Dense Interspecies Face Embedding

Sejong Yang
Yonsei University
sejong.yang@yonsei.ac.kr

Subin Jeon
Yonsei University
subin.jeon@yonsei.ac.kr

Seonghyeon Nam
York University
snam0331@gmail.com

Seon Joo Kim
Yonsei University
seonjookim@yonsei.ac.kr

Abstract

Dense Interspecies Face Embedding (DIFE) is a new direction for understanding faces of various animals by extracting common features among animal faces including human face. There are three main obstacles for interspecies face understanding: (1) lack of animal data compared to human, (2) ambiguous connection between faces of various animals, and (3) extreme shape and style variance. To cope with the lack of data, we utilize multi-teacher knowledge distillation of CSE and StyleGAN2 requiring no additional data or label. Then we synthesize pseudo pair images through the latent space exploration of StyleGAN2 to find implicit associations between different animal faces. Finally, we introduce the semantic matching loss to overcome the problem of extreme shape differences between species. To quantitatively evaluate our method over possible previous methodologies like unsupervised keypoint detection, we perform interspecies facial keypoint transfer on MAFL and AP-10K. Furthermore, the results of other applications like interspecies face image manipulation and dense keypoint transfer are provided. The code is available at https://github.com/kingsj0405/dife.

1 Introduction

Face understanding is a longstanding research topic in computer vision, and there have been extensive studies and applications such as face recognition, facial expression estimation, face photo manipulation, and virtual humans [6, 52, 4, 57, 3]. The vast majority of successful works on faces have focused on human faces due to its practical implication for a variety of applications and the accessibility to a large amount of annotated data. There have been few attempts for understanding faces of other animal species to facilitate more applications such as ecological research, pet identification, and farm automation [15, 38, 13]. However, the task of understanding animal faces is more challenging than that for the human face due to the lack of sufficient data with annotations. In the literature on full-body understanding, joint learning of interspecies data has been studied as a remedy by transferring the knowledge of well-annotated domain such as human to other species [37, 30, 31]. However, the existing approaches are focused on full-body data, therefore their performance on facial data is limited.

In this work, we are interested in the problem of learning the representation of interspecies facial semantics, which is a common representation shared across interspecies facial data. Learning such representation allows us to extend face understanding beyond the human face, as it enables discovering facial semantics of other species without enough annotations through transferring knowledge from well-annotated human data. As many animal species share similar face geometry, the representation can also benefit from learning multiple species jointly. Since sharing photos of animals through social networks has become more popular, learning the interspecies face representation can be used for

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

(a) Dense Interspecies Face Embedding (DIFE) (b) Applications of DIFE

Figure 1: (a) Dense Interspecies Face Embedding (DIFE) is a representation that contains common characteristics of faces of various species embedded to pixels in the feature space. (b) DIFE allows us to find visual correspondence across multiple face domain, so we can transfer facial keypoints between animals or generate different animal faces with the same posture.

various applications. For instance, one can search animal photos using a photo of human face as a query to find animal faces with similar pose. An animal face animation can be synthesized by using that of a human as a template, which can be useful for media production.

To achieve our goal, we propose to learn Dense Interspecies Face Embedding (DIFE) as shown in Fig. 1(a). Our approach is inspired by CSE [31], which was proposed to learn continuous surface embedding (CSE) of full-body shared across interspecies data. While the interspecies embedding in [31] is promising, it is difficult to use it for faces as the embedding in the facial region is too coarse to discriminate facial landmarks accurately. To overcome this problem, we propose a multi-teacher knowledge distillation method for learning DIFE. Our key idea is to enforce a student network to learn a fine-grained interspecies facial embedding from two teacher networks: CSE and StyleGAN2 [22]. Since the CSE has a useful representation of interspecies data, we use it as a teacher to have the student network learn a common embedding space shared in multiple species. For a fine-grained face embedding, we simultaneously exploit StyleGAN2 as the secondary teacher. StyleGAN2 is originally trained to generate high-quality face images [22]. Recent studies have demonstrated that the features of StyleGAN2 have abundant information on facial attributes such as postures and expressions [14, 40, 41]. Thus, the goal of our knowledge distillation is to transfer such information to the student network for learning fine-grained facial semantics for multiple species.

