

**A Metric for Keypoint Transfer**

In the keypoint transfer experiment, we calculate the pixel error between the transferred keypoint \( \hat{u}_t \) and the ground truth target keypoint \( u_t \) to evaluate facial embedding quantitatively. First, source embedding \( E^s \) and target embedding \( E^t \) are extracted from the source image and target image. Then, the ground truth source keypoint \( u_s \) is transferred to the target via cosine distance matching of pixel embedding formulated as

\[
\sigma(E^s, E^t, u_i, u_j) = \frac{E^s_{u_i} \cdot E^t_{u_j}}{\|E^s_{u_i}\| \cdot \|E^t_{u_j}\|},
\]

(1)

\[
\hat{u}_t = \arg \min_{u \in \Omega} \sigma(E^s, E^t, u, u),
\]

(2)

with 2D cartesian coordinate system \( \Omega = \{0, 1, ..., H' - 1\} \times \{0, 1, ..., W' - 1\} \).

After the keypoint transfer, pixel error \( e_p \) is calculated by the euclidean distance between the transferred keypoint and the label of target formulated as following:

\[
e_p = \|u_t - \hat{u}_t\|.
\]

(3)

With lower pixel error, we can assure the learned embedding works well as visual descriptors as matching reliably different face geometry of various species [10].

**B The reproducibility of our work**

For reproducibility of our work, we provide the code blocks in the zipped supplementary file. The codes related to dataset, model, loss, training pipeline and experiment are enclosed. To make sure the training stability of our pipeline, we execute multiple training and interspecies keypoint transfer experiment with respect to the random seed. All values for Section 4 are computed in three randomly seeded training and the statistics for some experiments are shown in Table 1.
Table 2: The human landmark detection results.

<table>
<thead>
<tr>
<th>Category</th>
<th>Method</th>
<th>Unsup.</th>
<th>Cross-Domain</th>
<th>MAFL</th>
<th>AFLW_M</th>
<th>AFLW_R</th>
<th>300W</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised learning</td>
<td>TCDCN [13]</td>
<td>X</td>
<td>X</td>
<td>7.95</td>
<td>7.65</td>
<td>-</td>
<td>5.54</td>
</tr>
<tr>
<td></td>
<td>MTCNN [12]</td>
<td>X</td>
<td>X</td>
<td>5.39</td>
<td>6.90</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Wing Loss [3]</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td></td>
<td>4.04</td>
</tr>
<tr>
<td>Generative modeling based</td>
<td>Deforming AE [9]</td>
<td>O</td>
<td>X</td>
<td>5.45</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>ImGen. [4]</td>
<td>O</td>
<td>X</td>
<td>5.34</td>
<td>6.31</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>ImGEN.++ [5]</td>
<td>O</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td></td>
<td>5.12</td>
</tr>
<tr>
<td></td>
<td>DVE [10]</td>
<td>O</td>
<td>X</td>
<td>2.86</td>
<td>7.53</td>
<td>6.54</td>
<td>4.65</td>
</tr>
<tr>
<td></td>
<td>DIFE</td>
<td>O</td>
<td>O</td>
<td>3.40</td>
<td>10.11</td>
<td>8.68</td>
<td>7.57</td>
</tr>
</tbody>
</table>

Figure 2: The qualitative results of the human keypoint regression.

C The human landmark detection

We carry out the keypoint regression experiment setup following DVE [10] and On Equivariant And Invariant [1] to evaluate the robustness and the accuracy of DIFE on intra-species landmark detection. In the experiment, we train a single FC layer with a frozen pre-trained embedder to predict keypoint. The NME values of previous methods are also brought from On Equivariant And Invariant. Even though our embedder is trained on synthesized datasets, not the target dataset, DIFE shows compatible performance with the early study results of each category meaning DIFE is the apposite baseline for cross-domain face understanding. The qualitative results are also provided in Fig. 2.

