Escaping Data Scarcity for High-Resolution Heterogeneous Face Hallucination

Yiqun Mei*, Pengfei Guo*, Vishal M. Patel
Johns Hopkins University

2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 978-1-6654-6946-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/CVPR52688.2022.01812

Abstract

In Heterogeneous Face Recognition (HFR), the objective is to match faces across two different domains such as visible and thermal. Large domain discrepancy makes HFR a difficult problem. Recent methods attempting to fill the gap via synthesis have achieved promising results, but their performance is still limited by the scarcity of paired training data. In practice, large-scale heterogeneous face data are often inaccessible due to the high cost of acquisition and annotation process as well as privacy regulations. In this paper, we propose a new face hallucination paradigm for HFR, which not only enables data-efficient synthesis but also allows to scale up model training without breaking any privacy policy. Unlike existing methods that learn face synthesis entirely from scratch, our approach is particularly designed to take advantage of rich and diverse facial priors from visible domain for more faithful hallucination. On the other hand, large-scale training is enabled by introducing a new federated learning scheme to allow institution-wise collaborations while avoiding explicit data sharing. Extensive experiments demonstrate the advantages of our approach in tackling HFR under current data limitations. In a unified framework, our method yields the state-of-the-art hallucination results on multiple HFR datasets.

1. Introduction

Deep convolutional neural networks have led to unprecedented success on visual face recognition [5, 58, 61, 66], where state-of-the-art methods achieve more than 99% accuracy on multiple benchmarks. These near-perfect performances come from both well-elaborated architectures and exhaustive training on massive datasets. Nevertheless, in many real-world scenarios with low-visibility, such as low-light and night-time, it is often infeasible to obtain clear visible (VIS) images. Under these circumstances, sensors deployed for other imaging spectra, e.g. Thermal (TH), can capture more discriminative information and serve as a more reliable solution. This raises a great need of heterogeneous face recognition (HFR) [32, 33, 40], an important task in computer vision and biometrics, that matches images from TH modality to its VIS counterpart [1]. The HFR problem has numerous applications in surveillance, monitoring and security.

Unfortunately, due to the large domain discrepancy, naively deploying a state-of-the-art face recognition algorithm trained on VIS images often leads to poor performance on a TH dataset [21]. Over the past decade, tremendous efforts have been spent to address the HFR problem by either learning domain-invariant features [10, 14, 15, 37] or finding a common subspace [28, 49, 59, 72]. Owing to the rapid progress in Generative Adversarial Networks (GANs) [16], most recent methods [7, 8, 12, 60, 73, 75] reformulate HFR as a face synthesis/translation problem. The resulting "recognition via hallucination" scheme embraces a huge benefit that any off-the-shelf recognizer can be directly applied on the synthesised images.

While these synthesis-based approaches fill the gap to some extent, the produced VIS images are still unsatisfactory, often accompanied with distorted and incorrect facial structures (shown in Figure 4), which significantly degrades the recognition accuracy. We found that the bottleneck is likely due to the limited size of the dataset which fails to offer sufficient information to guide image synthesis. Unlike visible images that are easy to obtain and widely available over the Internet [20], collecting and annotating a large-scale high-quality TH dataset is difficult. Challenges stem from many aspects. First, the acquisition process is both time-consuming and costly, which often requires laborious setup and non-trivial calibrations [52, 63]. Second, the diversity of collected data can be limited. Due to the physical constraints, it is typically infeasible for a single institution to collect a comprehensive dataset that covers a diverse set of identities with various attributes, such as race, gender and age. Most existing datasets [1, 9, 42, 47] are confined to a small number of subjects, leading to biased results and over-fitting. Third, face data is privacy sensitive. Since it contains subjects' personal identification information, one has to deal with privacy concerns when collecting and storing them, making it

*equal contribution

[1] Note that HFR is a general term that is used for matching two face images taken in two different domains such as TH, VIS or sketch. In this paper, we refer to HFR as a specific problem of matching TH images with VIS images.

978-1-6654-6946-3/22/$31.00 ©2022 IEEE
DOI 10.1109/CVPR52688.2022.01812

difficult to share the data with other institutions.

Besides poor synthesis quality, most existing methods can only process images at a resolution no more than 128×128. This not only leads to visually unappealing results, but also reduces their applicability in many downstream tasks that depend on high-resolution inputs, such as face parsing [38, 39], editing [74] and reenactment [53].

In this paper, we present a unified hallucination framework for HFR, that is capable of synthesizing high-resolution visible faces (512×512) from low-resolution heterogeneous data (i.e. smaller than 128×128), with superior realness and higher fidelity. Our approach consists of two separate strategies. The first one comes with a new generation paradigm inspired by the recent success in GAN inversion [68]. The core idea is to leverage rich and diverse facial priors from the visible domain to eschew the need of learning generation from scratch. This is accomplished by embedding a pre-trained GAN (e.g. StyleGAN [25, 26]) as a facial decoder which hallucinates visible faces conditioned on the latent representations of a U-shaped encoder. The encoder is carefully designed with a novel Multi-scale Contexts Aggregation (MSCA) mechanism which merges scale-wise information to enhance representation. MSCA offers better fine-grained generation control and is pivotal for preserving identity information. The proposed method, called ***Visual Prior enhanced GAN (VPGAN)***, can break the underlying data limitation by producing faces with state-of-the-art accuracy and photo-realism.

Deep models are data hungry and VPGAN may be further improved with large-scale training. However, in practice, HFR data tends to be separately collected and dispersed among different intuitions. Due to privacy concerns, one cannot simply share the data for centralized training. To this end, our second strategy introduces Federated Learning (FL) [29] to further improve HFR, which enables collaborative model training while avoiding explicit data exchanging. Specifically, we allow each institution to perform local training on their private HFR data and deploy a centralized server to periodically communicate with each institution, aggregating local models and updating a global model. The whole process does not involve any data transfer but benefits deep models a lot by integrating information from a significantly broader range of data. To make our approach more suitable for real-world HFR, we have to tackle the heterogeneous data distributions across institutions. This challenge, which may be caused by differences in sensor types or acquisition protocols, can lead to locally skewed updates, resulting in slow convergence and sub-optimal performance [31, 35]. To tackle this issue, we build our FL algorithm based on a new Model Proximity Regularization (MPR), which corrects local gradient updates by constraining the discrepancy between the latent representations from the global and local models. As a result, our approach achieves superior robustness towards non-ideal data distribution, implying its applicability for solving real-world HFR problems.

In our unified framework, we integrate VPGAN as the basic component and use it in the proposed FL scheme. We term this new framework VPFL. In the experiment section, we demonstrate that our approach generates VIS faces at high-resolution with superior realness and accuracy.