To further improve the performance of DIFE, we propose a data augmentation method to synthesize pseudo-paired images, and the semantic matching loss to enhance the semantic correspondence between images. For the data augmentation, we synthesize images, where the face poses of the synthesized images are similar to that of input images using StyleGAN2. This can be achieved in our framework as DIFE is tied to the latent features of multiple StyleGAN2 networks in different domains. Thus, DIFE of the input image can be used as a face geometry condition for the latent space exploration of StyleGAN2 to generate images for both intra- and inter-domain pairs. In addition, inspired by [44], a semantic matching loss is proposed to use the pseudo-paired data to enforce the DIFE to learn pixel-level fine-grained matching.

We demonstrate the effectiveness of our method on the cross-domain keypoint transfer task using interspecies data such as people, dogs, cats, and wild animals in MAFL [56] and AP-10K [50]. We evaluate our method on both sparse and dense keypoint transfers. In addition, we apply our method for an image editing task of overlaying a virtual object on various animal faces.

Our contribution can be summarized as follows:

- We present a cross-domain face understanding study that considers not only human faces but also the faces of animals. Our approach makes it possible to avoid tedious and expensive animal data collection through domain adaptation, setting the stage for a variety of cross-domain applications.
- We present a novel multi-teacher knowledge distillation paradigm that extracts and combines the information from models with different architectures and data domains. Our framework is designed to learn continuous face embedding across interspecies data from CSE and fine-grained face embedding from StyleGAN2. Combining the knowledge from the two, our method learns dense interspecies face embedding (DIFE).

- We further introduce a method for synthesizing paired data for learning the semantic matching using the latent space exploration of StyleGAN2. With the novel approach, we can refine the representation of our face embedding for semantic correspondence between various instances and species.

2 Related Works

2.1 Domain-specific Face Understanding

We first review works that aimed to identify the visual correspondence between various instances and scenes within one data category, human or animal. For faces, detecting facial landmarks such as the eyes and the nose has been studied extensively to find correspondences between subjects. Initial works [43, 28, 51, 54, 48, 5] find facial landmarks using facial landmark annotations in a fully supervised way. Regarding animal faces, Yang et al. [49] detected sheep facial landmarks based on 6 annotated keypoints. Although most works show impressive results under extreme conditions, the lack of annotated animal data limits the application compared to human faces.

Recently, there have been active researches [45, 53, 16, 44, 18, 7, 20] on finding landmarks from a collection of images in an unsupervised way, which is applicable to various domains, including human body, face, etc. This stream of works relies on the fact that landmarks are spatially equivariant to the transformation. Some works learn landmarks by comparing warped images with the original ones [45, 53], and others cast the problem into a conditional image generation using video frames depicting the same instances in different poses [16, 18]. The most related work to our method is DVE [44] in that they consider not only equivariance between the same instances but also intra-category variations. All of aforementioned methods can detect landmarks within the same object categories, but the landmarks are not convertible between different object categories (e.g. humans and cats).

2.2 Interspecies Understanding

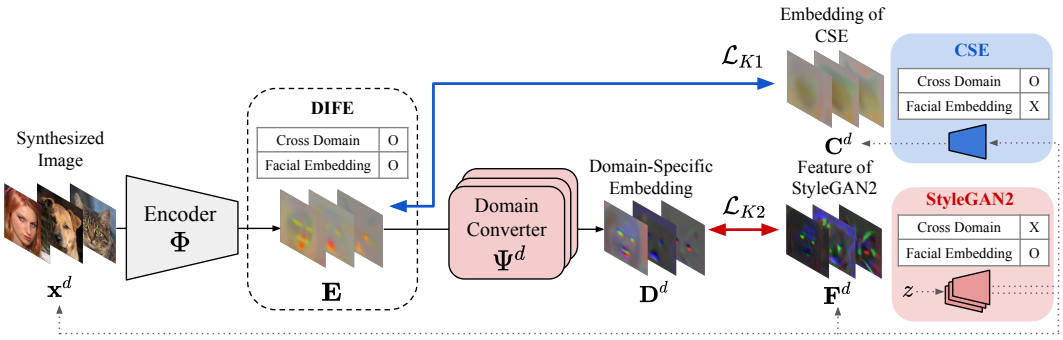
There are studies that aim to discover the relationship between animal species, whose structures are related but geometrically diverse. To the best of our knowledge, [36] is the only work that conducted semantic matching between animal faces. Animal facial landmark detector is trained by transferring the knowledge from the pre-trained human landmark detector. The network is finetuned using horses and sheep data with 5 keypoints annotations. In comparison, we propose an unsupervised approach for interspecies facial semantic matching with the higher granularity of the keypoints.