D Detailed Explanation for Method

Algorithm[1] shows our detailed training pipeline as pseudo-code. From line 6 to line 18, the pseudo-code for multi-teacher knowledge distillation is provided. At the start of training, latent code $z$ is sampled with normal distribution. The synthesized image $\mathbf{x}^{d_1}$, $\mathbf{x}^{d_2}$ and the feature of StyleGAN2 $\mathbf{F}^{d_1}$, $\mathbf{F}^{d_2}$ is generated with the StyleGAN2 $\mathbf{G}^{d_1}$, $\mathbf{G}^{d_2}$. Here $d_1$ and $d_2$ means the the domain of instance. The continuous surface embedding $\mathbf{C}^{d_1}$, $\mathbf{C}^{d_2}$ is extracted with the CSE $\mathbf{E}$. Therfore, DIFE $\mathbf{E}^{d_1}$, $\mathbf{E}^{d_2}$ and domain-specific embedding $\mathbf{D}^{d_1 \rightarrow d_1}$, $\mathbf{D}^{d_2 \rightarrow d_2}$ are extracted with the encoder $\Psi$ and the domain converter $\Psi^{d_1}$, $\Psi^{d_2}$. We utilize superscript $d \rightarrow d'$ to express the instance of domain $d'$ is originated from the image of $d$ which is useful to describe pseudo-paired data synthesis later. Finally, CSE knowledge distillation loss $\mathbf{L}_{K1}$ and StyleGAN2 knowledge distillation loss $\mathbf{L}_{K2}$ are calculated with
Algorithm 1 The Pipeline of Training For Two Domains

1: (0) Prepare DIFE training
2: Set the learning rate \( \eta \)
3: Load pre-trained weights For parameters of CSE \( \mathcal{E} \), StyleGAN2 \( G^{d_1}, G^{d_2} \)
4: Initialize network parameters \( \theta_\Phi, \theta_{\psi d_1}, \theta_{\psi d_2} \)
5: for the number of iterations do
6: (1-1) Synthesize data and labels with teacher models
7: Sample latent code \( z \sim \mathcal{N}(0, 1) \)
8: \( x^{d_1}, F^{d_1} \leftarrow G^{d_1}(z) \)
9: \( x^{d_2}, F^{d_2} \leftarrow G^{d_2}(z) \)
10: \( C^{d_1}, C^{d_2} \leftarrow \mathcal{E}(x^{d_1}, x^{d_2}) \)
11: (1-2) Extract embedding
12: \( E^{d_1}, E^{d_2} \leftarrow \Phi(x^{d_1}, x^{d_2}) \)
13: \( D^{d_1\rightarrow d_1} \leftarrow \psi_d^{d_1}(E^{d_1}) \)
14: \( D^{d_2\rightarrow d_2} \leftarrow \psi_d^{d_2}(E^{d_2}) \)
15: (1-3) Calculate losses For multi-teacher knowledge distillation
16: \( \mathcal{L}_{K_1} = \mathcal{L}_K(E^{d_1}, C^{d_1}) + \mathcal{L}_K(E^{d_2}, C^{d_2}) \)
17: \( \mathcal{L}_{K_2}^{d_1} = \mathcal{L}_K(D^{d_1\rightarrow d_1}, F^{d_1}) \)
18: \( \mathcal{L}_{K_2}^{d_2} = \mathcal{L}_K(D^{d_2\rightarrow d_2}, F^{d_2}) \)
19: if not initialization period then
20: (2-1) Intra pseudo-paired data synthesis
21: \( x^{d_1\rightarrow d_1} \leftarrow G^{d_1}(D^{d_1\rightarrow d_1}) \) with latent space exploration
22: \( x^{d_2\rightarrow d_2} \leftarrow G^{d_2}(D^{d_2\rightarrow d_2}) \) with latent space exploration
23: (2-2) Inter pseudo-paired data synthesis
24: \( D^{d_2\rightarrow d_1} \leftarrow \psi_d^{d_2}(E^{d_2}) \)
25: \( D^{d_1\rightarrow d_2} \leftarrow \psi_d^{d_1}(E^{d_1}) \)
26: \( x^{d_2\rightarrow d_1} \leftarrow G^{d_1}(D^{d_2\rightarrow d_1}) \) with latent space exploration
27: \( x^{d_1\rightarrow d_2} \leftarrow G^{d_2}(D^{d_1\rightarrow d_2}) \) with latent space exploration
28: (2-3) Extract embedding
29: \( E^{d_1\rightarrow d_1}, E^{d_1\rightarrow d_2}, E^{d_2\rightarrow d_1}, E^{d_2\rightarrow d_2} \leftarrow \Phi(x^{d_1\rightarrow d_1}, x^{d_2\rightarrow d_1}, x^{d_1\rightarrow d_2}, x^{d_2\rightarrow d_2}) \)
30: (2-3) Calculate losses For semantic matching
31: \( \mathcal{L}_{M_1} = \mathcal{L}_M(E^{d_1}, E^{d_1\rightarrow d_1}) + \mathcal{L}_M(E^{d_2}, E^{d_2\rightarrow d_2}) \)
32: \( \mathcal{L}_{M_2} = \mathcal{L}_M(E^{d_1}, E^{d_1\rightarrow d_2}) + \mathcal{L}_M(E^{d_2}, E^{d_2\rightarrow d_1}) \)
33: else
34: \( \mathcal{L}_{M_1} = 0 \)
35: \( \mathcal{L}_{M_2} = 0 \)
36: end if
37: (3) Update network parameters
38: \( \theta_\Phi = \theta_\Phi + \eta \nabla \theta_\Phi (\mathcal{L}_{K_1} + \mathcal{L}_{K_2} + \mathcal{L}_{M_1} + \mathcal{L}_{M_2}) \)
39: \( \theta_{\psi d_1} = \theta_{\psi d_1} + \eta \nabla \theta_{\psi d_1} (\mathcal{L}_{K_2}^{d_1}) \)
40: \( \theta_{\psi d_2} = \theta_{\psi d_2} + \eta \nabla \theta_{\psi d_2} (\mathcal{L}_{K_2}^{d_2}) \)
41: end for