Within the infrared spectrum, various modalities have been explored for thermal-to-visible (TH-VIS) face recognition. These include Near Infrared (NIR), Short-Wave Infrared (SWIR), Mid-Wave Infrared (MWIR) and Long-Wave Infrared (LWIR). The use of a particular thermal modality depends on the application. For instance, in long-range surveillance applications, SWIR or MWIR modalities are often used. Unlike NIR images, which are close to the VIS spectrum images, LWIR images are often acquired in low-resolution with many facial details missing on the captured imagery. This is well-reflected by the performance drop of the existing recognition methods on such datasets [23, 51]. Also, greater than 99% accuracy has been reported on many NIR face datasets [8, 12]. However, the performance of various HFR methods on LWIR data is significantly low [23, 51].

In summary, the main contributions of our paper are:

- We propose a new data-efficient generation scheme for HFR. Thanks to the powerful visual priors, it manages to alleviate fundamental data challenges, resulting in superior hallucination results.
- We introduce a unified framework to make large-scale training possible in real-world scenarios. VPFL makes multi-institutional collaborations possible without raising any privacy concerns.
- Extensive experiments show that our method can produce faces with state-of-the-art photo-realism and fidelity, which in turn significantly boosts the recognition accuracy. These merits show its great potential to serve as a universal solution towards real-world applications.

2. Related Work

Heterogeneous Face Recognition. Compared to the methods relying on feature/subspace learning, recognition via synthesis [8, 11–13, 41, 75, 76] has received significant attention, because off-the-shelf face recognition algorithms can be applied to the synthesized images. Our discussion will focus on these types of methods. Early deep learning-based approaches directly learn a CNN for cross-spectral mapping. For example, Lezama *et al*. [30] train a CNN for NIR-VIS and improve results with low-rank assumption. Riggan *et al*. [56] enhance the discriminative quality of the synthesized images by modeling both global and local regions. Recent methods leverage GANs to improve the hallucination qualities. Zhang *et al*. [76] propose GAN-VFS that jointly learns semantic rich features and facial reconstruction, which re-

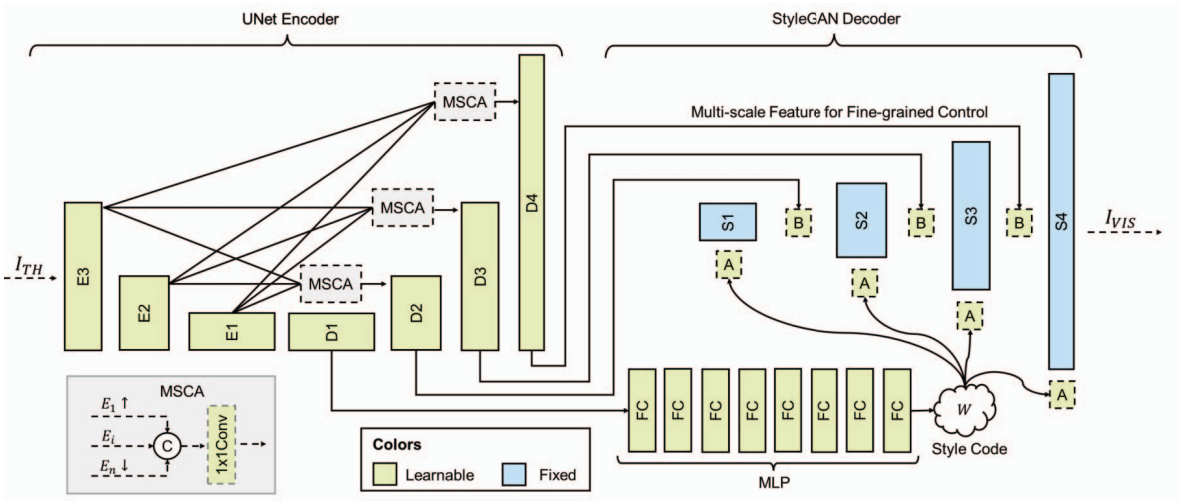

Figure 1. The proposed VPGAN. It adopts an encoder-decoder structure. The UNet encoder extracts style codes as well as multi-scale representations and then transmits them to the decoder for generation control. MSCA enhances the encoder by merging multi-scale information, which proves to be crucial for accurate hallucination. A pre-trained StyleGAN [25, 26] serves as the facial decoder and generates the desired visible face. A is the linear transformation of the style codes in [25, 26] and B plays a similar role as that of *noise injection*.

sults in more photo-realistic and accurate generation. Further improvements are based on cycle-consistency [60], more advanced loss [7], and attention mechanism [6, 23, 73].

GAN Inversion. Our method is related to GAN inversion [4, 17, 44, 55, 62, 71] which relies on pre-trained GAN priors for better image manipulation and restoration. Early approaches [17, 44] explicitly "invert GANs", which iteratively find the closest latent code of a targeting image. For example, PULSE [44] for photo upsampling gradually searches the correct latent code of a StyleGAN [25, 26] by optimizing a downsampling loss. More recent methods [4, 55, 62, 71] use a DNN encoder and learn to predict the latent code in one forward pass. Our work is inspired by these approaches but exploits its ability in transferring visual priors for data-efficient heterogeneous face hallucination.

Federated Learning. Federated learning is a decentralized machine learning framework which leverages data from multiple institutions or users to collaboratively train a global model without directly sharing their local data. Addressing heterogeneous data distribution across devices or institutions in the real-world deployment of FL applications draws emerging attention. Several FL methods [2, 18, 19, 34, 36, 46, 54, 69] targeting on this issue are built upon FedAvg [43]. FedProx [34] and Agnostic Federated Learning (AFL) [46] introduce an additional regularization on weights during the local training to alleviate the learning bias issue of the global model. FedDyn [2] is proposed to address the issue that there is an inconsistency between minima of the local model loss and those of the global loss by introducing a dynamic regularizer in each client. While those works conduct rigorous theoretical analysis, their performance is not validated on practical applications. Recently, Aggarwal *et al*. [3] proposed face recognition methods based on the FL framework. However, it is worth noting that the multi-institutional collaborative approach based on FL for face hallucination has not been well studied in the literature.

3. Proposed Method

In this section, we first formulate the TH-VIS face hallucination problem and then describe our method in detail. Given a TH image $\mathcal{I}_{\mathcal{TH}}$, our goal is to reconstruct a VIS face $\mathcal{I}_{\mathcal{VIS}}$ by learning a mapping function $\mathcal{I}_{\mathcal{VIS}} = \mathcal{F}(\mathcal{I}_{\mathcal{TH}})$. The synthesized face $\mathcal{I}_{\mathcal{VIS}}$ should be both visually realistic and accurate, and thus can be used for face matching. As discussed earlier, due to various reasons, $\mathcal{I}_{\mathcal{TH}}$ is often captured in low-resolution. Unlike existing methods (which only synthesize faces at 128×128), our work performs joint translation and upsampling. To the best of our knowledge, this is the first work that can hallucinate high-resolution faces (512×512) in HFR.

