We thank all the reviewers for their helpful comments. We first response to some key concerns and the others will be addressed point by point. **Confusions of the LPIPS metric.** We apologize for the lack of explanation of LPIPS metric. Here LPIPS is the diversity metric that measures the perceptual difference of generated images. This diversity metric was proposed by BicycleGAN and adopted by DRIT and MUNIT. Specifically, we compute the average LPIPS distance between pairs generated by the same input image and different sampled styles. 10 image pairs are generated for each test image. Thus the higher the LPIPS, the better the diversity. We will include the above explanation to improve our work. **The reason for performance improvement.** We would like to clarify that FID and LPIPS are metrics for quality and diversity, respectively. In this work, we argue that learning the styles among different domains results in more diverse sampling space than that learning from one specific domain. From Tab. 2, we observer that DMIT-based models have a significant improvement of diversity over the multi-modal baselines when there is a T-path to encourage cross-domain translation. It suggests that the supervision from multi-domain is beneficial to multi-modal translation. But improving diversity does not result in the improvement of FID, since artifacts may be introduced. Thus the capacity of the discriminator is important for producing realistic images. Please refer to line 216-222 for more analysis. **The performance gap between different tasks.** In season transfer, the main difference between DMIT and baselines is that DMIT aligns the styles among different domains. So there is a significant improvement in diversity. In semantic image synthesis, previous works focus on modeling the foreground and background separately in terms of training losses. Without reasonable representations, these methods are difficult to produce high-quality images. By learning disentanglement, we observe that the style $\mathcal{S}$ is associated with background and the content $\mathcal{C}$ is related to the foreground. The disentangled representations enable DMIT to perform finer manipulation and achieve better results than the baselines.

To Reviewer #1: Number of domains. In addition to facial attribute transfer, semantic image synthesis contains more than two domains as we introduced at line 97-100 and 179-181 of the paper. Since we treat the image set with the same text description as an image domain, there are countless domains. **Compare with StarGAN on CelebA.** As shown in Fig. A (a), all of the methods can produce images that correspond to expected attributes. But the styles of images generated by StarGAN are monotonous, despite the injection of noise vector. The quantitative results also confirm our observation. We will include more comparisons in the supplementary.

To Reviewer #2: How does Eq.(2) help to disentangle different domain styles? Eq.(2) encourages to minimize the mutual information of $\mathcal{X}$ and $\mathcal{S}$ (refer to [1] in the paper). Thus E_s is enforced to model the efficient disentangled representations. Besides, note that we assume the styles among different domains can be aligned (*e.g.*summer nightfall and winter nightfall), which suggests that the representations are domain invariant. To achieve this goal, we utilize a unified (weight sharing) encoder E_s to map images of different domains onto the same space. Thus similar images will have similar representations. But only sharing the mapping function cannot guarantee to eliminate the distribution shift of representations among different domains. Therefore, we encourage the style representations of all domains to be as close as possible to the same distribution to eliminate the domain bias. **Why does DMIT need the encoder E_d?** Combined with the above analysis, since we eliminate the domain-specific information of $\mathcal{S}$, we need the domain label to indicate the mapping of the target domain. **Why is there only one generator?** Previous methods do not have aligned styles, so they need multiple domain-specific generators. Our method assumes that both $\mathcal{C}$ and $\mathcal{S}$ can be shared among different image domains, so we can use one generator to perform multi-mapping translation. **Why does DMIT w/o D-Path achieve the best LPIPS score?** Without D-Path, DMIT cannot learn effective representations and produces blurry images. Although the artifacts produce meaningless diversity (LPIPS), the quality of generated images is poor. Without T-Path, DMIT lacks incentives for the use of styles and produces monotonous images that only a subset of real data. Combining both paths allows DMIT to learn effective representations for diverse cross-domain translation.

To Reviewer #3: Can DMIT perform content transfer? Yes. We have evaluated DMIT on three additive facial attributes: facial hair, glasses, and smile. As shown in Fig.A (b), we observe that DMIT can add or remove specified facial attributes arbitrarily. **Limitations and future works.** Although DMIT can perform the content transfer, we observe that the style representations tend to model some global properties rather than specific contents, *e.g.*skin color and scene lighting. We agree that the problem is caused by spatial pooling used in E_s, as discussed in ContentDisentanglement[1]. To verify the above conjecture, we construct a simple variant of DMIT (DMIT-CD) according to ContentDisentanglement. As shown in Fig.A (c), although there is still room for improvement, DMIT-CD has great potential for multi-domain content transfer. Besides, we observe that the convergence rates of different domains are generally different, *e.g.*adding glasses is more difficult than changing hair color. Thus a domain-adaptive learning strategy may help to improve training stability and performance. We will include these valuable discussions in our work.

(a) Comparison with StarGAN. The styles of StarGAN* and DMIT are sampled from random noise.

(b) Content transfer.

(c) Content guided transfer.

Figure A: Visual and quantitative results of facial attribute transfer.

[1]Emerging Disentanglement in Auto-Encoder Based Unsupervised Image Content Transfer. Press et al. ICLR 2019.