the training objectives. Consequently, a lightweight MLP is sufficient within the decoupling module.

As shown in Fig. 3 (c), the decomposed features $\mathcal{F}_{\text{cnt}}$ and $\mathcal{F}_{\text{frg}}$ are used to reconstruct the original image $\mathbf{X}$ and DCT coefficients $\mathbf{X}_{\text{dct}}$. To reduce the content influence in the forgery features $\mathcal{F}_{\text{frg}}$, we ensure that the content information is primarily captured by $\mathcal{F}_{\text{cnt}}$. This is achieved by randomly shuffling the spatial arrangement of $\mathcal{F}_{\text{frg}}$ to form $\tilde{\mathcal{F}}_{\text{frg}}$ before fusing it back with $\mathcal{F}_{\text{cnt}}$, and then requiring the network to reconstruct $[\mathbf{X}, \mathbf{X}_{\text{dct}}]$. Specifically, $\mathcal{F}_{\text{rec}} = \{\mathbf{F}^i_{\text{cnt}} + \mathbf{F}^i_{\text{frg}}\}_{i=1}^L$ and $\tilde{\mathcal{F}}_{\text{rec}} = \{\mathbf{F}^i_{\text{cnt}} + \tilde{\mathbf{F}}^i_{\text{frg}}\}_{i=1}^L$ serve as two distinct inputs to the reconstruction decoder D_{rec} for generating $\mathbf{X}_{\text{rec}} = D_{\text{rec}}(\mathcal{F}_{\text{rec}})$ and $\tilde{\mathbf{X}}_{\text{rec}} = D_{\text{rec}}(\tilde{\mathcal{F}}_{\text{rec}})$, which are respectively the reconstructed versions of $\mathbf{X}$ and $\mathbf{X}_{\text{dct}}$.

Remark: Existing content disentanglement methods [17, 20, 23, 42, 44] are mainly designed for face forgery detection. Motivated by our observed text-BG bias in Fig. 2 (e), we extend content disentanglement to dense prediction tasks with multi-modal inputs by proposing the HCD module. This module hierarchically decouples content, preventing data leakage through shortcuts in the U-shaped network and effectively disentangling multi-scale features to accommodate varying text scales in documents.

3.3. Forgery Localization

We introduce the PPE module, which injects prior knowledge of untampered regions to enhance performance. PPE can be selectively employed when high-confidence pristine areas exist (*e.g.* the BG of deepfake portraits and forged documents, which is less informative and predominantly pristine). As shown in Fig. 3 (d), an OCR model f_{ocr} is used to extract the BG of the image $\mathbf{X}_{\text{bg}} = f_{\text{ocr}}(\mathbf{X}) \in \{0,1\}^{H\times W}$, in which "0" is the text area while "1" is the BG. The estimated pristine prototype on level-i feature is computed by

$$\mathbf{p}^i_{\text{prs}} = \frac{\sum_{h,w} \mathbf{X}_{\text{bg}}(h,w)\hat{\mathbf{F}}^i_{\text{frg}}(h,w)}{\sum_{h,w} \mathbf{X}_{\text{bg}}(h,w)}, \tag{3}$$

in which h and w are the index of height and width, and $\hat{\mathbf{F}}^i_{\text{frg}}$ is the output of the level-i block in D_{frg}. Then, the estimated pristine map can be obtained by

$$\mathbf{S}^i_{\text{prs}}(h,w) = \frac{\hat{\mathbf{F}}^i_{\text{frg}}(h,w)\cdot \mathbf{p}^i_{\text{prs}}}{\|\hat{\mathbf{F}}^i_{\text{frg}}(h,w)\|\|\mathbf{p}^i_{\text{prs}}\|}. \tag{4}$$

As can be observed in Fig. 2 (f), by incorporating the HCD, the pristine prototype (in blue cross) is more accurately located in the pristine cluster. The pristine map is computed at multiple scales, resulting in $\mathcal{S}_{\text{prs}} = \{\mathbf{S}^i_{\text{prs}}\}_{i=1}^L$. The maps $\mathcal{S}_{\text{prs}}$ modulate the penultimate feature via element-wise scaling and biasing. Two MLP layers, f_{pps} and f_{ppb}, convert $\mathcal{S}_{\text{prs}}$ into a scale and a bias, respectively. This yields

$$\hat{\mathbf{Y}} = f_{\text{flh}}\Big(\hat{\mathbf{F}}^L_{\text{frg}}\cdot f_{\text{pps}}(\mathcal{S}_{\text{prs}}) + f_{\text{ppb}}(\mathcal{S}_{\text{prs}})\Big), \tag{5}$$

where $\hat{\mathbf{F}}^L_{\text{frg}}$ is the penultimate feature and f_{flh} is the segmentation head producing the predicted tampered map $\hat{\mathbf{Y}}$.