3.1. VPGAN

Conventional methods learn synthesis entirely from paired datasets. Due to data limitations, they can hardly output clear and high-quality images. Our method instead only learns to control generation by making use of diverse visual priors encapsulated in a pre-trained GAN. We leverage off-the-shelf StyleGAN [25, 26] pretrained on FFHQ [25], which contains 70,000 high-resolution faces. As shown in Figure 1, VPGAN adopts an encoder-decoder architecture. Only the encoder is required to train with the HFR dataset for guiding the hallucination. To output a visible face, we first extract global style codes

$$w = MLP(UNet_E(\mathcal{I}_{\mathcal{TH}})), \tag{1}$$

where $\mathcal{U}Net_E$ and MLP are the encoder part of the UNet and fully connected layers, respectively. Then we compute a

set of multi-scale features F_i from every stage of the decoder for fine-grained generation control, *i.e.* $F_i = \mathcal{U}Net_{D_i}(I_{TH})$. A visible face can then be produced via

$$\mathcal{I}_{\mathcal{VIS}} = \mathcal{S}(w, \{F_1, ..., F_n\}), \tag{2}$$

where $\mathcal{S}$ is the StyleGAN decoder. Thanks to diverse visual priors such as face geometry, color and texture, VPGAN is able to alleviate the need of large datasets and yields more faithful results.

Improved UNet Encoder. U-shaped structure has shown great capability in obtaining semantic-rich representations. However, we found that a naive UNet [57] is incapable of generating accurate local structures, which are crucial for face-matching. This is because features in the decoder are upsampled from coarser scales, and thus lack sufficient fine-grained information. Errors from early stages are also transmitted to the subsequent layers, leading to incorrect synthesis. To this end, we improve the conventional UNet [57] with a new Multi-Scale Contexts Aggregation (MSCA) mechanism, which provides a comprehensive encoding of the input image at multiple scales. This ensures the resulting features at all levels contain both coarse and fine information and thus leads to better generation control. As shown in Figure 1, MSCA computes an output by adaptively merging multi-scale features from the UNet encoder (after first upscaling and down-scaling to match the spatial dimensions). Formally, the output features at i-th level can be expressed as follows:

$$out_i = MSCA(E_1 \uparrow, ..., E_i, ..., E_n \downarrow), \tag{3}$$

where $\uparrow$ and $\downarrow$ denote the upscaling and downscaling operation, respectively. We will demonstrate that this simple design is crucial for accurate face matching in Section 4.2.1.

Embedded Visual Prior. Our key design is to utilize visual priors embedded in a pre-trained GAN. The StyleGAN [25, 26] decoder stores diverse facial knowledge and acts similar to a memory bank or dictionary, where the extracted style codes (from encoder) query desired faces. The style codes w can be incorporated either in a similar way to [26], by directly applying the modulation and demodulation operations on the convolution kernel of each style block, or through AdaIN [22] as used in [25]. Either operation is easy to implement and results in good generation quality. The multi-scale features from the UNet act in a similar way as the *noise injection* in the original StyleGAN. But rather than achieving localized variation, here we want to control detailed facial components to be consistent with the input image. We achieve this via a modulation operation (B in Figure 1) similar to [48, 64]. Specifically, we compute pixel-wise scale and shift parameters γ_i and β_i from the i-th multi-scale feature F_i via a simple 1×1 convolution layer. And then we use it to calibrate the output feature from the i-th StyleGAN layer:

$$S_i^+ = \gamma_i \odot S_i^- + \beta_i, \tag{4}$$

where $\odot$ denotes Hadamard product and S_i^- is the pre-modulated feature. After calibration, the resulting S_i^+ is passed to the next stage for subsequent generation.

Training Objectives. To ensure both realness and fidelity, our training objective consists of four terms: (1) reconstruction loss L_r, (2) adversarial loss L_{adv}, (3) perceptual loss L_p, and (4) Identity Loss L_{id}. The overall loss can be expressed as follows:

$$L_{gen} = L_r + \lambda_a L_{adv} + \lambda_b L_p + \lambda_c L_{id}, \tag{5}$$

where λ_a, λ_b and λ_c are corresponding balancing parameters. We define the reconstruction loss as the standard L_1 distance between the synthesized image $\mathcal{I}_{\mathcal{VIS}}$ and ground-truth image $\mathcal{I}_{\mathcal{GT}}$ to ensure content consistency. The adversarial loss is directly inherited from StyleGAN [25, 26] for more sharp generation. To improve visual quality while preserving identity, we further adopt the perceptual loss and the identity loss. Both can be expressed as feature-wise distance of a given CNN (*e.g.*, a pre-trained VGG):

$$L_p, L_{id} = \frac{1}{H_i W_i C_i} \|V_i(\mathcal{I}_{\mathcal{GT}}) - V_i(\mathcal{I}_{\mathcal{VIS}})\|_1, \tag{6}$$

where V is the corresponding CNN. H_i,W_i,C_i are the height, width and channel number of the i-th feature map in V. Here, we use an ImageNet pre-trained VGG for L_p and a simple ArcFace [5] for L_{id}.

3.2. Face Hallucination with Federated Learning

Even with an advanced design like VPGAN, training with limited data is still an essential shortcoming for deep models. To this end, we consider Federated Learning (FL) and introduce a novel Model Proximity Regularization (MPR). In this section, we will first describe a vanilla FL framework and then detail the proposed MPR.

A Vanilla FL Framework. We start by recalling issues in the standard local training. Suppose we have K HFR datasets $\mathcal{D}^1, \mathcal{D}^2, \ldots, \mathcal{D}^K$ dispersed at different institutions. These institutions can be different universities, government agencies or private companies. In a conventional non-collaboration scheme, a local model (parameterized by Θ_k) at institution k is trained with its own private data D^k, by optimizing $L_{gen,k}$ defined in Eq. 5:

$$\Theta_k^{p+1} \leftarrow \Theta_k^p - \gamma \nabla L_{gen,k}. \tag{7}$$

After several local gradient updates (i.e. P steps), institution k can obtain its local model. However, as discussed earlier, local data not only tend to have a limited capacity, but also may display unique characteristics due to discrepancies in the acquisition protocols. Therefore, the resulting model inevitably suffers from insufficient representation ability and low generalizability to other datasets. Ideally, one can mitigate such issue by training on a diversity-rich multi-source dataset or simply constructing a global dataset $\mathcal{D}$ from all available sources. Nevertheless, due to emerging privacy concerns, it is usually not the case for HFR.