Closely relevant to our work is the stream of works in [37, 30, 31], as they find the relationship of body poses between animals. Sanakoyeu et al. [37] transfers the knowledge of the human body to the chimpanzee using DensePose [11]. The CSE [30] utilizes the manually annotated 3D meshes of each category and continuous surface embedding to relate several animal categories flexibly. In the following work [31], they utilize the cycle-consistency between meshes and images to further the performance. Although they show impressive results on the complete body, they have difficulty in achieving a fine-level sharpness on faces, making it impossible to discern different parts of faces. While our work also takes the advantage of CSE [31] to detect interspecies correspondences, we improve the accuracy of facial correspondence with the help of the latent features of StyleGAN2 [22] without additional annotations.

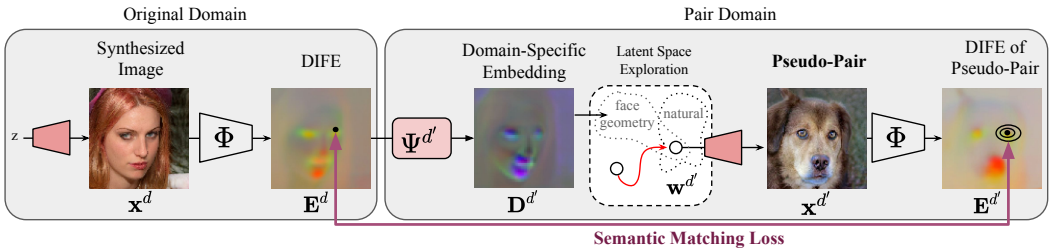
2.3 StyleGAN2 Utilization

Our work is also related to StyleGAN utilization works in that we use features of pretrained StyleGAN2 [22] to obtain fine-grained facial features. StyleGANs [21, 22] show highly photo-realistic results in the unconditional image generation problem with the power of well disentangled latent spaces. There have been abundant works to exploit the pre-trained StyleGAN for various tasks, including image manipulation [46, 14, 41, 39], 3D reconstruction [34], image segmentation [25, 2], and semantic matching [35].

Because StyleGAN2 has plenty of face geometry many works try to extract the information with knowledge distillation [2, 33, 46]. Labels4Free [2] and Segmentation in style [33] utilizes the feature of StyleGAN2 to train segmentation model in unsupervised manner. In StyleGAN2 distillation [46],



(a) Multi-Teacher Knowledge Distillation



(b) Pseudo-Pair Data Synthesis and Utilization

Figure 2: (a) Our encoder learns the cross domain association and the geometry of face with the responses and the intermediate feature from CSE and multiple StyleGAN2's. The converters for each domain is necessary to generate the domain-specific embeddings which is compatible with the intermediate features of StyleGAN2s. (b) The pseudo pair data of different domains can be synthesized with the domain-specific embedding and the latent exploration of StyleGAN2. The semantic matching loss is utilized to overcome the shape difference between faces of different species.

they generate images and labels with StyleGAN2 generator and face attribute classifier, and image manipulation model is trained with synthesized data. However we can tell no studies have used StyleGAN models learned in different data domains at the same time, while our work learns the relationship between separate StyleGAN2 models, which have different data domains.

3 Method

Let $\mathbf{x}^d \in \mathbb{R}^{H \times W \times 3}$ denote an image of a face in a domain $d \in \{0, 1, ..., D-1\}$, where H and W are the height and the width of the image and D denote the number of domains. The goal of our task is to obtain a dense interspecies face embedding $\mathbf{E} \in \mathbb{R}^{H' \times W' \times n}$ of the image $\mathbf{x}^d$ through an encoder Φ, which is described as $\mathbf{E} = \Phi(\mathbf{x}^d)$. Here, n is the embedding dimension for dense interspecies face embedding. To train the encoder Φ in an unsupervised manner, we propose two training approaches: the multi-teacher knowledge distillation, and the semantic matching through pseudo-paired data synthesis. Since all images and labels are synthesized by the teacher models whose training data are not suitable for our task, the training can be considered as an unsupervised learning. In the following, we describe our dense interspecies face embedding (DIFE) and its training in detail.

3.1 Multi-Teacher Knowledge Distillation

Fig. 2(a) illustrates the overview of the multi-teacher knowledge distillation on multiple domains consisting of humans, dogs, and cats. Desirable properties of the interspecies face embedding are two folds. First, the representation is shared by various face images of different domains. Faces in multiple domains should have similar embeddings if their geometry is similar. Second, the embedding contains fine-grained details of facial semantics, so that different parts of a face are distinguishable.

