R1: Motivation Learning disentangled representations in cross-domains is useful for real-world problems such as image translation (demonstrated in the paper) and language translation. CdDN is one of the recent promising work on the cross-domain disentanglement task. However, it is a GAN-based architecture with the gradient reversal layer, which is not ideal for training in our opinion. Our work, IIAE, has a much simpler architecture with a more direct training scheme. The advantage of IIAE can be appreciated by the quality of results, compared to CdDN in Tables 6 and 7.

R1: Learned representations for image retrieval We think we can still check whether the learned representations are disentangled in the image retrieval task. In Table 2, we also report the retrieval accuracy using the exclusive representation (numbers in the parenthesis), which is closed to a random guess (100/N%) showing the successful disentanglement. In contrast, the results from CdDN are noticeably high or low, suggesting that it was relatively unsuccessful in disentangling the representations. Please see below for additional experiments.

R1: Clarification on the data We assume that the pairing is not unique, as in the CdDN paper.

R1: Quantitative evaluation We report quantitative evaluation on the quality of samples, as request by R1. We followed the exact experimental setting for the Cars dataset as in [12], except we use freshly generated training data (the data from [12] was unavailable) and the updated version of the evaluation metric LPIPS. Thus, please understand that the numbers here do not exactly match those in [12]. The results show that the sample quality of IIAE clearly exceeds the quality of GAN-based methods.

Translation	pix2pix [23]	CdDN [12]	IIAE
$X \to Y$	0.24987 ± 0.00780	0.23517 ± 0.00799	$\mathbf{0.21478} \pm 0.00844$
$Y \to X$	0.21524 ± 0.00704	0.19295 ± 0.00687	$\mathbf{0.15277} \pm 0.00774$

R2: Limitations For Cars and Sketchy datasets, we randomly paired samples within categories, which is a straightforward way to use our method for unpaired samples. Extending to semi-supervised learning tasks and scaling to multiple domains remain as future work. As for the quantitative comparison with SOTA, please see our response to R1 above.

R2: Correctness of the lower bound optimization Our training objective is II minus MI, whose lower bound (ELBO with regularization) is derived taking the standard steps for obtaining variational lower bounds. Thus, this lower bound has the same tightness property as ELBO and VIB.

Please note that our approach is independent of the choice of the prior, although all the experiments used the Gaussian prior for the simplicity in the implementation. Yet, in all of our experiments, we were not able to observe any of the sample diversity issues even under the Gaussian prior. Please refer to Table 1, 4, and 6 demonstrating that our method generates diverse samples depending on z^x and z^y.

R3: Comparison to TC regularization FactorVAE and β-TCVAE are for the *single-domain* disentanglement task, which minimize the total correlation (TC) among all dimensions of the latent variable to make them independent. The cross-domain disentanglement aims to decompose domain-specific and domain-invariant factors of variation into three latent variables (one shared and two exclusives for two domains). Minimizing TC is not directly applicable to cross-domain disentanglement. To the best of our knowledge, our work is the first to introduce the notion of interaction information for the cross-domain disentanglement task.

R3: Ablation Study We conducted ablation study on the effect of terms in the IIAE objective, using the ZS-SBIR dataset. II represents optimizing interaction information only, and II-MI is the objective in (11). Last two columns represent taking weighted sum with the ELBO, treating $\lambda = 2$ as the hyperparameter. The final column is the objective of IIAE. Comparing to Table 3, all settings significantly outperform SOTA, and shows that subtracting MI from II always help.

Metric	II	II-MI	ELBO+λII	ELBO+λ(II-MI)
mAP	0.517	0.534	0.516	**0.573**
P@100	0.605	0.616	0.595	**0.659**

R3: Additional comments on Table 9 Please note that we re-trained DRIT using the *paired* data in order to make a fair comparison (stated in the text), via minor modification to the author's code to take advantage of the paired data. Regarding the numbers from the Facades dataset, they are different from the original paper since we used test set rather than the validation set (stated in the footnote). The table on the right shows the result on the validation set, which matches the numbers in the original paper. Finally, thank you very much for catching typos, which will be fixed in the final version of the paper.

Facades(Val)	BicycleGAN	CdDN	IIAE
F $\to$ L (%)	-	95.0 (1.0)	**100.0** (1.0)
L $\to$ F (%)	45.0	97.0 (1.0)	**100.0** (0.0)