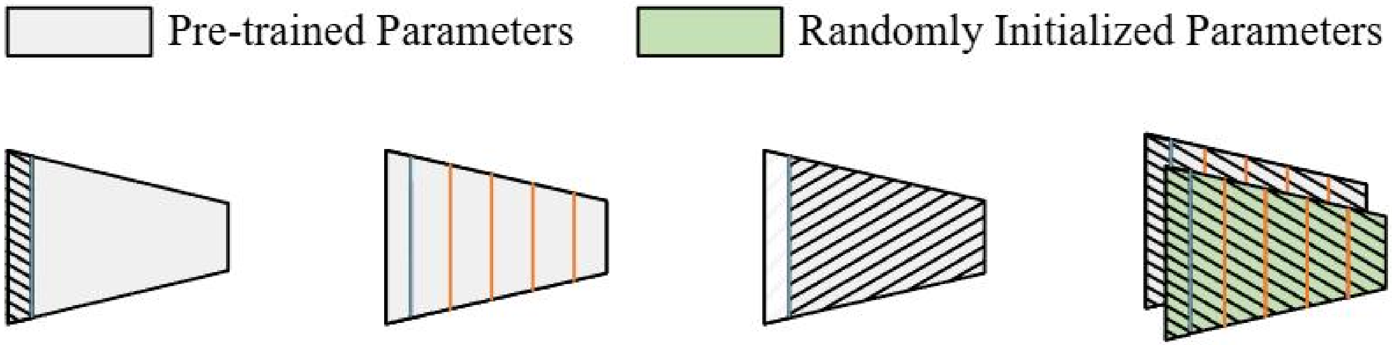
Improving the Transferability of Adversarial Attacks on Face Recognition with Diverse Parameters Augmentation

Fengfan Zhou[1], Bangjie Yin[2], Hefei Ling[1]*, Qianyu Zhou[3], Wenxuan Wang[4]
[1]School of Computer Science and Technology, Huazhong University of Science and Technology;
[2]Shanghai Shizhuang Information Technology Co., Ltd;
[3] Department of Computer Science and Engineering, Shanghai Jiao Tong University;
[4] School of Computer Science, Northwestern Polytechnical University.
[1]{ffzhou, lhefei}@hust.edu.cn, [2]jamesyin10@gmail.com,
[3]zhouqianyu@sjtu.edu.cn, [4]wxwang@nwpu.edu.cn