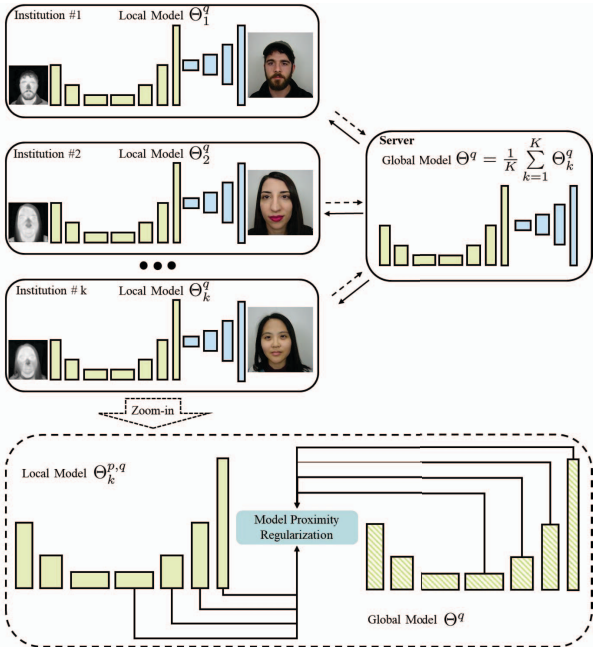

Figure 2. An overview of the proposed FL framework. Through q rounds of communication between data centers and the server, the collaboratively trained global model parameterized by Θ^q can be obtained in a data privacy-preserving manner. The zoomed-in subplot shows the proposed Model Proximity Regularization (MPR) in a local institution k.

Algorithm 1: VPFL with MPR

Input: $\mathcal{D}_s^1, \mathcal{D}_s^2, \ldots, \mathcal{D}_s^K$, K dispersed datasets; P, local update steps; Q, communication rounds; γ, learning rate; $\Theta_1, \ldots, \Theta_K$, local models; Θ, global model.

▷ parameters initialization

for *q = 0 to Q* **do**

  for *k = 0 to K in parallel* **do**

    ▷ deploy weights Θ^q to local model

    for *p = 0 to P* **do**

      VPGAN Face Hallucination:

      ▷ compute loss $L_{gen,k}$ using Eq. 5

      Model Proximity Regularization:

      ▷ compute the proximal term with respect to $\Theta_k^{p,q}$ and Θ^q

      ▷ compute the final local objective using Eq. 9 and update $\Theta_k^{p,q}$

    end

    ▷ upload weights to the central server

  end

  ▷ update the global model using Eq. 8

end

return Θ^Q

To maximize data utilization and learn a more generic model, we propose a vanilla FL framework based on FedAvg [43]. Rather than directly sharing private datasets, we leverage a centralized server to indirectly harvest information from all available institutions. This is achieved by periodically aggregating local models and broadcasting the updated results to all participants. A global update in the central server is calculated as follows:

$$\Theta^q = \frac{1}{K}\sum_{k=1}^{K}\Theta_k^q, \tag{8}$$

where q represents the q-th communication round. The final trained global model Θ^Q is obtained after Q rounds of client-server communications.

Model Proximity Regularization. While our vanilla FL algorithm manages to scale-up training, in real-world applications, the non-i.i.d. data distribution among institutions will still inevitably hurt the performance [34]. Due to the dissimilar local objectives $L_{gen,k}$ resulting from heterogeneous data distribution, local updates without proper constraints will cause the resulting model to skew toward the optima of its local objective, leading to inconsistency with the global one. Previous works [18, 50] circumvent this issue via FL adversarial training between source and target domains. However, such methods require directly sharing the latent features between participants, which compromises the privacy-preserving principle.

Inspired by [2, 34, 46], here we introduce a new Model Proximity Regularization (MPR) to correct local updates which can be easily combined with VPGAN. As shown in Figure 2, rather than solely minimizing the local objective $L_{gen,k}$, MPR introduces an extra proximal term for each local solver to force the proximity of latent codes from the current local model and the initial global model. The final local objective $\mathcal{L}_{gen,k}$ is adjusted to

$$\mathcal{L}_k = L_{gen,k} + \lambda_d\left(\|w - w^q\|^2 + \sum_{i=1}^{n}\|F_i - F_i^q\|^2\right), \tag{9}$$

where λ_d is a balancing parameter. In our unified framework VPFL, we integrate VPGAN as the base model and use it in the FL framework. VPFL thus can jointly enjoy benefits of strong visual priors and large-scale training for more realistic and faithful hallucination. The detailed training process can be found in Algorithm 1.

4. Experiments

In our experiments, we focus on synthesizing 512×512 VIS faces from 128×128 TH images, with an emphasis on both recognition accuracy and image quality. More analysis, discussion and additional results for resolutions less than 128×128 can be found in the supplementary material.

Dataset and Evaluation Metrics. So far, there are no standardized protocols in this field. Existing methods report results trained and tested on custom datasets/splits. This paper selects two common datasets (VIS-TH [42] and ARL-VTF [52]) where high-resolution VIS images are available. To study the effect of data restriction, we intentionally choose one dataset to be more challenging than the other. We create VIS-TH image pairs by cropping 512×512 and 128×128 face regions respectively.

VIS-TH is a challenging dataset containing data from 50 subjects. Images from each subject contain 21 faces varying significantly in pose, expression and light conditions. VIS-TH images are captured via a dual-sensor camera in LWIR

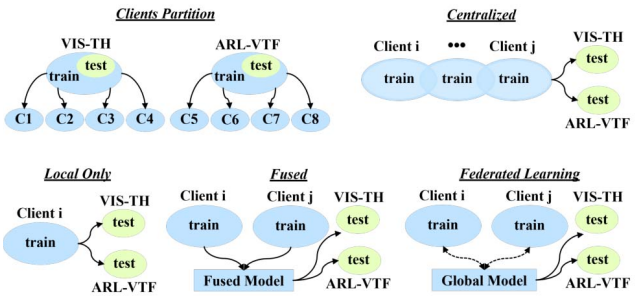

Figure 3. The schematics of the clients partition and different training strategies.

modality and thus naively aligned. We construct the training set by randomly selecting data from 40 subjects. The remaining data from 10 subjects are used as the testing set.

ARL-VTF provides subjects' data in LWIR modality with annotations for alignment. We create a dataset by randomly selecting a subset of 160 subjects with variations only in expressions as the training set, 20 subjects' data as the validation set, and 40 subjects' data as the testing set. The resulting data split contains 3,200 training pairs, 400 validation pairs, and 985 testing pairs. The color adjustment is applied to mitigate overexposure of the VIS images.

