A. Dataset Details

We evaluate our method on six benchmark datasets:

- **Multi-MNIST** dataset [38], a multi-task variant of MNIST dataset where the input features are composed of two digit pictures in the left and right position, respectively. Following MT-CRL [14], we randomly shuffle the label pairs and split the training and testing set such that every label co-occurrence in the training set will no longer appear in the testing set. The tasks are to predict the left digits and right digits. We use a 3-layer CNN as the encoder, and one linear layer as per-task predictor. We choose classification accuracy as the evaluation metric for both tasks.
- **CelebA** dataset [25], which contains $202,599$ face images, each of resolution 178×218, with 40 binary attributes. We consider four classification tasks, including gender, smiling, age and attractiveness. The testing data is evenly sampled across different combinations of task labels to construct covariate shift between training and testing set. We use ResNet-18 as the encoder, and one linear layer as per-task predictor. We choose classification accuracy as the evaluation metric for the four tasks.
- **Taskonomy** dataset [45], a MTL benchmark dataset of indoor scene images from various buildings. We consider three tasks, including semantic segmentation, scene classification and object classification. Following MT-CRL [14], we select images from non-overlapping 48 and 6 buildings as the training and testing set. We use ResNet-50 as the encoder, one linear layer as per-task predictor for classification tasks and a 15-layer CNN with upsampling blocks for semantic segmentation. We choose classification accuracy as the evaluation metric for scene and objective classification, and cross-entropy loss as the evaluation metric for semantic segmentation.
- **MetaShift** dataset [22], a large-scale dataset for evaluating distribution shifts composed of $12,868$ sets of natural images across 410 classes. Following the authors' construction, we consider two tasks: 10-class animal classification and scene classification. On training data, the cats and dogs classes are spuriously correlated with 'indoor/outdoor' scene, while on the testing data, the task labels are evenly distributed across different combinations of animal and scene classes. We use ResNet-18 as the encoder, and one linear layer as per-task predictor. We choose classification accuracy as the evaluation metric for both tasks.
- **NYU-V2** dataset [35], which consists of 1449 RGB-D indoor scene images. We We consider two tasks, including semantic segmentation and depth estimation. On training data, the foreground objects are spuriously correlated with shallower depths, while on the testing data, the objects are evenly distributed across different depths. We choose mean IoU and relative error as evaluation metrics for semantic segmentation and depth estimation, respectively.
- **CityScape** dataset [7], which consists of 3475 street-view images. We consider two tasks, including semantic segmentation and depth estimation. On training data, the foreground objects are spuriously correlated with shallower depths, while on the testing data, the objects are evenly distributed across different depths. We choose mean IoU and relative error as evaluation metrics for semantic segmentation and depth estimation, respectively.

B. Additional Results

Results on CityScape dataset are shown in Tab. 11. Our method shows better preformance over baselines on both datasets, validating the effectiveness of our method.

C. Proof of Lemma 1

Proof. Let $p_{y^i,y^j} := P(Y^i = y^i, Y^j = y^j)$ be the joint probability of task labels, we have

$$p_{00} = \alpha_0^{ij} p_{01},\ p_{10} = \alpha_1^{ij} p_{11},$$

where α_0^{ij} and α_1^{ij} remains constant before and after resampling since the random sampling does not alter the corresponding labels for task b. Since $\sum_{\{y^i,y^j\}} p_{y^i,y^j} = 1$, based on the proportionality, we have

$$p_{01} = \frac{1 - (1 + \alpha_1^{ij}) p_{11}}{1 + \alpha_0^{ij}}. \tag{3}$$

Since $p^i = p_{10} + p_{11}$ and $p^j = p_{11} + p_{01}$, based on Eq. (3) we have

$$p^j = \frac{1}{1+\alpha_0^{ij}} + \left(\frac{1}{1+\alpha_1^{ij}} - \frac{1}{1+\alpha_0^{ij}}\right) p^i = (\hat{\alpha}_1^{ij} - \hat{\alpha}_0^{ij}) p^i + \hat{\alpha}_0^{ij}, \tag{4}$$

where $\hat{\alpha}_0^{ij} = \frac{1}{1+\alpha_0^{ij}}$ and $\hat{\alpha}_1^{ij} = \frac{1}{1+\alpha_1^{ij}}$. Eq. (4) indicates that the expectations of Y^i and Y^j remains linearly proportional before and after resampling. Accordingly, we have the correlation coefficient between Y^i and Y^j as

$$\begin{aligned}\rho_{Y^i,Y^j} &= \frac{\mathbb{E}[Y^i Y^j] - \mathbb{E}[Y^i]\mathbb{E}[Y^j]}{\sqrt{\mathbb{E}\left[(Y^i)^2\right] - (\mathbb{E}[Y^i])^2}\sqrt{\mathbb{E}\left[(Y^j)^2\right] - (\mathbb{E}[Y^j])^2}} \\ &= \frac{\frac{1}{\hat{\alpha}_1^{ij}} p^i - p^i p^j}{\sqrt{p^i(1-p^i)p^j(1-p^j)}},\end{aligned}$$

where the simplification is due to $\mathbb{E}\left[(Y^i)^2\right] = p^i$ and $\mathbb{E}\left[(Y^j)^2\right] = p^j$.

