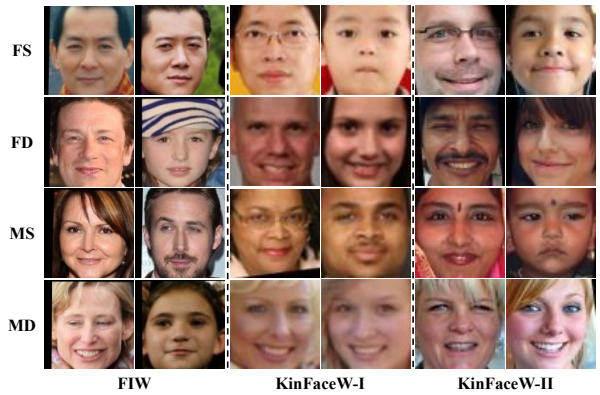
GeneFlow: Modeling Heredity and Variation via Flow Matching Transformers for Kinship Verification

Supplementary Material

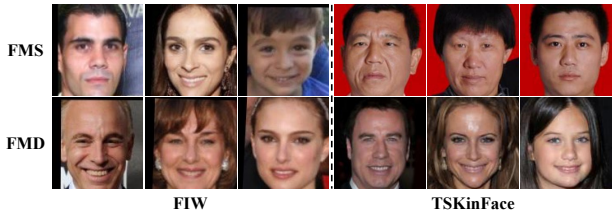
A. Samples of the Kinship Dataset

In the training of kinship verification (task1), we use three kinship datasets: FIW, KinFaceW-I, and KinFaceW-II. The FIW dataset contains seven relationship types—FS (Father–Son), FD (Father–Daughter), MS (Mother–Son), MD (Mother–Daughter), BB (Brother–Brother), SS (Sister–Sister), and SIBS (S iblings)—as illustrated in Figs. 1a and 1b. Both KinFaceW-I and KinFaceW-II include four relationship types—FS (Father–Son), FD (Father–Daughter), MS (Mother–Son), and MD (Mother–Daughter)—as shown in Fig. 1a. For the training of tri-subject verification (task2), we use two datasets: FIW and TSKinFace, each containing two relationship types—FMS (Father–Mother–Son) and FMD (Father–Mother–Daughter)—illustrated in Fig. 1c.



(a) FS, FD, MS and MD kin relations in FIW, KinFaceW-I and KinFaceW-II databsets.

(b) BB, SS and SIBS kin relations in FIW dataset.



(c) FMS and FMD tri-subject kin relations in FIW and TSKinFace databsets.

Figure 1. Examplar images from FIW, KinFaceW-I, KinFaceW-II and TSKinFace databsets.

B. Vector Field Predictor Architecture

In this work, we adopt the DiT [36] architecture as the backbone of our vector field predictor for flow matching in latent facial gene space. As show in Fig. 2, the only modification is that the input and output dimensions are adjusted to match the facial gene dimension. We also adopt the conditioning configuration to accommodate different kinship verification tasks. Finally, the global genes of the parents and the timestep t are separately embedded and then summed as the conditioning vector, which is injected into each DiT block via AdaLN, following the original DiT design.

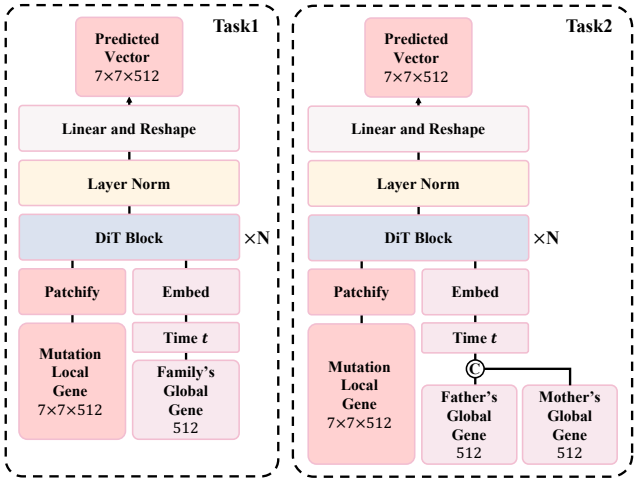


Figure 2. The vector field predictor architecture. The **left** architecture corresponds to kinship verification (task1), where the global gene of a single family member (e.g., father, mother, brother, or sister) is used as the conditioning input. The **right** architecture corresponds to tri-subject verification (task2), where the global genes of both parents are concatenated and used jointly as the conditioning input.

C. Sensitivity Analysis

In this section, we analyze two hyper-parameters used in our model. The first one is the proportion r of gene mutations to control the degree of gene mutation. The other is the inference steps to decide how many steps are used to generate the child's local genes.

C.1. Effect of Mutation Proportion

To enhance genetic diversity, we design a gene mutation process controlled by the mutation ratio r, which deter-

mines the degree of genetic variation in the generated descendants during training. We evaluate the effect of r on kinship verification by computing the tri-subject verification accuracy on the FIW dataset under different settings. As shown in Fig. 3, the best performance is achieved when $r = 0.6$. This indicates that a moderate mutation ratio provides a better balance between genetic inheritance and diversity, leading to improved kinship verification performance.

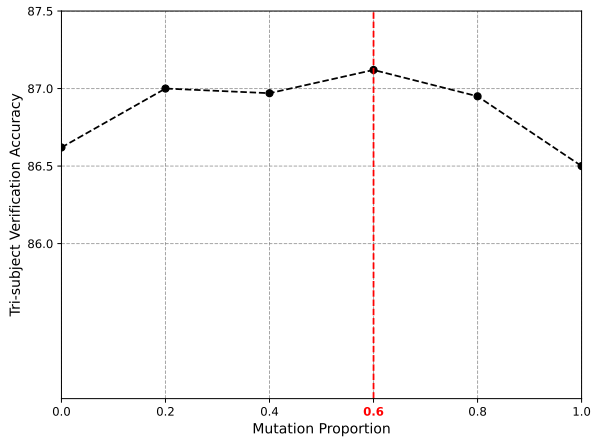


Figure 3. The effect of mutation proportion for tri-subject verification.