Motivation

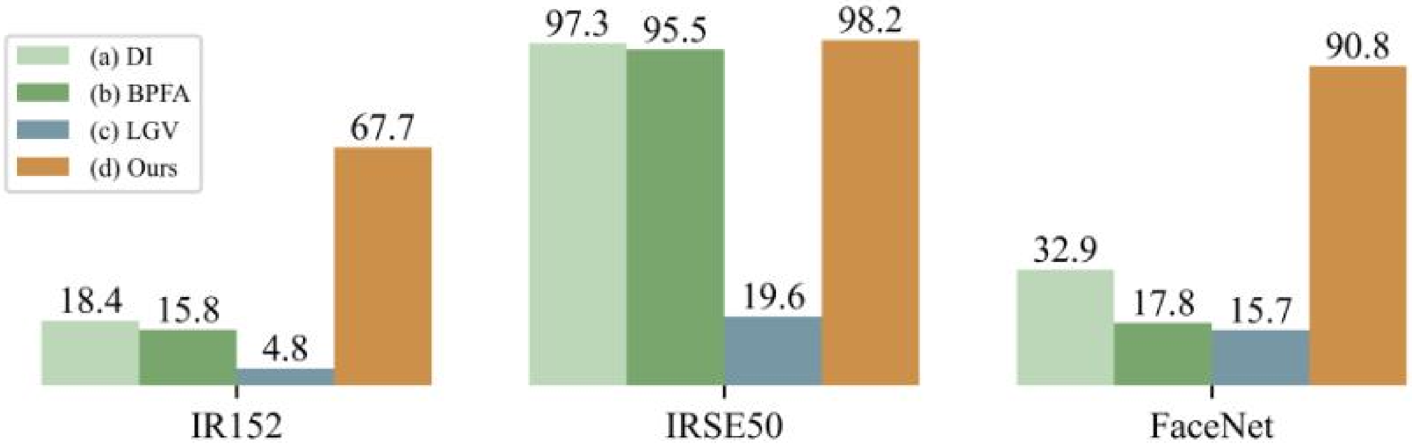


(a) Input-based Augmentation

(b) Feature-based Augmentation

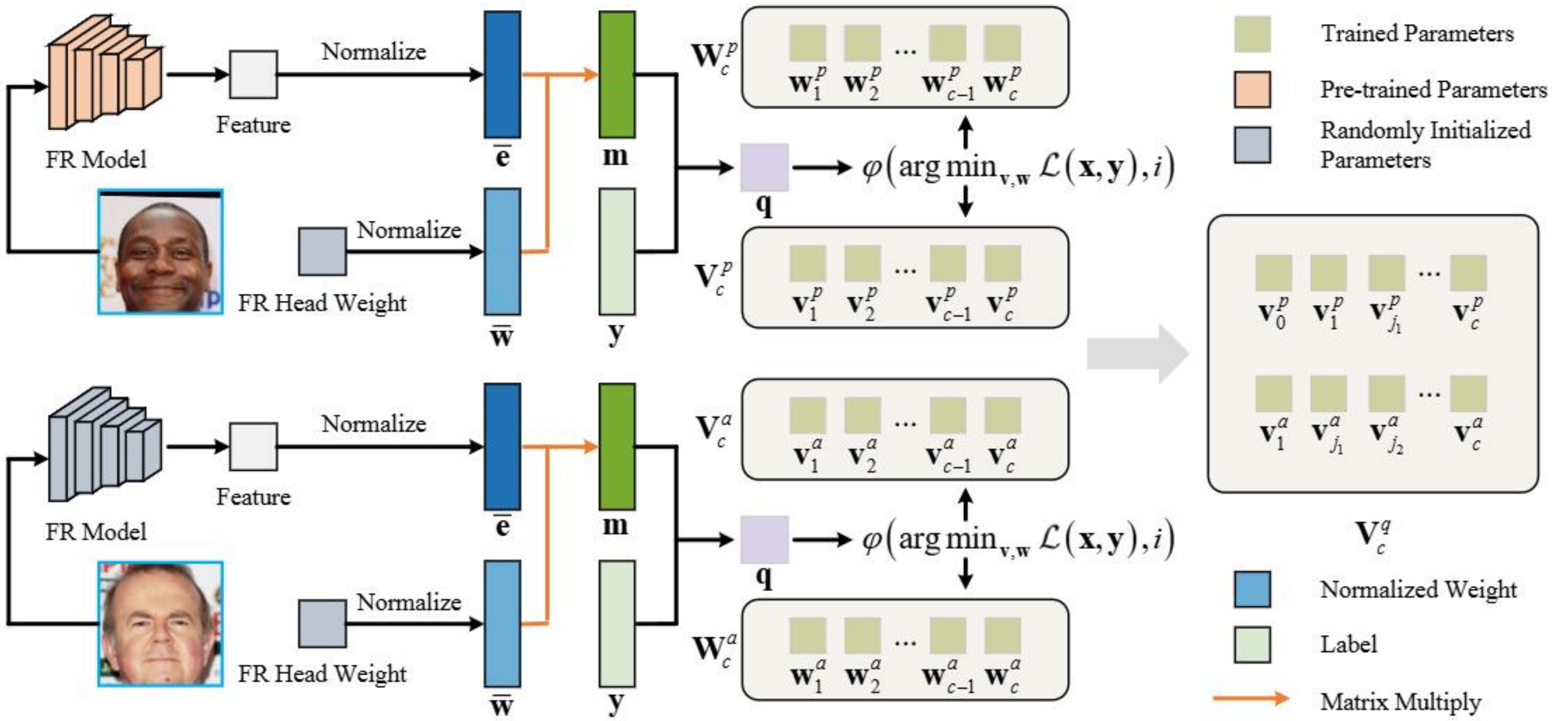
(c) Parameter-based Augmentation

(d) Ours



Methodology

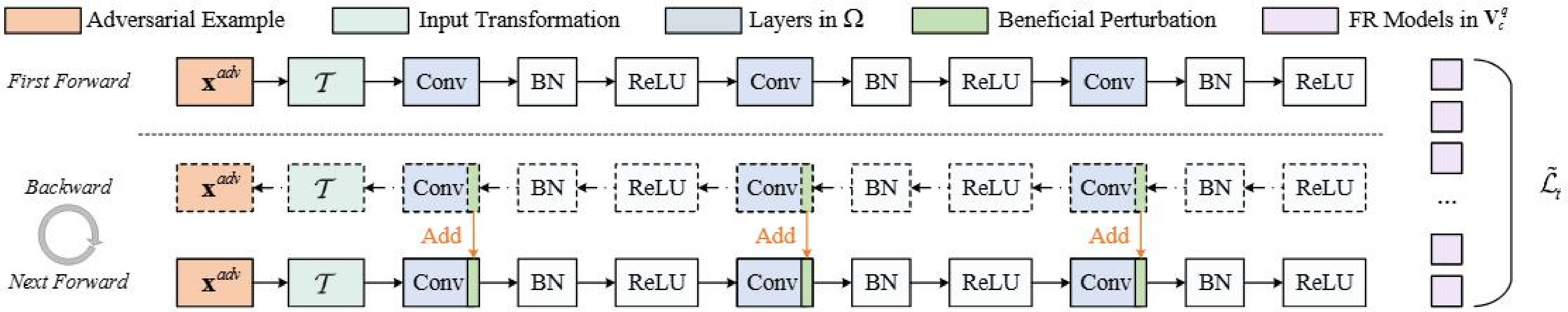
Diverse Parameters Optimization


FR Model
Feature
Normalize
$\bar{\mathbf{e}}$
$\mathbf{m}$
FR Head Weight
Normalize
$\bar{\mathbf{w}}$
$\mathbf{y}$
$\mathbf{W}_c^p$
$\mathbf{w}_1^p$ $\mathbf{w}_2^p$ … $\mathbf{w}_{c-1}^p$ $\mathbf{w}_c^p$
$\mathbf{q}$
$\varphi\left(\arg\min_{\mathbf{v},\mathbf{w}} \mathcal{L}(\mathbf{x},\mathbf{y}), i\right)$
$\mathbf{V}_c^p$
$\mathbf{v}_1^p$ $\mathbf{v}_2^p$ … $\mathbf{v}_{c-1}^p$ $\mathbf{v}_c^p$
FR Model
Feature
Normalize
$\bar{\mathbf{e}}$
$\mathbf{m}$
FR Head Weight
Normalize
$\bar{\mathbf{w}}$
$\mathbf{y}$
$\mathbf{V}_c^a$