Evaluation Metrics. This paper extends the existing verification protocol with image quality measurements. For verification, we follow [8] and report Rank-1 accuracy, Verification Rate (VR) @ False Accept Rate (FAR)=1% and VR@FAR=0.1%. One VIS image of each subject is added to the gallery set and the probe set contains all TH images. To measure image quality, we report perceptual metrics LPIPS [77], NIQE [45], identity metric Deg (cosine distance between LightCNN [67] features), and pixel-wise PSNR and SSIM [65].

4.1. Implementation and Training Details

For VPGAN, we adopt off-the-shelf StyleGAN2 [26] as our facial decoder. The UNet encoder contains 5 downsample stages and 7 upsample stages for joint face translation and upsampling. Feature at the lowest level has a spatial size of 4×4. The network is trained using the Adam [27] optimizer with the following hyperparameters: initial learning rate of 2e-3 for the first 140K iterations then reduced to 1e-3; 150K maximum iterations; batch size of 4; $\lambda_a = 1$; $\lambda_b = 10$; $\lambda_c = 100$; λ_d =10e-4 if applicable. We implement the proposed model using PyTorch on Nvidia RTX8000 GPUs.

For VPFL, we adopt the same hyperparameters as that in VPGAN except 80K maximum iterations. The periodical communication between clients and the server is set to 200 iterations. Two training sets are further split into 4 subsets by sampling from a Dirichlet distribution (α=0.3) to simulate heterogeneous data distribution in the FL scheme, resulting in 8 independent clients as shown in Figure 3. Detailed dataset statistics in each client is provided in the supplementary material. For experiments in the FL setting, we not only compare different FL algorithms but also privacy-preserving alternative strategies. Models of **Local Only** are trained only with data from a single client and evaluated on two testing datasets. We denote the method that obtains independently trained models from all local clients and fuses their outputs as **Fused**, which does not violate privacy regulations. In addition, we can obtain a model trained by all available data, which is denoted by **Centralized**. Since it is prohibited in FL, we treat it as an upper bound. Figure 3 provides the schematics of different training and evaluation strategies in the FL setting.

Table 1. Image quality results on the **VIS-TH** dataset. **Red** and blue indicates the best and the second best performance.

Methods	LPIPS↓	NIQE↓	Deg.↑	PSNR↑	SSIM↑
TH	0.7147	10.666	36.13	6.41	0.3619
Pixel2Pixel [24]	0.3837	6.642	43.97	16.64	0.6818
HiFaceGAN [70]	0.3769	5.973	51.03	15.75	0.6794
GANVFS [76]	0.4012	6.314	43.95	16.69	0.6569
SAGAN [6]	0.2786	5.899	62.35	18.15	0.7179
AxialGAN [23]	0.2688	5.761	62.66	**19.02**	0.7190
VPGAN (ours)	**0.2253**	**5.508**	**68.36**	18.96	**0.7456**

4.2. Evaluations for VPGAN

Results on the VIS-TH dataset. To demonstrate the effectiveness of our VPGAN, we first report results on the challenging VIS-TH dataset and compare it with 5 representative methods: Pixel2Pixel [24], HiFaceGAN [70], GANVFS [76], SAGAN [6] and AxialGAN [23]. Pixel2Pixel is a well-known image-to-image translation method. HiFaceGAN is the state-of-the-art approach for face restoration. For TH-VIS hallucination, we select three leading methods GANVFS, SAGAN, and AxialGAN. Note: only AxialGAN has made their code publicly available. For GANVFS and SAGAN, implementations are acquired from authors.

Visual results are shown in Figure 4. As can be seen from this figure, previous approaches fail to generate clear visible faces. Specifically, Pixel2Pixel, HifaceGAN, GANVFS show strong artifacts and distortions in the generated faces. SAGAN and AxialGAN improve hallucination by adapting self-attention mechanism, but the produced images are still very blurry. In contrast, VPGAN outperforms previous methods by a large margin and it synthesizes the most faithful and accurate faces. Results for quantitative quality assessment are reported in Table 1. Our approach achieves the best performance in almost all metrics. VPGAN also obtains the highest Deg. value, indicating its superior ability in preserving identity. All of these results demonstrate the huge benefits of using visual priors for HFR.

We report verification results in Table 2. Given the superior generation quality, there is no surprise that VPGAN achieves the best performance on all metrics. Specif-

Figure 4. Visual comparison on the TH-VIS and ARL-VTF datasets. Low-resolution TH inputs are attached at the bottom right corner of the GT images with the real scale ratio (128:512) preserved. Our VPGAN can synthesize high-quality faces even with challenging expressions and large poses. Best viewed by zooming to **400%** in the screen.

Table 2. Verification results on the **VIS-TH** dataset.

Method	Rank-1	VR@FAR=1%	VR@FAR=0.1%
LightCNN [67]	30.48	8.57	2.86
Pixel2Pixel [24]	15.24	2.21	0.07
HiFaceGAN [70]	44.76	10.95	2.86
GANVFS [76]	18.11	7.29	1.90
SAGAN [6]	63.33	23.81	17.62
AxialGAN [23]	66.67	24.76	13.81
VPGAN (ours)	**76.67**	**45.71**	**20.00**

Table 3. Image quality results on the **ARL-VTF** dataset.

Methods	LPIPS↓	NIQE↓	Deg.↑	PSNR↑	SSIM↑
TH	0.6721	10.176	42.34	5.63	0.2940
Pixel2Pixel [24]	0.2038	6.298	70.67	19.46	0.7759
HiFaceGAN [70]	0.2166	7.274	70.11	19.67	**0.7954**
GANVFS [76]	0.2433	6.679	67.26	19.76	0.7511
SAGAN [6]	0.1925	6.155	71.12	20.11	0.7772
AxialGAN [23]	0.1998	6.223	69.75	20.17	0.7770
VPGAN (ours)	**0.1713**	**6.059**	**72.00**	**20.29**	0.7883

ically, our method significantly improves the baseline LightCNN [67] by 46%, and previous state-of-the-art AxialGAN by 10% in Rank-1 accuracy. In contrast, low-quality hallucinations, e.g., from Pixel2Pixel and GANVFS, can also impair the performance.

Results on the ARL-VTF dataset. To study the effect of data restriction, we further report results on the ARL-VTF dataset. This dataset contains $\times 4$ more subjects (160) than VIS-TH with only slight variations in expressions. Note that it is non-trivial for a single institution to achieve diversity at this scale. As shown in Figure 4, data simplicity allows previous methods to yield better visual results as expected. While most of them are able to produce a face outline, they struggle to create detailed facial components. In contrast, our approach can produce realistic and faithful facial details. Its superior hallucination ability is further verified by the quantitative results in Table 3. Face verification results are shown in Table 4. While previous methods can achieve reasonable performance, VPGAN still reaches the best performance given the more clear and accurate face details.

