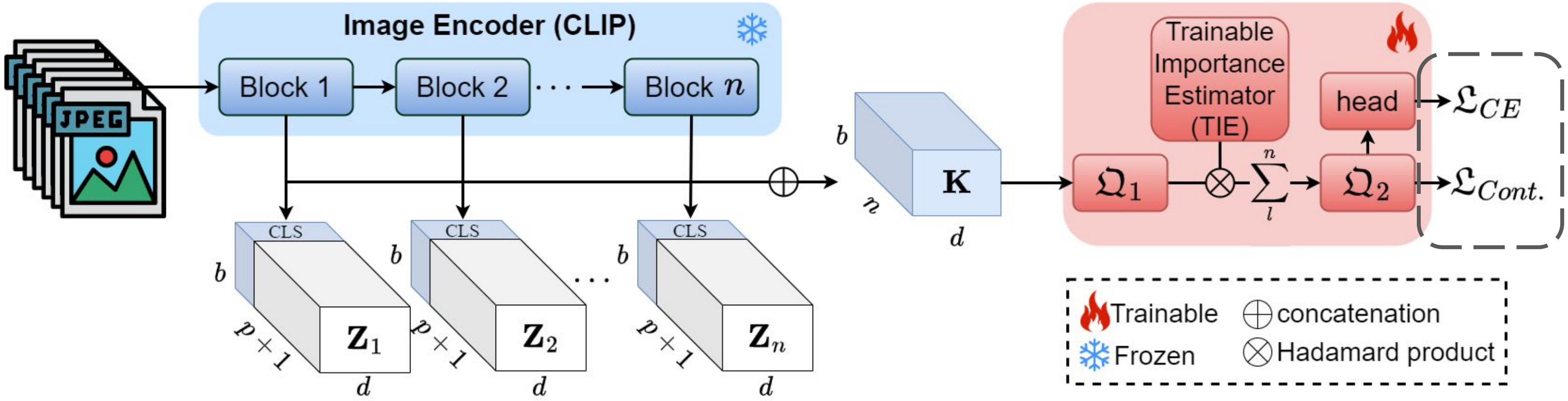
Methodology



The Trainable Importance Estimator (TIE) module $\mathbf{A}$, estimates the importance of feature k at processing stage l and applies the weights accordingly:

$$\tilde{\mathbf{K}}^{(ik)} = \sum_{l}^{n} \mathcal{S}(\mathbf{A})^{(lk)} \cdot \mathbf{K}_q^{(ilk)}$$

Methodology



The objective function is the combination of binary cross-entropy and Supervised Contrastive Learning (Khosla et al. 2020).

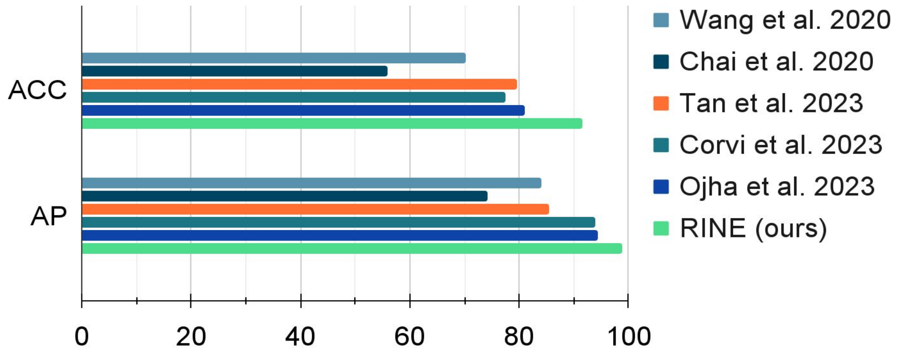
$$\mathfrak{L}_{CE} = -\sum_{i=1} y_i \log \hat{y}_i + (1 - y_i) \cdot \log(1 - \hat{y}_i)$$

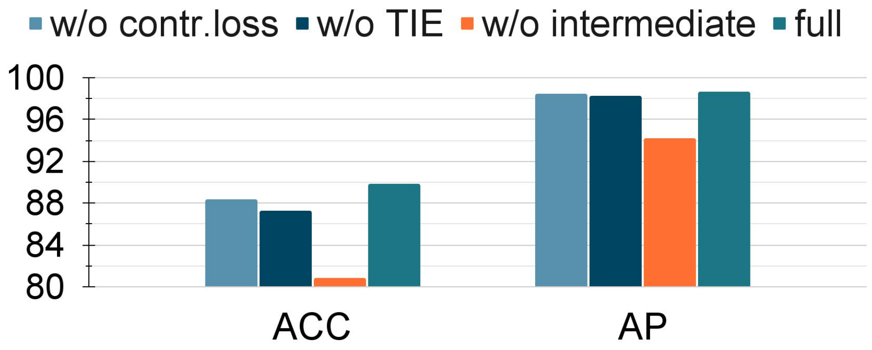
$$\mathfrak{L}_{Cont.} = -\sum_{i=1}^{b} \frac{1}{G(i)} \sum_{g \in G(i)} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_g / \tau)}{\sum_{a \in A(i)} \exp(\mathbf{z}_i \cdot \mathbf{z}_a / \tau)}$$

$$\mathfrak{L} = \mathfrak{L}_{CE} + \xi \cdot \mathfrak{L}_{Cont.}$$

Khosla et al. (2020). Supervised contrastive learning. NeurIPS (pp. 18661-18673).