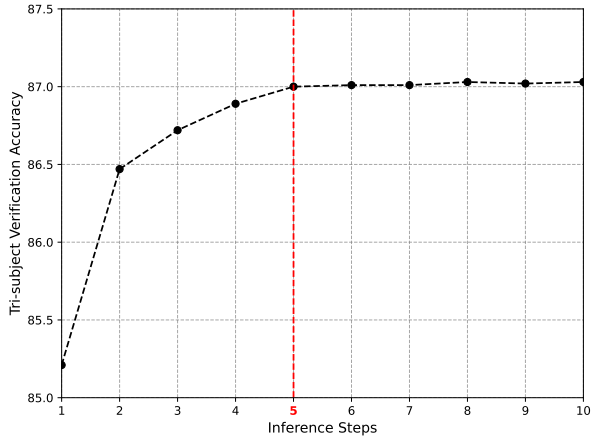


Figure 4. The effect of inference steps for tri-subject verification.

C.2. Effect of Inference Steps

Increasing the number of inference steps in the flow matching model typically improves the quality of the generated results, at the expense of higher computational cost and longer inference time. So we perform an experiment to evaluate the tri-subject verification performance on FIW with different numbers of inference steps. As shown in Fig. 4, the accuracy increases with the number of inference steps. The improvement is significant when the number of steps is fewer than 5, but the growth slows down beyond 5 steps. Therefore, considering the trade-off between accuracy and inference time, we set the number of inference steps to 5 for kinship verification.

D. Computation Analysis

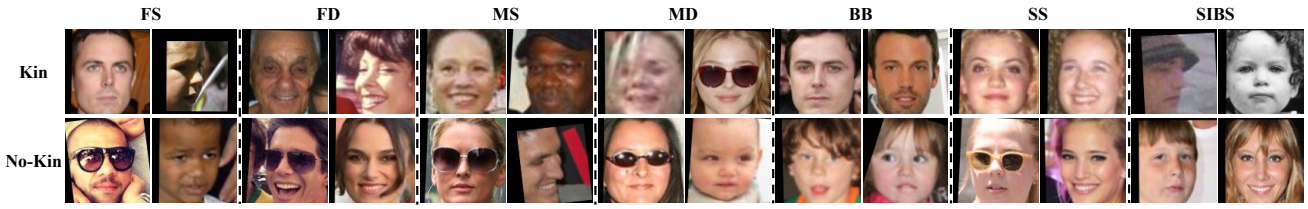
Tab. 1 compares the performance and computational efficiency of our model on tri-subject verification of the FIW dataset, and also reports the computational cost of existing open-source kinship face synthesis models. Since the training code for these synthesis models is not publicly available—and integrating them into a kinship verification framework would require retraining—we do not include their verification performance in our comparison. Introducing flow-matching–based generative modeling into a kinship verification system inevitably increases computation and inference latency. However, compared with directly incorporating existing kinship face synthesis models, the additional overhead introduced by GeneFlow is acceptable and consistently yields improved verification performance.

Table 1. The computational complexity and performance comparison of *Tri-subject Verification* on FIW dataset. The **Time** in the table refers to the time it takes to complete all samples in the test set.

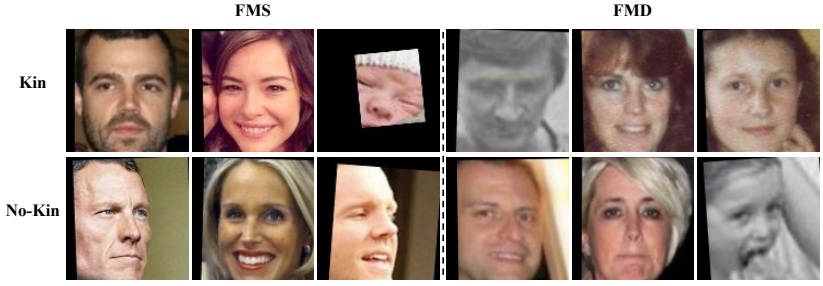
Model	**Params (M)**	**FLOPs (G)**	**Time (s)**	**ACC (%)**
FaCoRNet [44]	72.56	48.62	25	85.1
GeneFlow(Ours)	130.72	76.66	32	87.1
	+58.16	+28.04	+7	+2
StyleDNA [27]	75.32	140.64	-	-
ChildPredictor [50]	114.41	52.69	-	-
StyleGene [21]	356.67	274.92	-	-

E. Limitations and Future Work

Although GeneFlow achieves strong performance, several limitations remain. As shown in Figs. 5a and 5b, the model is more likely to produce errors on low-resolution, heavily occluded, or otherwise degraded facial images. This suggests that its robustness to real-world image quality variations is still limited. In future work, we plan to pretrain GeneFlow on large-scale face datasets with diverse imaging conditions, enabling the model to acquire implicit facial restoration and normalization capabilities. Another limitation is the reliance on visual information alone. To further improve prediction reliability—especially in challenging or ambiguous cases—we aim to incorporate complementary multimodal cues such as gender, age, and textual attributes.



(a) Mispredictions on hard samples in kinship verification on FIW.



(b) Mispredictions on hard samples in tri-subject verification on FIW.

Figure 5. Mispredictions on FIW dataset. The **Kin** row shows true kin pairs misclassified as non-kin (false negatives), while the **No-Kin** row shows non-kin pairs misclassified as kin (false positives).