To learn such properties, we first train DIFE by distilling knowledge from CSE [31]. As the representation of CSE has high-quality co-embeddings of diverse animal species, we exploit it to

learn a common space for cross-domain faces. Specifically, we optimize a knowledge distillation loss that minimizes the distance between the DIFE and the CSE embedding of the same input. Formally, the loss is described as

$$\mathcal{L}_{K1}^{d} = \sum \left\| \mathbf{C}^{d} - \Phi(\mathbf{x}^{d}) \right\|_{2}, \tag{1}$$

where $\mathbf{C}^d$ is the CSE embedding of $\mathbf{x}^d$. Note that $\mathbf{x}^d$ is an image synthesized by StyleGAN2 described as dotted arrow in Fig. 2(a).

Although the above loss enforces our embedding to learn a shared space for multiple domains of faces, the representation of CSE is coarse in the region of the face since it is originally learned to distinguish different parts of a full body. To further learn fine-grained details of faces, we add another knowledge distillation loss using the features of StyleGAN2. Specifically, we exploit the semantic information of faces extracted from a pre-trained StyleGAN2 to enhance the representation of DIFE. Using synthetic images generated by the StyleGAN2 as input, we minimize the distance between the embedding of the input and its corresponding StyleGAN2 feature map, which is formulated as

$$\mathcal{L}_{K2}^{d} = \sum \left\| \mathbf{F}^{d} - \mathbf{D}^{d} \right\|_{2}, \tag{2}$$

where $\mathbf{F}^d$ is the StyleGAN2 feature map, and $\mathbf{D}^d$ is the domain-specific embedding of $\mathbf{x}^d$ for the domain d. Since the dimension of DIFE is different from that of StyleGAN2 and each feature of StyleGAN2s is in the different domain, we convert DIFE to a domain-specific embedding using a domain converter Ψ^d, as follows: $\mathbf{D}^d = \Psi^d(\mathbf{E}^d)$. When learning from multiple domains, we optimize $\mathcal{L}_{K1} = \sum_d^D \mathcal{L}_{K1}^d$ and $\mathcal{L}_{K2} = \sum_d^D \mathcal{L}_{K2}^d$ for all of the domains.

3.2 Semantic Matching

Pseudo-paired data synthesis. Fig. 2(b) shows how to synthesize pseudo-paired data $\mathbf{x}^{d'}$, which has the identical face geometry with the original image $\mathbf{x}^d$. The paired domain d' can be different from the original domain d. From a given human face image $\mathbf{x}^d$ generated by StyleGAN2, we first extract DIFE using the encoder Φ. Then, with the domain converter $\Psi^{d'}$, we transform the interspecies embedding into a domain-specific embedding $\mathbf{D}^{d'}$. In Fig. 2(b), we convert DIFE into the dog-specific embedding. Note that the domain-specific embedding lies in the similar vector space with StyleGAN2, since the domain converter is trained by the StyleGAN2 knowledge distillation. Finally, we synthesize a pseudo-paired image $\mathbf{x}^{d'}$ by feeding the domain-specific embedding $\mathbf{D}^{d'}$ to StyleGAN2. Rather than directly feeding the embedding, we introduce a latent space exploration method, which generates more realistic images by adjusting the latent space of the domain-specific embeddings into the natural image manifold.

In the latent space exploration, given a domain-specific embedding $\mathbf{D}^{d'}$, the goal is to find a latent code $\mathbf{w}^{d'} \in \mathcal{W}^{d'}$, which satisfies the following two conditions. Here, $\mathcal{W}^{d'}$ is a latent space of StyleGAN2 trained on the data domain d'. First, the pseudo-paired data $\mathbf{x}^{d'}$ generated from the found latent code $\mathbf{w}^{d'}$ should have the same face geometry with the original image $\mathbf{x}^d$. This is achieved by finding $\mathbf{w}^{d'}$ with small difference between the StyleGAN2 feature $\mathbf{F}^{d'}(\mathbf{w}^{d'})$ that is generated from $\mathbf{w}^{d'}$, and the domain-specific embedding $\mathbf{D}^{d'}$. Second, the pseudo-paired data needs to be realistic and natural. To this end, following [1], we take the mean of latent codes $\bar{\mathbf{w}}$, and enforce our latent code to be similar to $\bar{\mathbf{w}}$. $\bar{\mathbf{w}}$ is calculated by sampling 4096 latent codes. With the regularization, the generated image looks natural by locating the latent code in the realistic image manifold. Finally, the latent code $\mathbf{w}^{d'}$ is found by optimizing the following equation:

$$\mathbf{w}^{d'} = \underset{\mathbf{w} \in \mathcal{W}^{d'}}{\arg\min} \left[\left\| \mathbf{F}^{d'}(\mathbf{w}) - \mathbf{D}^{d'} \right\|_{2} + \left\| \mathbf{w} - \bar{\mathbf{w}} \right\|_{2} \right]. \tag{3}$$