Table 4. Verification results on the **ARL-VTF** dataset.

Method	Rank-1	VR@FAR=1%	VR@FAR=0.1%
LightCNN [67]	11.07	9.24	4.57
Pixel2Pixel [24]	70.96	56.35	33.60
HiFaceGAN [70]	70.15	56.65	32.18
GANVFS [76]	70.76	45.99	22.03
SAGAN [6]	71.16	54.11	38.07
AxialGAN [23]	71.57	57.16	37.36
VPGAN (ours)	**74.16**	**59.96**	**41.27**

4.2.1 Ablation Study

Effects of Visual Priors. The key design of VPGAN is to leverage rich and diverse visual priors for better hallucina-

Table 5. Ablation study on the VIS-TH dataset.

	Rank-1	VR@FAR=1%	Deg.	LPIPS↓	PSNR↑
w/o VP	52.85	19.04	57.69	0.2948	18.23
w/o MSCA	71.90	31.43	63.15	0.2460	18.74
VPGAN	**76.19**	**38.10**	**65.87**	**0.2381**	**18.85**

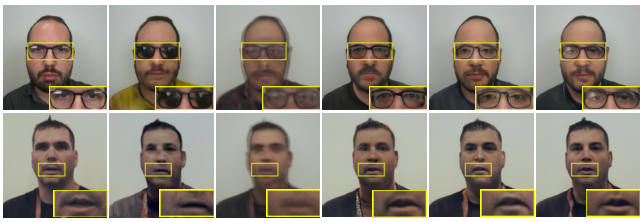

Figure 5. Visual comparison under the FL setting. VPFL is able to recover the most accurate face components.

tion. Here, we investigate its effectiveness. We construct a baseline model by removing the pre-trained decoder, resulting in a U-Shaped generator. As shown in Table 5, the performance is substantially improved by incorporating visual priors, demonstrating our design is indeed beneficial and helpful.

Multi-Scale Contexts Aggregation. VPGAN improves the standard UNet encoder with Multi-Scale Contexts Aggregation module. Table 5 shows its effect. Adding the MSCA can obviously improve results in all metrics, especially in terms of recognition accuracy and Deg. These results indicate that MSCA can provide more accurate generation control and preserve the identity information better.

4.3. Evaluations for VPFL

Here we show the benefits of collaborative training for HFR by revealing performance and generalizability gaps. Table 6 presents the verification and quality assessment results of different privacy-preserving strategies on three sub-tables VIS-TH, ARL-VTF, and *Global Test Avg*. *Global Test Avg.* refers to the average performance on the two datasets and reflects generalizability.

The first 8 rows of each sub-table report the results of locally trained models (**Local Only**). We treat them as baselines. Due to the data heterogeneity, all locally trained models exhibit low generalizability on the data from another distribution. For example, C1-C4 achieve low performance on ARL-VTF and vice versa. This can also be verified by their poor results on the Global Test Avg. In addition, the naive **Fused** strategy cannot consistently improve the performance. In contrast, apparent improvements can be observed by introducing the FL scheme, which demonstrates the necessity of collaborative training. When compared with the strong FL baseline FedProx [34] (with the empirically best μ=10e-4), our VPFL achieves the significantly better overall performance and is closest to the **Centralized** upper bound. These results indicate that VPFL can generalize well and is more robust towards heterogeneous data distribution. As

Table 6. Verification and image quality comparisons with different methods under the FL setting. VR1% denotes VR@FAR=1%.

VIS-TH								
Methods	Rank-1↑	VR1%↑	VR0.1%↑	Deg.↑	PSNR↑	SSIM↑	LPIPS↓	NIQE↓
Local Only C1	59.05	30.95	17.62	60.69	18.05	0.723	0.262	5.903
Local Only C2	32.86	1.90	0.48	51.52	16.56	0.699	0.315	5.809
Local Only C3	50.95	20.00	7.14	55.49	17.65	0.718	0.283	6.403
Local Only C4	46.19	17.62	10.00	56.82	17.01	0.700	0.294	5.748
Local Only C5	47.62	17.14	8.10	51.64	14.94	0.664	0.383	5.911
Local Only C6	42.86	11.43	6.19	53.17	14.75	0.664	0.385	6.171
Local Only C7	40.95	15.24	5.24	51.30	15.01	0.668	0.385	6.420
Local Only C8	40.95	15.71	5.71	52.20	14.80	0.665	0.384	6.426
Fused	37.62	16.19	7.14	55.83	17.36	**0.732**	0.328	6.934
FedProx [34]	66.19	30.95	20.00	61.99	17.86	0.718	0.262	**5.565**
VPFL w/o MPR	70.95	30.95	22.38	61.16	18.19	0.719	0.254	5.579
VPFL	**73.81**	**35.71**	**25.71**	**65.81**	**18.81**	0.728	**0.245**	5.651
Centralized	76.67	39.05	24.76	66.63	18.71	0.743	0.232	5.729
ARL-VTF								
Methods	Rank-1↑	VR1%↑	VR0.1%↑	Deg.↑	PSNR↑	SSIM↑	LPIPS↓	NIQE↓
Local Only C1	17.77	10.66	2.44	52.44	16.78	0.699	0.325	6.708
Local Only C2	20.51	11.78	2.74	47.65	16.67	0.704	0.330	6.821
Local Only C3	16.85	14.31	2.84	49.41	16.67	0.713	0.332	6.777
Local Only C4	23.55	10.66	3.96	53.30	16.51	0.690	0.335	6.771
Local Only C5	54.11	39.80	20.10	64.96	19.36	0.764	0.211	6.770
Local Only C6	54.31	36.65	**22.34**	65.66	19.28	0.768	0.213	6.294
Local Only C7	40.91	33.91	18.07	63.45	19.51	0.776	0.211	6.335
Local Only C8	54.82	37.77	20.91	64.05	19.01	0.762	0.222	6.205
Fused	37.16	26.50	10.56	62.45	**19.96**	**0.789**	0.260	6.840
FedProx [34]	57.77	37.36	15.94	67.21	19.60	0.770	0.212	6.022
VPFL w/o MPR	62.03	36.14	16.45	67.46	19.51	0.770	0.209	6.019
VPFL	**65.79**	**40.71**	22.23	**67.68**	19.69	0.773	**0.203**	**6.013**
Centralized	69.34	57.77	28.63	71.28	20.08	0.785	0.186	6.106
Global Test Avg.								
Methods	Rank-1↑	VR1%↑	VR0.1%↑	Deg.↑	PSNR↑	SSIM↑	LPIPS↓	NIQE↓
Local Only C1	38.41	20.81	10.03	56.57	17.41	0.711	0.293	6.306
Local Only C2	26.69	6.84	1.61	49.59	16.61	0.701	0.322	6.315
Local Only C3	33.90	17.16	4.99	52.45	17.16	0.716	0.308	6.590
Local Only C4	34.87	14.14	6.98	55.06	16.76	0.695	0.314	6.259
Local Only C5	50.87	28.47	14.10	58.30	17.15	0.714	0.297	6.341
Local Only C6	48.59	24.04	14.27	59.42	17.01	0.716	0.299	6.233
Local Only C7	40.93	24.58	11.66	57.38	17.26	0.722	0.298	6.378
Local Only C8	47.89	26.74	13.31	58.13	16.90	0.713	0.303	6.316
Fused	37.39	21.35	8.85	59.14	18.66	**0.761**	0.294	6.887
FedProx [34]	61.98	34.16	17.97	64.60	18.73	0.744	0.237	**5.794**
VPFL w/o MPR	66.49	33.55	19.42	64.31	18.85	0.745	0.231	5.799
VPFL	**69.80**	**38.21**	**23.97**	**66.75**	**19.25**	0.751	**0.224**	5.832
Centralized	73.01	48.41	26.70	68.96	19.40	0.764	0.209	5.918

