APPENDIX

In this appendix, we include the following additional information, which we could not fit in the main paper due to space constraints:

- Additional implementation details
- More details on related work
- Empirical evidence on the relationship between confounding and correlation
- Ablation studies
- Counterfactual images of MNIST variants for various methods

A ADDITIONAL IMPLEMENTATION DETAILS

The generators used in CycleGAN are U-Networks Ronneberger et al. (2015) made of a 4-layer encoder and a 4-layer decoder. The discriminators are 4-layer encoders that output the probability of an image being a particular class. Since all the contrastive loss terms in Equation 4 work together towards removing the confounding effect of a confounding edge (unlike loss functions where each term has a different purpose). Hence, we use a common weighting term α in Equation 4. For all the experiments, batch size of 256 is used to train classifier (Equation 8). In all of the architectures, *leaky-relu* activation is used as an activation and *sigmoid* activation is used in the final layer to get probabilities. *Adam* optimizer is used in all the experiments.

B MORE ON RELATED WORK

In this section, we continue with the related work presented in the Section 2 of the main paper. Joshi & He (2022) discusses a potential issue for a counterfactual data augmentation method, viz.: if counterfactual data augmentation does not consider/augment counterfactuals w.r.t. all robust features that are spuriously correlated with non-robust features, then the performance of a model may drop in unseen distributions. To contrast this with our work, since we are able to quantify the confounding and hence correlation between any pair of generative factors, CONIC can generate all possible counterfactuals, which may in fact help in generate counterfactual images w.r.t. all robust/causal features. Hence, models trained on the counterfactual images generated using CONIC are more robust.

Idrissi et al. (2022) is similar to our work in performing data augmentation (for e.g., results on CelebA) with a difference that our method can be extended to the performance on the entire test set instead of on the worst group (e.g., MNIST results). Also, Hu & Li (2021) is similar in a sense to our work, but aims at controllable generation and counterfactual generation in the natural language setting. dif (2022) also generates counterfactual images but it assumes the availability of the data generation process and does not explicitly tackle confounding. Our work only assumes access to the attribute information but not the data generating process.

C RELATIONSHIP BETWEEN CORRELATION AND CONFOUNDING

Table 3 shows the empirical evidence that confounding is directly proportional to correlation between generative factors in CM-MNIST dataset.

Correlation Coefficient (color, digit)	Confounding (color, digit) (Defn 3)
0.10	0.072
0.20	0.249
0.50	1.244
0.90	3.585
0.95	4.041

Table 3: Relationship between correlation coefficient and confounding between color and digit in CM-MNIST dataset. Correlation is directly proportional to confounding.

D ABLATION STUDIES

In this section, we present the results on some ablation studies to understand the usefulness of the proposed regularizers. Without the additional regularizers in Eqn 4, the accuracy on the downstream classifier on CelebA is 73.69 ± 1.10. However, using the additional regularizer, the accuracy improves upto $\mathbf{79.56 \pm 1.28}$. The additional contrastive loss in Eqn 8 brings a slight improvement over Eqn 7. In CelebA experiments, while accuracy obtained when using Eqn 7 is around $78.73 \pm 1.22\%$ but using Eqn 8, the accuracy improved to $\mathbf{79.56 \pm 1.28}$.

Also, to understand the performance of ERM model using only the 5% unconfounded data, we experimented on MNIST variants and observed that the accuracy on CM-MNIST, DCM-MNIST, and WLM-MNIST are 34.39 ± 0.02, 17.22 ± 0.02, 17.72 ± 0.05, respectively. These results are worse than ERM model trained on entire training dataset (Table 1). We believe that the reason for these poor results by ERM model is due to the presence of multiple confounders. In MNIST variants, along with the confounding between color and digit, there is another feature called thickness that is challenging to learn, especially when the digits are thin (Figure 3 left). When we take only 5% unconfounded data, the train set size is very small, with many thin digits, making it difficult for ERM to learn the features.

E COUNTERFACTUAL IMAGES BY VARIOUS METHODS

Figure 5 and 6 shows the counterfactual images generated by various methods on MNIST variants.