$\mathbf{v}_1^a$ $\mathbf{v}_2^a$ … $\mathbf{v}_{c-1}^a$ $\mathbf{v}_c^a$
$\mathbf{q}$
$\varphi\left(\arg\min_{\mathbf{v},\mathbf{w}} \mathcal{L}(\mathbf{x},\mathbf{y}), i\right)$
$\mathbf{W}_c^a$
$\mathbf{w}_1^a$ $\mathbf{w}_2^a$ … $\mathbf{w}_{c-1}^a$ $\mathbf{w}_c^a$
Trained Parameters
Pre-trained Parameters
Randomly Initialized Parameters
$\mathbf{v}_0^p$ $\mathbf{v}_1^p$ $\mathbf{v}_{j_1}^p$ … $\mathbf{v}_c^p$
$\mathbf{v}_1^a$ $\mathbf{v}_{j_1}^a$ $\mathbf{v}_{j_2}^a$ … $\mathbf{v}_c^a$
$\mathbf{V}_c^q$
Normalized Weight
Label
Matrix Multiply

We enhance the diversity of the surrogate model parameters by integrating both pre-trained and random initializations. The method yields a diverse set of surrogate model parameters, which enhances the parameter diversity of the surrogate FR models and consequently improves transferability of the crafted adversarial examples.