Here, the feature map $\mathbf{F}^{d'}(\mathbf{w})$ is extracted from StyleGAN2 by feeding the latent code $\mathbf{w}$ to optimize. After the optimization, the pseudo-paired data $\mathbf{x}^{d'}$ is synthesized by StyleGAN2 with $\mathbf{w}^{d'}$.

Semantic matching loss. The purpose of the semantic matching loss is to boost the matching performance of DIFE at pixel level with the synthesized pseudo-paired data. The pixel location $u \in \Omega$ on 2D cartesian coordinate system $\Omega = \{0, 1, ..., H' - 1\} \times \{0, 1, ..., W' - 1\}$ is used to explain

the loss. Theoretically, the pixel embedding $\mathbf{E}_u^d$ at location u in DIFE of the original image has the same semantic meaning with the corresponding pixel embedding $\mathbf{E}_u^{d'}$ of the pseudo-paired image. However, the shape difference between different domains makes it infeasible to exactly match the pixel locations. For example, the eyes of the original image $\mathbf{x}^d$ and the eyes of the pseudo-paired data $\mathbf{x}^{d'}$ in Fig. 2(b) do not perfectly match. To alleviate this issue, we introduce a soft distance measure similar to [44]. Intuitively, we compare the embedding from a pixel on the original image to the weighted embedding of a small region on the pseudo-paired image, as described as the contour in Fig. 2(b). We find the region to compare based on the similarity $\sigma(\mathbf{E}^d, \mathbf{E}^{d'}, u_i, u_j)$ between the pixel embeddings as:

$$\sigma(\mathbf{E}^d, \mathbf{E}^{d'}, u_i, u_j) = \frac{\mathbf{E}_{u_i}^d \cdot \mathbf{E}_{u_j}^{d'}}{\left\|\mathbf{E}_{u_i}^d\right\| \left\|\mathbf{E}_{u_j}^{d'}\right\|}, \tag{4}$$

where u_i and u_j are the pixel coordinates of $\mathbf{E}^d$ and $\mathbf{E}^{d'}$, respectively. Finally, the semantic matching loss $\mathcal{L}_M^{d,d'}$ is defined as follows:

$$\mathcal{L}_M^{d,d'} = \sum_{u_i \in \Omega} \left\| u_i - \sum_{u_j \in \Omega} \sigma(\mathbf{E}^d, \mathbf{E}^{d'}, u_i, u_j) u_j \right\|_2. \tag{5}$$

When learning from multiple domains, we optimize intra semantic matching loss $\mathcal{L}_{M1} = \sum_d^D \mathcal{L}_M^{d,d}$ and inter semantic matching loss $\mathcal{L}_{M2} = \sum_d^D \sum_{d' \neq d}^D \mathcal{L}_M^{d,d'}$ for all permutations of the domains.

3.3 Training Objective

We can formulate our final training objective with hyper-parameters as:

$$\mathcal{L} = \lambda_1 \cdot \mathcal{L}_{K1} + \lambda_2 \cdot \mathcal{L}_{K2} + \lambda_3 \cdot \mathcal{L}_{M1} + \lambda_4 \cdot \mathcal{L}_{M2}. \tag{6}$$

The matching losses $\mathcal{L}_{M1}$ and $\mathcal{L}_{M2}$ cannot be applied in the early stage of training, since the domain-specific embedding is different from the features of StyleGAN2s. Therefore, we first impose $\mathcal{L}_{K1}$ and $\mathcal{L}_{K2}$ on the encoder and domain converter, and after initialization, we start to generate pseudo-paired data and utilize semantic matching losses $\mathcal{L}_{M1}$ and $\mathcal{L}_{M2}$.

4 Experiments

4.1 Interspecies Keypoint Transfer

Given an image in the source domain with keypoints, the goal of interspecies keypoint transfer is to find the corresponding locations of the keypoints on another image in the target domain. We mainly evaluate our method on this task to demonstrate that our embedding is not only a high quality representation of cross-domain faces, but also beneficial for knowledge transfer to unlabeled domains. The task is basically achieved by finding the pixel-level correspondence between the embeddings of the source and the target images. The performance is quantitatively evaluated by calculating the error between the positions of the transferred keypoints and their corresponding ground truth. We further describe mathematical formulations of the evaluation in detail in the supplementary material.