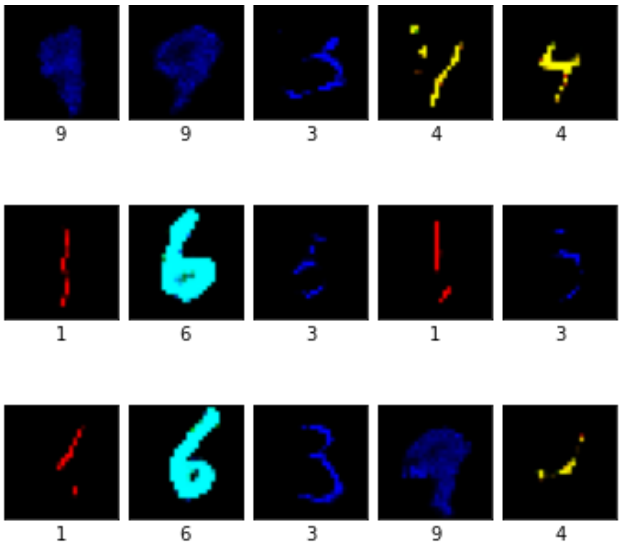

Figure 5: Conditional GAN generations and their conditioned value on CM-MNIST dataset. Because of extreme confounding, digit and shape are not de-confounded by Conditional GAN model.

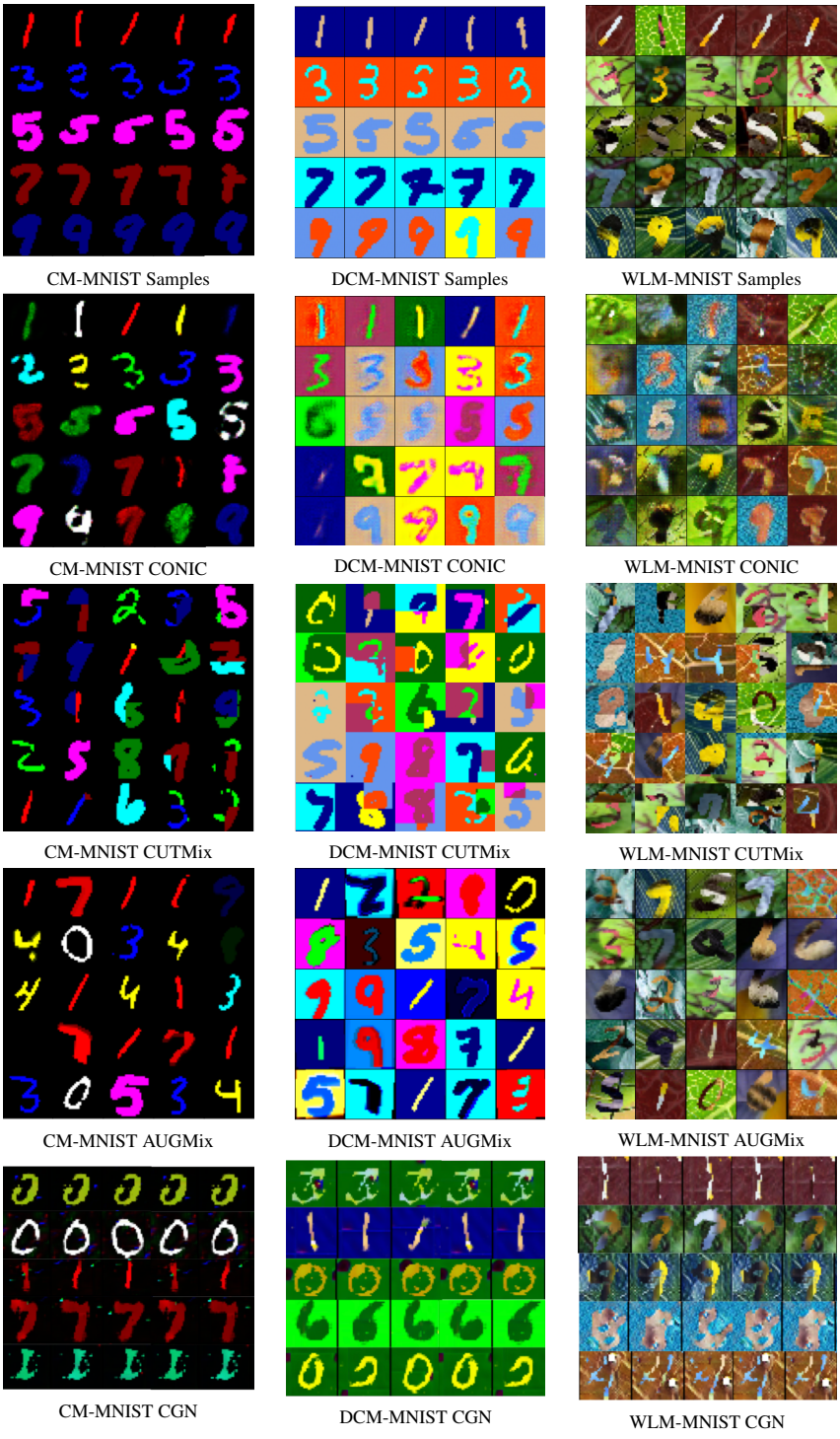

Figure 6: Sample images from MNIST variants and augmented images by various methods.