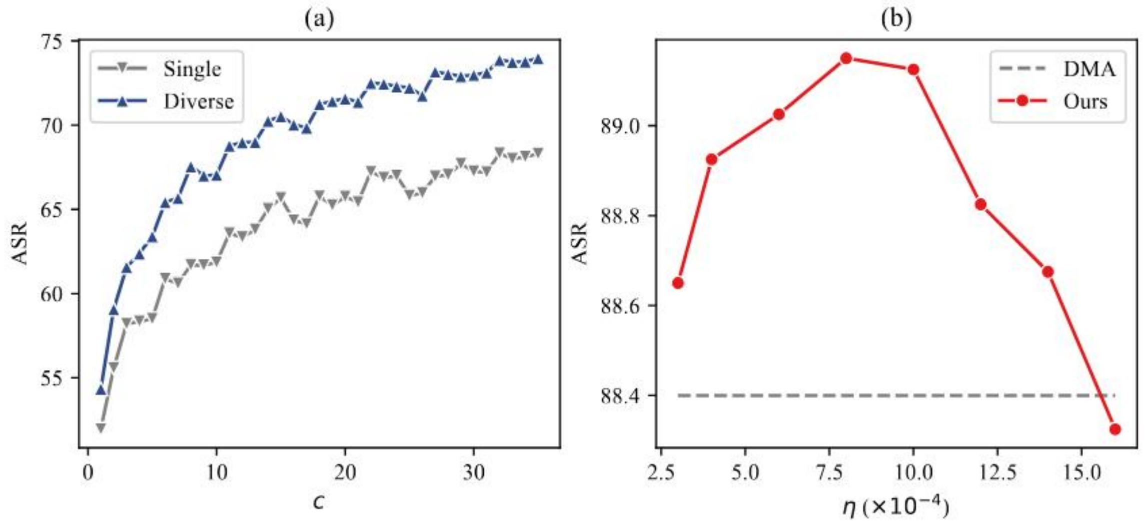
Methodology

Hard Model Aggregation

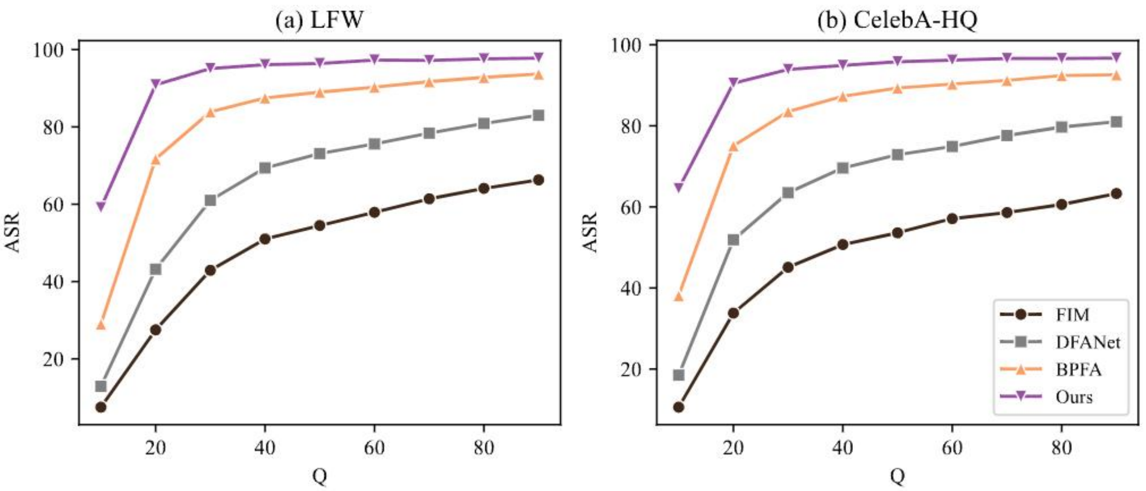


After acquiring a surrogate model set with diverse parameters, we introduce beneficial perturbations with the optimization direction opposite to that of adversarial perturbations onto the feature maps of these diversified surrogate models, transforming them into hard models and aggregate the hard models to increase the transferability.

Experiments


(a)

(b)


(a) LFW

(b) CelebA-HQ

Thank you