shown in the last three rows of each sub-table, the advantages come from the newly designed MPR. These quantitative results are also aligned with the visual comparison in Figure 5. One can see that VPFL yields the most accurate and faithful hallucination results.

5. Conclusion

In this paper, we proposed a unified framework VPFL for heterogeneous face hallucination. VPFL consists of a novel VPGAN and a new Federated Learning (FL) scheme. VPGAN introduces powerful visual priors to avoid learning hallucination from scratch, resulting in more accurate generation under the current data limitations. With the consideration of practical privacy issues, the proposed FL scheme allows institution-wise collaborations without sharing data, making large-scale training possible. Extensive experiments demonstrate that VPFL can significantly boost HFR by synthesizing accurate and realistic visible faces at a resolution unseen in the literature. Discussion of limitations can be found in the supplementary material.

Acknowledgments This work was supported by NSF CARRER award 2045489.

References

[1] The buaa-visnir face database instructions. 2012. 1

[2] Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2020. 3, 5

[3] Divyansh Aggarwal, Jiayu Zhou, and Anil K Jain. Fedface: Collaborative learning of face recognition model. *arXiv preprint arXiv:2104.03008*, 2021. 3

[4] Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 14245–14254, 2021. 3

[5] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 1, 4

[6] Xing Di, Benjamin S Riggan, Shuowen Hu, Nathaniel J Short, and Vishal M Patel. Polarimetric thermal to visible face verification via self-attention guided synthesis. In *IEEE International Conference on Biometrics*, pages 1–8, 2019. 3, 6, 7

[7] Xing Di, He Zhang, and Vishal M Patel. Polarimetric thermal to visible face verification via attribute preserved synthesis. In *IEEE International Conference on Biometrics Theory, Applications and Systems*, pages 1–10, 2018. 1, 3

[8] Boyan Duan, Chaoyou Fu, Yi Li, Xingguang Song, and Ran He. Cross-spectral face hallucination via disentangling independent factors. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7930–7938, 2020. 1, 2, 6

[9] Virginia Espinosa-Duró, Marcos Faundez-Zanuy, and Jiří Mekyska. A new face database simultaneously acquired in visible, near-infrared and thermal spectrums. *Cognitive Computation*, 5(1):119–135, 2013. 1

[10] Chaoyou Fu, Yibo Hu, Xiang Wu, Hailin Shi, Tao Mei, and Ran He. Cm-nas: Cross-modality neural architecture search for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11823–11832, 2021. 1

[11] Chaoyou Fu, Yibo Hu, Xiang Wu, Guoli Wang, Qian Zhang, and Ran He. High-fidelity face manipulation with extreme poses and expressions. *IEEE Transactions on Information Forensics and Security*, 16:2218–2231, 2021. 2

[12] Chaoyou Fu, Xiang Wu, Yibo Hu, Huaibo Huang, and Ran He. Dual variational generation for low shot heterogeneous face recognition. *Advances in Neural Information Processing Systems*, 32:2674–2683, 2019. 1, 2

[13] Chaoyou Fu, Xiang Wu, Yibo Hu, Huaibo Huang, and Ran He. Dvg-face: Dual variational generation for heterogeneous face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2021. 2

[14] Hamed Kiani Galoogahi and Terence Sim. Face sketch recognition by local radon binary pattern: Lrbp. In *IEEE International Conference on Image Processing*, pages 1837–1840, 2012. 1

[15] Dihong Gong, Zhifeng Li, Weilin Huang, Xuelong Li, and Dacheng Tao. Heterogeneous face recognition: A common encoding feature discriminant approach. *IEEE Transactions on Image Processing*, 26(5):2079–2089, 2017. 1

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1

[17] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3012–3021, 2020. 3

[18] Pengfei Guo, Puyang Wang, Jinyuan Zhou, Shanshan Jiang, and Vishal M Patel. Multi-institutional collaborations for improving deep learning-based magnetic resonance image reconstruction using federated learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2423–2432, 2021. 3, 5

[19] Pengfei Guo, Dong Yang, Ali Hatamizadeh, An Xu, Ziyue Xu, Wenqi Li, Can Zhao, Daguang Xu, Stephanie Harmon, Evrim Turkbey, et al. Auto-fedrl: Federated hyperparameter optimization for multi-institutional medical image segmentation. *arXiv preprint arXiv:2203.06338*, 2022. 3

[20] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, pages 87–102. Springer, 2016. 1

[21] Ran He, Xiang Wu, Zhenan Sun, and Tieniu Tan. Wasserstein cnn: Learning invariant features for nir-vis face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1761–1773, 2018. 1

[22] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 4

[23] Rakhil Immidisetti, Shuowen Hu, and Vishal M. Patel. Simultaneous face hallucination and translation for thermal to visible face verification using axial-gan. In *IEEE International Joint Conference on Biometrics*, pages 1–8, 2021. 2, 3, 6, 7

[24] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017. 6, 7

[25] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 2, 3, 4

[26] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 2, 3, 4, 6

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 6

[28] Brendan F Klare and Anil K Jain. Heterogeneous face recognition using kernel prototype similarities. *IEEE Transactions*

on Pattern Analysis and Machine Intelligence, 35(6):1410–1422, 2012. 1

[29] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016. 2

[30] José Lezama, Qiang Qiu, and Guillermo Sapiro. Not afraid of the dark: Nir-vis face recognition via cross-spectral hallucination and low-rank embedding. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6628–6637, 2017. 2

[31] Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study. *arXiv preprint arXiv:2102.02079*, 2021. 2