Datasets. At the training time, DIFE is trained with synthetic images and their target labels generated by pre-trained CSE and StyleGAN2 networks. We use four datasets for evaluation: MAFL [56], AP-10K [50], WFLW [47] and AnimalWeb [23]. The MAFL is a dataset of human faces that is composed of 20k images. The test set of MAFL excluding the shared images with the training set of CelebA is used (400 images). The AP-10K dataset is composed of 10k images of various animal species which is split into three disjoint subsets, i.e., train, validation, and test sets, with the ratio of 7:1:2 per animal species. As the images contain the full-body of animals, we crop the region of their faces by computing the bounding box using facial landmarks in the dataset. For the evaluation, we make pairs of two different species by randomly selecting images from each domain. As a result, we collect 100 pairs of humans and dogs, 89 pairs of human and cats, 47 pairs of humans and wild animals, and 89 pairs of dogs and cats. The same preprocessing is applied to WFLW and AnimalWeb to evaluate our method on more dense keypoint annotations including the mouth.

Table 1: Quantitative results of interspecies keypoint transfer on MAFL and AP-10K. The pixel error is obtained by calculating the euclidean distance between the ground truth and the transferred keypoints. * indicates that the model is trained by the authors of this paper, as the pre-trained weights are not provided. † indicates that the data domains for the learning and the evaluation do not match.

Method	Pixel Error				
	Human+Dog	Human+Cat	Dog+Cat	Human+Dog+Cat	Human+Wild †
CSE	19.00	18.15	8.00	15.20	17.69
DVE*	17.78	17.40	6.75	14.34	15.58
CATs	14.78	17.63	8.47	13.67	14.05
Ours(DIFE)	**11.73**	**11.00**	**6.51**	**13.16**	**10.37**

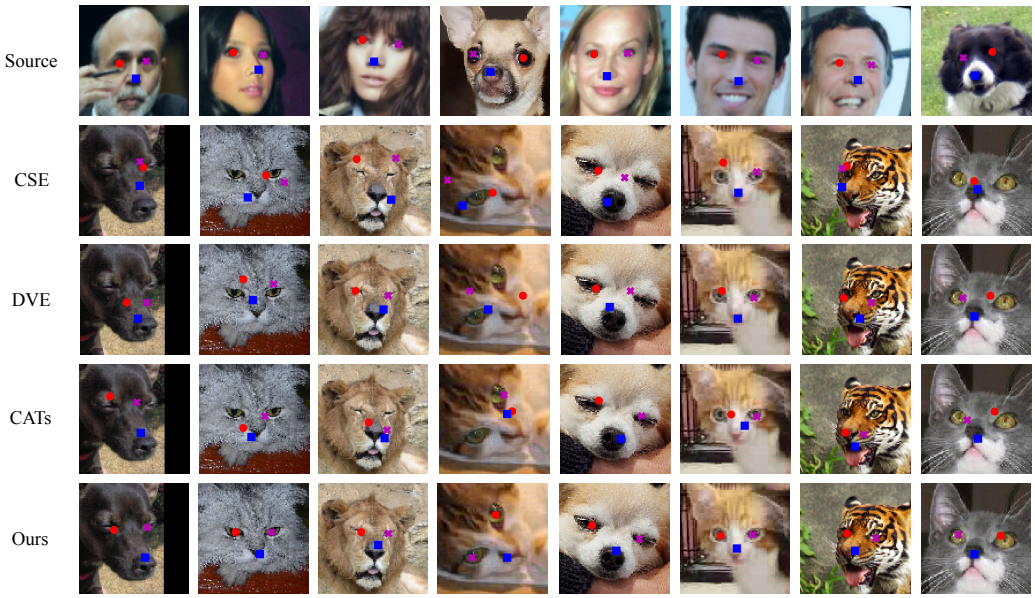


Figure 3: Qualitative results of interspecies keypoint transfer on MAFL and AP-10K.

