APPENDIX

Proof of Theorem 4.2. As the gradient matching objective is commonly calculated for each class separately, we define the class-wise loss for $\mathcal{T}$ and $\mathcal{S}$ as:

$$\begin{aligned}\mathcal{L}_i^{\mathcal{T}} &= \frac{1}{2}\left\|\Phi_\theta(X_i)\mathbf{W} - Y_i\right\|^2 + \lambda\|\mathbf{W}\|^2,\\ \mathcal{L}_i^{\mathcal{S}} &= \frac{1}{2}\left\|\Phi_\theta(X_i')\mathbf{W} - Y_i'\right\|^2 + \lambda\|\mathbf{W}\|^2,\end{aligned} \tag{23}$$

where X_i, X_i' denote the samples belonging to class i in $\mathcal{T}$ and $\mathcal{S}$, respectively, and Y_i, Y_i' are the corresponding class-wise label matrices. For brevity, we denote $\Phi_i := \Phi_\theta(X_i)$ and $\Phi_i' := \Phi_\theta(X_i')$.

$$\begin{aligned}\mathcal{L}_{\text{GM}} &= \sum_{i=0}^{n-1}\left\|\frac{1}{|n_i|}\nabla\mathcal{L}_i^{\mathcal{T}}(\mathbf{W}) - \frac{1}{|n_i'|}\nabla\mathcal{L}_i^{\mathcal{S}}(\mathbf{W})\right\|^2\\ &= \sum_{i=0}^{C-1}\left\|\frac{1}{|n_i|}\left(\Phi_i^\top\Phi_i\mathbf{W} - \Phi_i^\top Y_i\right) - \frac{1}{|n_i'|}\left(\Phi_i'^\top\Phi_i'\mathbf{W} - \Phi_i'^\top Y_i'\right)\right\|^2\\ &= \sum_{i=0}^{C-1}\left\|\left(\frac{1}{|n_i|}\Phi_i^\top\Phi_i - \frac{1}{|n_i'|}\Phi_i'^\top\Phi_i'\right)\mathbf{W} - \left(\frac{1}{|n_i|}\Phi_i^\top Y_i - \frac{1}{|n_i'|}\Phi_i'^\top Y_i'\right)\right\|^2\\ &\leq \sum_{i=0}^{C-1}\left\|\frac{1}{|n_i|}\Phi_i^\top Y_i - \frac{1}{|n_i'|}\Phi_i'^\top Y_i'\right\|^2 + \sum_{i=0}^{C-1}\left\|\frac{1}{|n_i|}\Phi_i^\top\Phi_i - \frac{1}{|n_i'|}\Phi_i'^\top\Phi_i'\right\|^2\|\mathbf{W}\|^2\\ &= \left\|\mathbf{P}\Phi(X) - \mathbf{P}'\Phi(X')\right\|^2 + \sum_{i=0}^{C-1}\left\|\frac{1}{|n_i|}\Phi_i^\top\Phi_i - \frac{1}{|n_i'|}\Phi_i'^\top\Phi_i'\right\|^2\|\mathbf{W}\|^2.\end{aligned} \tag{24}$$

This decomposition reveals that distribution matching, when combined with second-order embedding alignment, provides an upper bound for class-wise gradient matching. □