[32] Stan Z Li, RuFeng Chu, ShengCai Liao, and Lun Zhang. Illumination invariant face recognition using near-infrared images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):627–639, 2007. 1

[33] Stan Z Li, Zhen Lei, and Meng Ao. The hfb face database for heterogeneous face biometrics research. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2009. 1

[34] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, pages 429–450, 2020. 3, 5, 8

[35] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2019. 2

[36] Xiaoxiao Li, Meirui JIANG, Xiaofei Zhang, Michael Kamp, and Qi Dou. FedBN: Federated learning on non-IID features via local batch normalization. In *International Conference on Learning Representations*, 2021. 3

[37] Shengcai Liao, Dong Yi, Zhen Lei, Rui Qin, and Stan Z Li. Heterogeneous face recognition from local structures of normalized appearance. In *IEEE International Conference on Biometrics*, pages 209–218. Springer, 2009. 1

[38] Jinpeng Lin, Hao Yang, Dong Chen, Ming Zeng, Fang Wen, and Lu Yuan. Face parsing with roi tanh-warping. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5654–5663, 2019. 2

[39] Sifei Liu, Jianping Shi, Ji Liang, and Ming-Hsuan Yang. Face parsing via recurrent propagation. *arXiv preprint arXiv:1708.01936*, 2017. 2

[40] Sifei Liu, Dong Yi, Zhen Lei, and Stan Z Li. Heterogeneous face image matching using multi-scale features. In *IEEE International Conference on Biometrics*, pages 79–84, 2012. 1

[41] Khawla Mallat, Naser Damer, Fadi Boutros, Arjan Kuijper, and Jean-Luc Dugelay. Cross-spectrum thermal to visible face recognition based on cascaded image synthesis. In *IEEE International Conference on Biometrics*, pages 1–8, 2019. 2

[42] Khawla Mallat and Jean-Luc Dugelay. A benchmark database of visible and thermal paired face images across multiple variations. In *International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5, 2018. 1, 5

[43] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017. 3, 5

[44] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2437–2445, 2020. 3

[45] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 6

[46] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, pages 4615–4625. PMLR, 2019. 3, 5

[47] Karen Panetta, Qianwen Wan, Sos Agaian, Srijith Rajeev, Shreyas Kamath, Rahul Rajendran, Shishir Paramathma Rao, Aleksandra Kaszowska, Holly A Taylor, Arash Samani, et al. A comprehensive database for benchmarking imaging systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(3):509–520, 2018. 1

[48] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019. 4

[49] Chunlei Peng, Nannan Wang, Jie Li, and Xinbo Gao. Soft semantic representation for cross-domain face recognition. *IEEE Transactions on Information Forensics and Security*, 16:346–360, 2020. 1

[50] Xingchao Peng, Zijun Huang, Yizhe Zhu, and Kate Saenko. Federated adversarial domain adaptation. *arXiv preprint arXiv:1911.02054*, 2019. 5

[51] N. Peri, J. Gleason, C. D. Castillo, T. Bourlai, V. M. Patel, and R. Chellappa. A synthesis-based approach for thermal-to-visible face verification. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2021. 2

[52] Domenick Poster, Matthew Thielke, Robert Nguyen, Srinivasan Rajaraman, Xing Di, Cedric Nimpa Fondje, Vishal M Patel, Nathaniel J Short, Benjamin S Riggan, Nasser M Nasrabadi, et al. A large-scale, time-synchronized visible and thermal face dataset. In *IEEE Winter Conference on Applications of Computer Vision*, pages 1559–1568, 2021. 1, 5

[53] Albert Pumarola, Antonio Agudo, Aleix M. Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *European Conference on Computer Vision*. Springer, 2018. 2

[54] Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H Brendan McMahan. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*, 2020. 3

[55] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2287–2296, 2021. 3

[56] Benjamin S Riggan, Nathaniel J Short, and Shuowen Hu. Thermal to visible synthesis of face images using multiple regions. In *IEEE Winter Conference on Applications of Computer Vision*, pages 30–38, 2018. 2

[57] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 4

[58] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015. 1

[59] Ming Shao, Dmitry Kit, and Yun Fu. Generalized transfer subspace learning through low-rank constraint. *International Journal of Computer Vision*, 109(1-2):74–93, 2014. 1

[60] Lingxiao Song, Man Zhang, Xiang Wu, and Ran He. Adversarial discriminative heterogeneous face recognition. In *the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 1, 3

[61] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018. 1

[62] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9168–9178, 2021. 3

[63] Xiaogang Wang and Xiaoou Tang. Face photo-sketch synthesis and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):1955–1967, 2008. 1

[64] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 606–615, 2018. 4

[65] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 6

[66] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. Springer, 2016. 1

[67] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896, 2018. 6, 7

[68] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *arXiv preprint arXiv:2101.05278*, 2021. 2

[69] An Xu, Wenqi Li, Pengfei Guo, Dong Yang, Holger Roth, Ali Hatamizadeh, Can Zhao, Daguang Xu, Heng Huang, and Ziyue Xu. Closing the generalization gap of cross-silo federated medical image segmentation. *arXiv preprint arXiv:2203.10144*, 2022. 3

[70] Lingbo Yang, Shanshe Wang, Siwei Ma, Wen Gao, Chang Liu, Pan Wang, and Peiran Ren. Hifacegan: Face renovation via collaborative suppression and replenishment. In *ACM International Conference on Multimedia*, pages 1551–1560, 2020. 6, 7

[71] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 672–681, 2021. 3

[72] Dong Yi, Rong Liu, RuFeng Chu, Zhen Lei, and Stan Z Li. Face matching between near infrared and visible light images. In *IEEE International Conference on Biometrics*, pages 523–530. Springer, 2007. 1

[73] Junchi Yu, Jie Cao, Yi Li, Xiaofei Jia, and Ran He. Pose-preserving cross spectral face hallucination. In *International Joint Conference on Artificial Intelligence*, pages 1018–1024, 7 2019. 1, 3

[74] Yu Zeng, Zhe Lin, and Vishal M Patel. Sketchedit: Mask-free local image manipulation with partial sketches. *arXiv preprint arXiv:2111.15078*, 2021. 2

[75] He Zhang, Vishal M Patel, Benjamin S Riggan, and Shuowen Hu. Generative adversarial network-based synthesis of visible faces from polarimetrie thermal faces. In *IEEE International Joint Conference on Biometrics*, pages 100–107, 2017. 1, 2

[76] He Zhang, Benjamin S Riggan, Shuowen Hu, Nathaniel J Short, and Vishal M Patel. Synthesis of high-quality visible faces from polarimetric thermal faces using generative adversarial networks. *International Journal of Computer Vision*, 127(6):845–862, 2019. 2, 6, 7

[77] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 6