Implementation details. We pre-train CSE with DensePose-COCO [26] and DensePose-LVIS [12], which are the datasets for full-body keypoints of human and animal respectively. StyleGAN2 is pre-trained with FFHQ [22] for human and AFHQ [9] for animals. DIFE encoder uses the hourglass network [32] for a fair comparison with DVE [44], and the domain converter consists of three consecutive 7×7 conv layers. We use Adam optimizer [24] with the learning rate of 10^{-3}, batch size of 12, and max training step of 1M with early stopping. The lambda values for each loss functions are fixed to $\lambda_1 = 10^{-2}$, $\lambda_2 = 10^{0}$, $\lambda_3 = 10^{-2}$, and $\lambda_4 = 10^{-2}$ in all experiments. All experiments are carried out on one NVIDIA Titan XP.

Baselines. We compare our method with three baselines: CSE [31], DVE [44] and CATs [8]. Even though CSE is originally learned for full-body embeddings of interspecies data, we compare to the method to demonstrate the effectiveness of our method on face images. DVE is an unsupervised learning approach for learning pixel-level facial embeddings, which is designed for human faces in the original work. For the comparison, we train the method with interspecies data. Both the training set of AP-10K and synthesized images of StyleGAN2 were used, and the latter is selected for better performance. We implemented the baselines using the original source code available on their websites with the default hyperparameters. CATs is the state-of-the-art for finding visual semantic correspondence. We use the pre-trained model from the original paper trained on SPair-71k [29] that contains annotations including the landmarks of dogs, cats, and other animals.

Quantitative results. Table 1 presents the quantitative results of CSE, DVE, and DIFE on humans, dogs, cats, and wild animals showing that our method achieves better performance in all data domains.

Table 2: The interspecies keypoint transfer on WFLW [47] and AnimalWeb [23].

	Human+Dog	Human+Cat	Dog+Cat	Human+Wild
CSE	14.43	*14.95*	*16.85*	16.08
DVE	*14.16*	13.28	13.40	*15.74*
CATs	20.84	18.84	19.35	22.20
Ours	**12.01**	**12.84**	**11.70**	**14.03**

As can be seen in the table, transferring between human and other animals is more challenging due to the larger shape differences between faces of human and animals. Compared to other baselines, DIFE makes a significant improvement with the help of the semantic loss. The errors in all methods are lower for Dog+Cat, with DIFE performing better than other baselines. We have also tested a challenging case of the keypoint transfer between three domains by training the models at once with human, dogs, and cats, in which DIFE also outperforms other baselines. For the Human+Wild case, we utilized the model that was trained with human+cat, because it was difficult to train with human and wild animals due to the extreme shape and style variance. Since the wild animal domain includes animals that resemble cats, such as tigers, lions, and cheetahs, using our model trained with human and cats performed reasonably, showing that our method is applicable for out-of-domain cases. Our method also outperforms the CATs which indicates the applicability of our method to discovering landmarks in unlabeled animal domains.

Table 2 shows the quantitative result of the interspecies keypoint transfer on WFLW [47] and AnimalWeb [23] with 9 landmarks including the corners of the eye and mouth. Our method shows the best performance compared to previous methods on every domain pair.

Qualitative results. Fig. 3 shows some qualitative results of the keypoint transfer. As expected, CSE has difficulty in finely distinguishing different parts of faces as it has limited face prior for animals due to the sparse annotation of DensePose-LVIS. While DVE tends to recognize the eyes and the noses to some extent, it does not properly overcome the shape difference, showing some mismatches. DIFE, with its ability to understand cross-domain association and face embedding, can figure out exact visual correspondence on interspecies faces. More visualizations of the interspecies keypoint transfer including the results on WFLW and AnimalWeb can be found in the appendix.

Fig. 4 visualizes the embeddings of different methods. CSE shows similar values across domains, but less distinction between different parts of a face. On the other hand, StyleGAN2 features exhibit distinctiveness between different parts of a face, but corresponding parts between different domains show a large difference in values, making it difficult to align between different domains. DIFE makes the best of both worlds, showing distinctive values within a face, while having similar values between corresponding face parts between different domains.