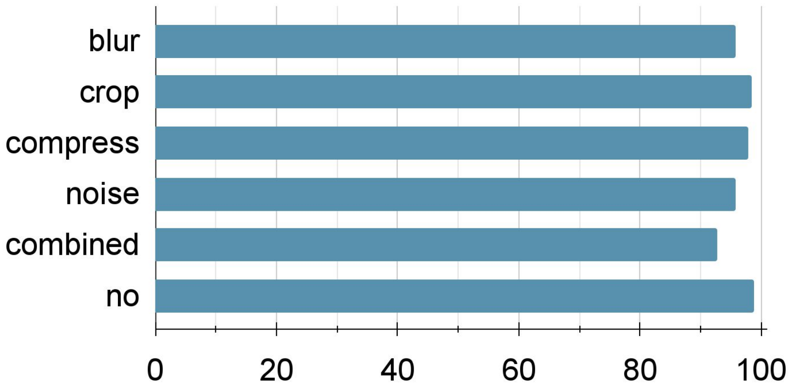
Results

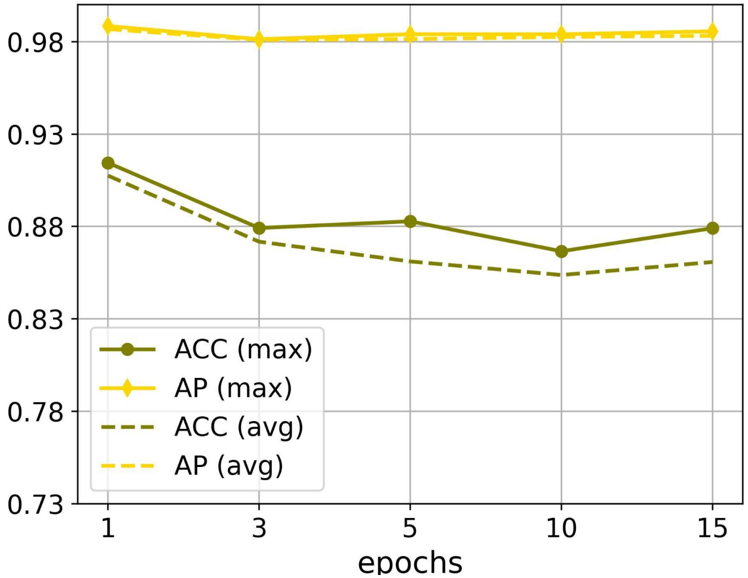


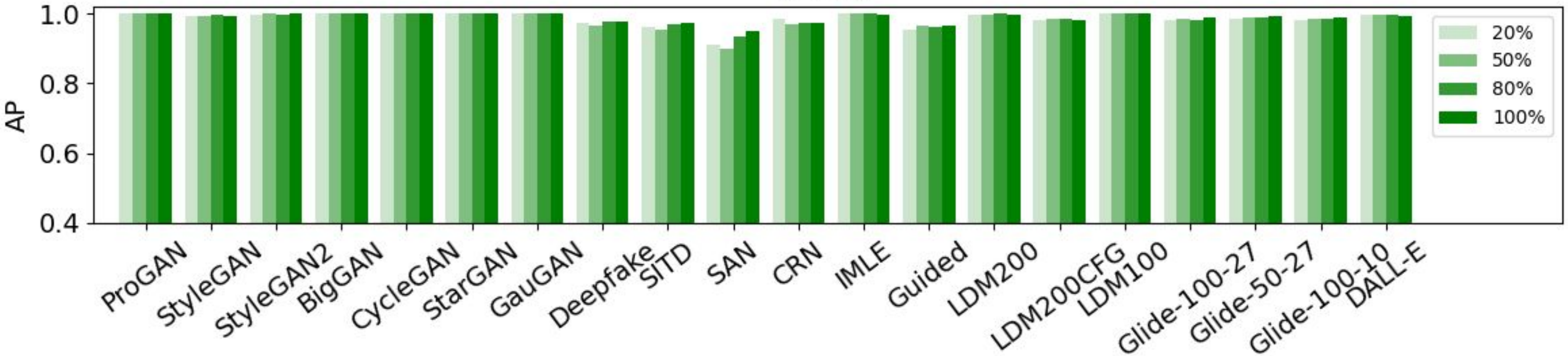


- Evaluation on data from 20 different GAN-based and Diffusion model-based generators
- RINE outperforms existing state-of-the-art methods
- The ablation analysis supports our methodological choices

Results


blur
crop
compress
noise
combined
no
0
20
40
60
80
100


0.98
0.93
0.88
0.83
0.78
0.73
ACC (max)
AP (max)
ACC (avg)
AP (avg)
1
3
5
10
15
epochs


1.0
0.8
0.6
0.4
AP
20%
50%
80%
100%
ProGAN
StyleGAN
StyleGAN2
BigGAN
CycleGAN
StarGAN
GauGAN
Deepfake
SITD
SAN
CRN
IMLE
Guided
LDM200
LDM200CFG
LDM100
Glide-100-27
Glide-50-27
Glide-100-10
DALL-E