knowledge distillation loss \( \mathcal{L}_K \) formulated as following:

\[
\mathcal{L}_K(A, B) = \|A - B\|_2.
\] (4)

The following codes from line 20 to line 32 presents the procedure of the pseudo-paired data synthesis and the semantic matching. The if statement on line 18 means pseudo-paired data is not synthesized in initialization period because the data synthesis requires proper face geometry of the domain-specific embedding. After the initialization period, the intra pseudo-paired data \( x^{d_1\rightarrow d_1}, x^{d_2\rightarrow d_2} \) and inter pseudo-paired data \( x^{d_1\rightarrow d_2}, x^{d_2\rightarrow d_1} \) are generated with latent space exploration described at Section 3.2. With encoder \( \Phi \), each DIFE for pseudo-paired data is extracted. At last semantic matching loss \( \mathcal{L}_M \) are utilized to get intra semantic matching loss \( \mathcal{L}_{M_1} \) and inter semantic matching loss \( \mathcal{L}_{M_2} \).
Here the equation for semantic matching loss with cosine similarity $\sigma$ defined at Appendix B:

$$L_M(A, B) = \sum_{u_i \in \Omega} \left\| u_i - \sum_{u_j \in \Omega} \sigma(A, B, u_i, u_j) u_j \right\|_2.$$  \hspace{1cm} (5)

At the end of iteration, the parameters of the encoder $\theta_\Phi$ is updated with all losses and the parameters of the domain converter $\theta_{\Psi_{d1}}, \theta_{\Psi_{d2}}$ are updated with StyleGAN2 knowledge distillation loss.

**E More Examples of Pseudo-Paired Data**

We provide more examples of pseudo-paired data on various combinations of original and pair domains in Fig. 3. Each three group of rows shows examples of pseudo-paired data for human+dog, human+cat, and dog+cat. With various pseudo-paired data, we observe the generated image has the same face geometry with the original image. In inter pseudo-paired data, there is a little difference like the location of the ear or the shape of the eye, but the posture of the face is always aligned. By matching them, ambiguous connections between the interspecies faces can be learned in unsupervised manner.
Figure 4: The qualitative results of the interspecies keypoint transfer on WFLW and AnimalWeb. The blue landmarks are predicted and the green landmarks are the ground truth.

F Intellectual property rights of assets

The detailed licenses of utilized codes and datasets are as followings:

- Detectron2, DensePose-CSE [8]: The license of code is Apache 2.0 License which is an open source license.
- StyleGAN, FFHQ [6]: The individual images were published in Flickr by their respective authors under either Creative Commons BY 2.0, Creative Commons BY-NC 2.0, Public Domain Mark 1.0, Public Domain CC0 1.0, or U.S. Government Works license. All of these licenses allow free use, redistribution, and adaptation for non-commercial purposes.
- AFHQ [2]: The dataset is available under Creative Commons BY-NC 4.0 license by NAVER Corporation. There is no problem to use, copy, transform and build upon the material for non-commercial purposes as long as giving appropriate credit by citing our paper, and indicating if changes were made.
- AP-10K [11]: The dataset is available under Creative Commons BY-NC 4.0 license which is same with AFHQ.
- CelebA [7], MAFL [14]: The name of license is unknown but datasets are available for non-commercial research purposes.
References


