Thanks all the reviewers for acknowledging our contributions and their valuable comments.

To Reviewer #1 Q1: More modalities? **R1**: Thanks. We add experiments on the NJU-ID (different resolutions) [1] and the Multi-PIE (different poses) [2] datasets. Details of these datasets are introduced as follows.

• The NJU-ID dataset consists of 256 identities with one ID card image (102×126 resolution) and one camera image (640×480 resolution) per identity. Considering the few number of images in the NJU-ID dataset, we use our collected ID-Photo dataset (1000 identities) as the training set and the NJU-ID dataset as the testing set.
• The Multi-PIE dataset contains 337 persons with different poses. We use profiles ($\pm 75^o$, $\pm 90^o$) and frontal faces as different modalities. 200 persons are used as the training set and the rest 137 persons are the testing set.

The examples of dual generation are shown in Fig. 1. For the recognition performance, on the NJU-ID dataset, we improve Rank-1 by 5.5% (DVG 96.8% - Baseline 91.3%) and VR@FAR=1% by 6.2% (DVG 96.7% - Baseline 90.5%) over the baseline LightCNN-29. On the Multi-PIE dataset, the Rank-1 of $\pm 90^o$ and $\pm 75^o$ is increased by 18.5% (DVG 83.9% - Baseline 65.4%) and 4.3% (DVG 97.3% - Baseline 93.0%), respectively. All experiments demonstrate the effectiveness of our method in other modalities.

Q2: More ablations? **R2**: For the generation model, the ablations of $\mathcal{L}_{\text{dist}}$, $\mathcal{L}_{\text{ip}}$ and $\mathcal{L}_{\text{div}}$ have been reported in Table-1. We add the ablation of $\mathcal{L}_{\text{adv}}$ in Eq. (10). That is, on the CASIA NIR-VIS 2.0 dataset, the Rank-1 decreases 0.5% if $\mathcal{L}_{\text{adv}}$ is not used. For the recognition model, the effect of $\mathcal{L}_{\text{pair}}$ in Eq. (13) can be found in Table-2 ('+DVG' means using $\mathcal{L}_{\text{pair}}$). All ablations reveal that each component of our method is useful. Especially for $\mathcal{L}_{\text{ip}}$, $\mathcal{L}_{\text{dist}}$ and $\mathcal{L}_{\text{pair}}$, the Rank-1 decreases 5.5%, 4.9% and 2.1% respectively on the ablations. Moreover, our method is not sensitive to the trade-off parameters in a large range. Please see Reviewer-2' R2 for details.

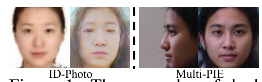

Figure 1: The examples of dual generation on the ID-Photo and the Multi-PIE datasets.

To Reviewer #2 Q1: The relationship between DVG and PIM? Some related works? **R1**: Thanks. The noise in PIM is to help recover invisible details. The generated 'many' faces are required to be consistent with one ground truth. Hence, PIM is still a conditional image-to-image translation method. As mentioned in the introduction, it faces diversity and uniqueness limitations. Differently, our method belongs to unconditional generation. That is, we generate diverse new paired faces from noise, which alleviates the above two limitations. We will cite these related works in our paper.

Q2: How to assign hyper-parameters? A sensitivity analysis? **R2**: The hyper-parameters are set by balancing the magnitude of each loss function. Fig. 2 presents the sensitivity studies of λ_1, λ_2 and λ_3 in Eq. (10). For α_1 in Eq. (13), when α_1 is set to 0.0025, 0.005, 0.01, 0.02 and 0.04, the Rank-1 is 98.9%, 99.1%, 99.2%, 99.2% and 98.8%, respectively. We can observe that our method is not sensitive to these hyper-parameters in a large range. For example, the Rank-1 only decreases 0.3% when λ_1 changes from 0.1 to 0.4.

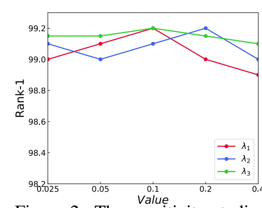

Figure 2: The sensitivity studies of trade-off parameters on the CASIA NIR-VIS 2.0 dataset. The backbone is LightCNN-9.

Q3: Complexity? **R3**: Thanks. Our method is computationally efficient. For instance, when using one Titan XP, training the generation model on the CASIA NIR-VIS 2.0 dataset spends 3 hours. Meanwhile, in the inference stage, generating a pair of heterogeneous faces only needs 3.2 ms. Furthermore, training the HFR network spends 1 hour.

Q4: Releasing the generated faces and the writing suggestions. **R4**: Thanks. We will release our codes and the generated data. The writing of our paper has been carefully revised according to your advice.

To Reviewer #3 Q1: How to tune these three parameters in Eq. (10)? **R1**: The trade-off parameters in Eq. (10) are tuned by balancing the magnitude of each loss function. In addition, the sensitivity studies of the trade-off parameters λ_1, λ_2 and λ_3 in Eq. (10) are shown in Fig. 2. We can observe that our method is not sensitive to these trade-off parameters in a large range. For instance, when λ_1 changes from 0.1 to 0.4, the Rank-1 only decreases 0.3%.

Q2: It is suggested to report the time cost. **R2**: Thanks for your advice. Our proposed framework is computationally efficient. For example, when using one Titan XP, training the generation model on the CASIA NIR-VIS 2.0 dataset only needs 3 hours. Meanwhile, generating a pair of heterogeneous faces in the inference stage needs 3.2 ms. Moreover, training the HFR network needs 1 hour.

Q3: Apply to other heterogeneous recognition problems? **R3**: We add experiments on other two datasets, including the NJU-ID (different resolutions) [1] and the Multi-PIE (different poses) [2] datasets. The results show that our method can be effectively applied to more modalities. Please see Reviewer-1' R1 for experimental details. Due to the limited time, we will explore other heterogeneous recognition tasks in the future work.

References: [1] Huo et al. Heterogeneous face recognition by margin-based cross-modality metric learning. IEEE Transactions on Cybernetics 2018. [2] Gross et al. Multi-PIE. Image and Vision Computing 2010.