Task (CityScape)	STL	SubSel	MT-CRL	Meta-learning	Ours
Segmentation (mean IoU, ↑)	+2.1%	+3.7%	+4.2%	+4.6%	**+6.4**%
Depth (relative error, ↓)	+2.6%	+2.6%	+2.3%	+3.0%	**+5.2**%

Table 11. Experimental results on CityScape datasets.

D. Proof of Theorem 1

We have the change in correlation coefficients before and after resampling as

$$\begin{aligned}&\left|\rho_{Y^i,Y^j}-\rho_{Y'^i,Y'^j}\right|\\=&\left|\frac{\frac{1}{\hat{\alpha}_1^{ij}}p^i-p^ip^j}{\sqrt{p^i(1-p^i)p^j(1-p^j)}}-\frac{\frac{1}{\hat{\alpha}_1^{ij}}p^{'i}-p^{'i}p^{'j}}{\sqrt{p'^i(1-p'^i)p'^j(1-p'^j)}}\right|.\end{aligned} \tag{5}$$

Since $\rho_{Y^i,Y^j},\rho_{Y'^i,Y'^j}>0$, WLOG, consider downsampling on class 1 w.r.t. Y^i such that $p^{'i}+p^i\leq 1$ and $p^{'j}+p^j\leq 1$ such that $\rho_{Y'^i,Y'^j}\geq\rho_{Y^i,Y^j}$, we have Eq. (5) as

$$\begin{aligned}&\left|\rho_{Y^i,Y^j}-\rho_{Y'^i,Y'^j}\right|\\\geq&\frac{1+\alpha_1^{ij}}{\sqrt{p^ip^j(1-p^i)(1-p^j)}}\left|p^{'i}-\alpha_1^{ij}p^{'i}p^{'j}-p^i+\alpha_1^{ij}p^ip^j\right|\\=&\tilde{\mathcal{K}}\left|(1-\hat{\alpha}_0^{ij}\hat{\alpha}_1^{ij})(p^{'i}-p^i)+\hat{\alpha}_1^{ij}(\hat{\alpha}_1^{ij}-\hat{\alpha}_0^{ij})((p^i)^2-(p^{'i})^2)\right|,\end{aligned} \tag{6}$$

where the last equality is due to Eq. (4) and $\tilde{\mathcal{K}}=\frac{1+\alpha_1^{ij}}{\sqrt{p^ip^j(1-p^i)(1-p^j)}}$. Since $p^{'i}\leq p^i$ and $p^{'i}+p^i\leq 1$, we can further write Eq. (5) as

$$\begin{aligned}&\left|\rho_{Y^i,Y^j}-\rho_{Y'^i,Y'^j}\right|\\\geq&\frac{1+\alpha_1^{ij}}{\sqrt{p^ip^j(1-p^i)(1-p^j)}}(1-(\hat{\alpha}_1^{ij})^2)\left|p^i-p^{'i}\right|\\\geq&\ \frac{1}{\sqrt{(1-p^i)(1-p^j)}}(\frac{1}{\hat{\alpha}_1^{ij}}-\hat{\alpha}_1^{ij})\left|p^i-p^{'i}\right|\\=&\ \mathcal{K}\left|p^{'i}-p^i\right|,\end{aligned}$$

where $\mathcal{K}=\frac{(\frac{1}{\hat{\alpha}_1^{ij}}-\hat{\alpha}_1^{ij})}{\sqrt{(1-p^i)(1-(\hat{\alpha}_1^{ij}-\hat{\alpha}_0^{ij})p^i-\hat{\alpha}_0^{ij})}}$.

Furthermore, when $Y^i\perp\!\!\!\perp Y^j$, it is easy to see that $\mathbb{E}[Y^iY^j]-\mathbb{E}[Y^i]\mathbb{E}[Y^j]=\mathbb{E}[Y^i]\mathbb{E}[Y^j]-\mathbb{E}[Y^i]\mathbb{E}[Y^j]=0$, and the correlation coefficient remains zero before and after resampling. When $Y^i=Y^j$, we have $\mathbb{E}[Y^iY^j]-\mathbb{E}[Y^i]\mathbb{E}[Y^j]=p^i-(p^i)^2$, and the correlation coefficient remains 1 before and after resampling. □

E. Computational Cost

We include the training time of different methods on NYU-V2 dataset in Tab. 12. Our method leads to a relatively smaller increase in computational cost since we only finetune per-task perdictors, validating the scalability to large-scale datasets.

Method	Meta-learning	MT-CRL	SubSel	Ours
Time	1.33	1.45	1.37	1.13

Table 12. The training time of different methods relative to vanilla MTL.