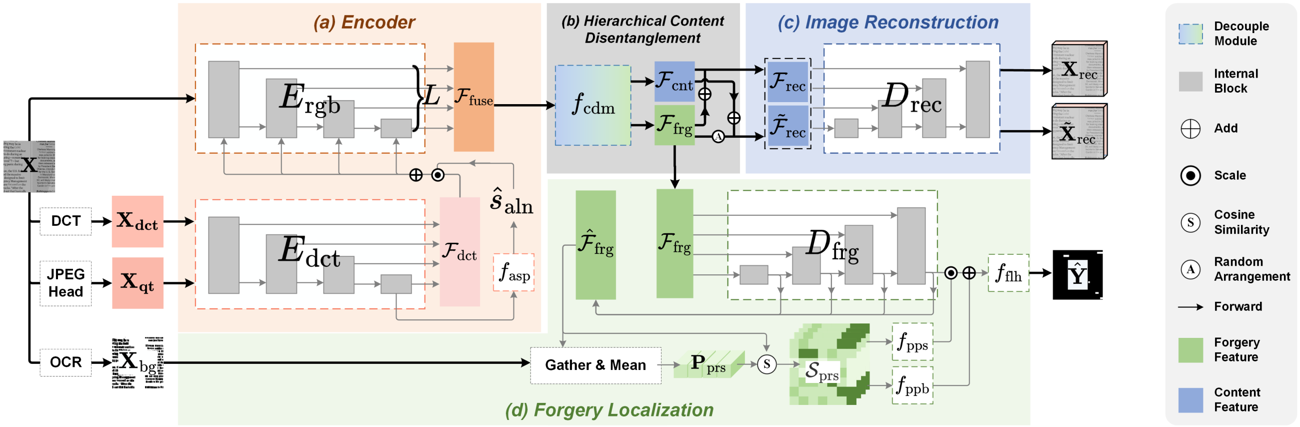

Figure 3. Overview of ACDC-Net, consisting of (a) adaptive RGB-DCT encoder E extracting multi-scale features from RGB and DCT domains; (b-c) HCD module f_{cdm} disentangling forgery and content features and the reconstruction branch reconstructs original image based on content features; (d) localization branch locates the tampered region based on decoupled forgery features.

27]; but fragile to various post-processing that disrupts the 8×8 block alignment [6, 12], such as cropping, resizing. Fig. 3 (a) illustrates our adaptive RGB-DCT encoder E, which leverages DCT domain features while mitigating their sensitivity. The encoder consists of two subnetworks: one extracts RGB features $E_{\text{rgb}}(\mathbf{X})$ to produce the fused feature set $\mathcal{F}_{\text{fuse}} = \{\mathbf{F}^i_{\text{fuse}}\}^L_{i=1}$, and the other extracts DCT features $E_{\text{dct}}(\mathbf{X}_{\text{dct}}, \mathbf{X}_{\text{qt}})$ to yield $\mathcal{F}_{\text{dct}} = \{\mathbf{F}^i_{\text{dct}}\}^L_{i=1}$. For each level i, the DCT feature $\mathbf{F}^i_{\text{dct}}$ is adaptively fused into the corresponding RGB feature using a predicted alignment score $\hat{s}_{\text{aln}} = f_{\text{asp}}(\mathbf{F}^L_{\text{dct}}) \in (0, 1)$, which controls its contribution. Specifically, the fusion is defined as

$$\mathbf{F}^{i+1}_{\text{fuse}} = f^i_{\text{fuse}}\Big(\mathbf{F}^i_{\text{rgb}} + f_{\text{asp}}(\mathbf{F}^L_{\text{dct}}) \cdot \mathbf{F}^i_{\text{dct}}\Big), \tag{1}$$

where f^i_{fuse} denotes the level-i block of E_{rgb} and L is the total number of blocks. The alignment score $\hat{s}_{\text{aln}}$ is predicted by a head appended to the last layer of E_{dct} and is applied uniformly across all scales. We cast its estimation as a classification task, with the output probability serving as $\hat{s}_{\text{aln}}$. Given forgery images with well-aligned DCT blocks, we disrupt the alignment using augmentations like random resizing, cropping, and pixel shifting[1] to create non-aligned samples. The ground-truth label s_{aln} can be readily assigned, with non-aligned samples labeled as "0", while aligned ones as "1".

Remark: Our adaptive DCT encoder builds on the MoE principle [13, 32], showing its potential to generalize across various forensic traces beyond just DCT features. In a typical MoE setup, specialized expert subnetworks are weighted by a gating mechanism. Here, E_{dct} targets compression artifacts while E_{rgb} captures robust and general forensic cues, with an adaptive routing function $f_{\text{asp}}(\mathbf{F}^L_{\text{dct}})$ controlling the DCT contribution for each sample. Unlike existing MoE methods that learn routing weights using the prediction-GT [13] or sparsity [32] objectives, we optimize f_{asp} with an explicit classification task to prioritize DCT alignment. Although both DTD [27] and our ADCD-Net leverage RGB and DCT traces, DTD overfits to DCT features and is vulnerable to post-processing (see Fig. 5 DTD). Even with RGB input, DTD is significantly less robust than other RGB-based detectors such as TruFor and ADCD-Net. This indicates that DTD's simple RGB-DCT concatenation fails to prioritize critical features. In contrast, our adaptive DCT encoder with the MoE structure enables the model to rely primarily on RGB cues when DCT traces are weak, and vice versa (see Fig. 7).

3.2. Hierarchical Content Decoupling

To mitigate our observed text-BG bias (see Fig. 2), we propose the HCD module f_{cdm} to disentangle content and forgery features at multiple scales. This disentanglement is possible because forgery cues stem from subtle inconsistencies or noise rather than the intrinsic image content, while content features are closely tied to the visual structure (see Fig. 8). As depicted in Fig. 3 (b), after we obtain $\mathcal{F}_{\text{fuse}}$, the decoupling module f_{cdm} is employed to map $\mathbf{F}^i_{\text{fuse}}$ into content $\mathbf{F}^i_{\text{cnt}}$ and forgery feature $\mathbf{F}^i_{\text{frg}}$ at each i scale, such that:

$$f_{\text{cdm}}(\mathbf{F}^i_{\text{fuse}}) = [\mathbf{F}^i_{\text{cnt}}, \mathbf{F}^i_{\text{frg}}] \in \mathbb{R}^{H \times W \times 2C}. \tag{2}$$

For efficiency, f_{cdm} at each scale is implemented as a multi-layer perceptron (MLP). The decoupled process is performed at each scale of the input features, resulting in $[\mathcal{F}_{\text{cnt}}, \mathcal{F}_{\text{frg}}] = f_{\text{cdm}}(\mathcal{F}_{\text{fuse}})$. Then, $\mathcal{F}_{\text{cnt}}$ and $\mathcal{F}_{\text{frg}}$ are passed to the following reconstruction and localization decoders. It is important to note that the decoupling process is not solely carried out by f_{cdm}; instead, it is achieved through the following reconstruction and localization decoders and

[1] We crop and shift pixels $n \bmod 8 \neq 0$ to avoid block-aligned output.