In the second step, we attempt to learn the content-aware feature vector for each pair of image and text samples. The challenge of cross-modal retrieval is that the cross-modal data has significant heterogeneity. We thus leverage label information to bridge the gap between image and text sample pairs by learning the content-aware feature vector in the common space. We employ the multi-modal conditional principal label space transformation (CPLST) [47], which falls into the range of label space dimension reduction (LSDR) paradigm, to obtain the content-aware feature vector S. It exploits both the label and the feature data from different modalities to compress the label space.

Since what we aim to do is to learn feature vectors in a d-dimensional common space, where $\mathrm{d} \leq \mathrm{c}$, we first shift each label information l_i to $y_i = l_i - \bar{l}, i = 1,2,\dots,n$, where $\bar{l} = \frac{1}{n}\sum_i^n l_i$ represents the estimated mean of vectors in label matrix L. Let Y contain y_i as columns. Then, we learn the best content-aware feature vector by simultaneously minimizing the prediction error and coding error:

$$\min_{W_v, W_t, W_Y W_Y{}^T = I} \left(\| W_t \mathrm{T} - W_Y \mathrm{Y} \|_F^2 + \| W_v \mathrm{V} - W_Y \mathrm{Y} \|_F^2 + \| \mathrm{Y} - W_Y{}^T W_Y \mathrm{Y} \|_F^2 \right),$$

where $\|\cdot\|_F$ denotes the Frobenius-norm, $\mathrm{I} \in \mathbb{R}^{d*d}$ is an identity matrix, $W_v \in \mathbb{R}^{d*d_v}$, $W_t \in \mathbb{R}^{d*d_t}$ and $W_Y \in \mathbb{R}^{d*c}$. In the formula, the first two items are the prediction error terms of image and text, and the last item is the coding error term. It is worth noting that the prediction error terms consider the feature and label information equally while the coding error item ensures that the semantic information can be reconstructed correctly during the decoding process. Thus, we ensure that the content-aware feature vector generated by the second step contains feature information from both image and text modalities. After achieving W_Y, we linearly map Y to the code vector S by $\mathrm{S} = W_Y Y$. The code vector S is the content-aware feature vector which will be used in the following step.