3.4. Training Objectives

ADCD-Net is trained in an end-to-end fashion by using the following loss function:

$$\mathcal{L} = \lambda_{\text{aln}}\mathcal{L}_{\text{aln}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{frg}}\mathcal{L}_{\text{frg}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}}. \tag{6}$$

Here, the alignment score loss $\mathcal{L}_{\text{aln}}$ ensures the accuracy of the predicted alignment score $\hat{s}_{\text{aln}}$. The image reconstruction loss $\mathcal{L}_{\text{rec}}$ maintains the quality of the reconstructed image, implicitly validating the feature disentanglement. The forgery localization loss $\mathcal{L}_{\text{frg}}$ ensures accurate prediction of the tampered mask. Lastly, the within-image contrastive loss $\mathcal{L}_{\text{con}}$ amplifies the distinction between pristine and forged pixels in the feature domain. The details on the calculation of these four loss terms are given below.

Alignment score loss $\mathcal{L}_{\text{aln}}$. The cross-entropy loss $\mathcal{L}_{\text{aln}}$ is computed between the prediction score $\hat{s}_{\text{aln}}$ and the ground-truth s_{aln}, to accurately and dynamically control the magnitude of the DCT features toward more robust localization.

Image reconstruction loss $\mathcal{L}_{\text{rec}}$. The ℓ_1 loss is used to ensure the reconstructed content is close to the original:

$$\begin{aligned}\mathcal{L}_{\text{rec}} = \|[\mathbf{X},\mathbf{X}_{\text{dct}}] - D_{\text{rec}}(\{\mathbf{F}^i_{\text{cnt}} + \mathbf{F}^i_{\text{frg}}\}_{i=1}^L)\| +\\ \|[\mathbf{X},\mathbf{X}_{\text{dct}}] - D_{\text{rec}}(\{\mathbf{F}^i_{\text{cnt}} + \tilde{\mathbf{F}}^i_{\text{frg}}\}_{i=1}^L)\|.\end{aligned} \tag{7}$$

Forgery localization loss $\mathcal{L}_{\text{frg}}$. The loss $\mathcal{L}_{\text{frg}}$ is used to compute the error between the prediction $\hat{\mathbf{Y}}$ and the ground-truth forgery mask $\mathbf{Y}$ with the cross-entropy loss and Lovase loss [3], in which $\mathcal{L}_{\text{frg}} = \lambda_{\text{ce}}\mathcal{L}_{\text{ce}}(\hat{\mathbf{Y}},\mathbf{Y}) + \mathcal{L}_{\text{lov}}(\hat{\mathbf{Y}},\mathbf{Y})$.

FOCAL loss $\mathcal{L}_{\text{con}}$. Inspired by FOCAL [40], with the idea that forged/pristine pixels are relative concepts within an image, we adopt the within-image contrastive loss to further enlarge the discrepancy between pristine and forged pixels. The contrastive loss is computed on multi-scale forgery features $\hat{\mathcal{F}}_{\text{frg}} = \{\hat{\mathbf{F}}^i_{\text{frg}}\}_{i=1}^L$. Given the extreme imbalance between pristine and forged areas, we use the SupCon loss [15, 50] to balance the influence of each class. Specifically, the contrastive loss for the i-th level feature $\hat{\mathbf{F}}^i_{\text{frg}}$ is formulated as:

$$\mathcal{L}^b_{\text{con},i} = \sum_{\mathbf{z}\in\mathcal{Z}} \frac{-1}{|\mathcal{P}|}\sum_{\mathbf{p}\in\mathcal{P}} \log\frac{\exp(\mathbf{z}\cdot\mathbf{p})}{\sum_{j\in\mathcal{Y}}\sum_{\mathbf{a}\in\mathcal{A}_j}\exp(\mathbf{z}\cdot\mathbf{a})}. \tag{8}$$

where $\mathcal{Z}$ denotes the entire pixel set for $\hat{\mathbf{F}}^i_{\text{frg}}$. The positive set $\mathcal{P}$ for a pixel $\mathbf{z}$ is defined as $\mathcal{P} = \{\mathbf{p}\in\mathcal{Z} \mid y_{\mathbf{p}} = y_{\mathbf{z}}\}\setminus\{\mathbf{z}\}$. The label set $\mathcal{Y} = \{0,1\}$ represents pristine