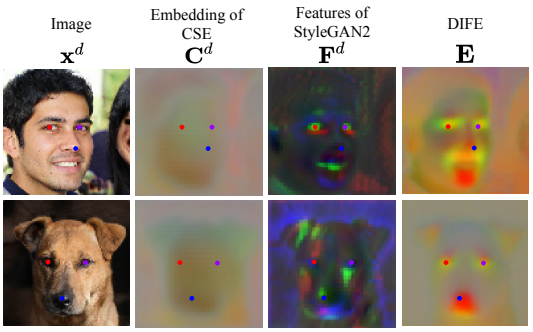


Figure 4: Visualization of embeddings

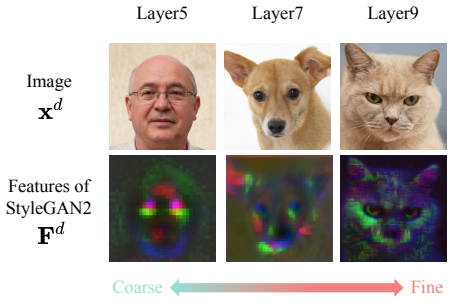


Figure 5: Visualization of features of StyleGAN2

Table 3: Ablation study on losses

$\mathcal{L}_{K1}$	$\mathcal{L}_{K2}$	$\mathcal{L}_{M1}$	$\mathcal{L}_{M2}$	Pixel Error
✓				19.33
✓	✓			13.59
✓	✓	✓	✓	**11.73**
	✓	✓	✓	17.45
✓		✓	✓	-
✓	✓		✓	13.03
✓	✓	✓		12.22

Table 4: Quantitative results of choosing layers of StyleGAN2 for the knowledge distillation.

Layer Number	n^d	Pixel Error
Layer5	512	13.80
Layer7	512	**11.73**
Layer9	256	14.78

Figure 6: The result of image manipulation

Figure 7: The face parsing results based on DIFE

4.2 Ablation Study

The effect of loss functions. Table 3 shows an ablation study to verify the role of each loss. Pixel errors are reduced by adding each loss, and any absence of a loss leads to the performance drop. As we cannot synthesize pseudo-paired data without the knowledge distillation of StyleGAN2, the result for the absence of $\mathcal{L}_{K2}$ is not provided. The absence of $\mathcal{L}_{K1}$ leads to a large performance drop of DIFE, and we find that the role of knowledge distillation of CSE is essential in the early stages of learning. The semantic matching loss also helps to increase the performance by learning more fine visual correspondence across the data domains.

Which layer in StyleGAN2 is good for the knowledge distillation? Table 4 shows the result of selecting a different layer of StyleGAN2 for acquiring the features. Fig. 5 shows that the features from the higher layer is more fine-grained but domain-specific, which makes the domain converter to become confused. Also, because the dimension of the domain-specific embedding follows the dimension of StyleGAN2 features n^d, selecting the ninth layer limits the representation power of the domain embedding. On the other hand, the feature from the lower layer is easy for the knowledge distillation with the abundant representation power with large n^d, but the information about the face is limited because of its coarseness. With this ablation study, we selected the seventh layer of StyleGAN2 to learn face prior of each data domain.

4.3 Applications

Image manipulation. Fig. 6 presents image manipulation examples based on DIFE. Putting objects like glasses and masks requires the positions of face parts, and the result shows that DIFE finds exact locations to put the objects onto the images, even on animal faces.

Interspecies face parsing Fig. 7 shows the results of interspecies face parsing based on DIFE. Following the segmentation in style [33], we use k-means clustering by DIFE for unsupervised face parsing. The eye, nose, mouth, and hairy parts are discovered with a simple method which means DIFE has proper semantic information for the interspecies face.

Dense keypoint transfer. Although DIFE extracts embeddings from all pixels, we cannot accurately evaluate all visual correspondences due to the lack of annotations. However, we can figure out more dense visual correspondence qualitatively by transferring generated grid keypoints . In Fig. 8,

Figure 8: The result of dense keypoint transfer

Figure 9: The examples of pseudo pair data

the result of transferring dense keypoints from a human source to another human, dog, cat, and wild animal targets can be found.

Pseudo-paired data synthesis. Fig. 9 shows examples of pseudo-paired data with various data domains and postures. Given original images, we perform pseudo-paired data synthesis on the intra-domain and the inter-domain. We can see that the generated paired images are in similar poses with the original ones, while having minor texture variations.

5 Conclusion

In this paper, we introduced a study for extracting common features from faces of several species, including humans, as a dense embedding. In this process, we presented a mechanism for extracting and learning desired information from separate teacher models, even though the teacher models have different architectures and data domains. Furthermore, we proposed a new data augmentation technique with a latent space exploration of StyleGAN2 and the semantic matching loss to overcome the extreme shape difference between species. Various experiments and applications were offered for the evaluation, and our approach has been shown to be superior to other baselines. For the future research directions, we need to deal with multiple animals more than three domains efficiently, which is not focused on this paper due to the absence of pre-trained StyleGAN2. A research on the generalization method for the out-of-domain data is needed to address the situation where data cannot be obtained or the animal species is unknown. Our method is dependent on the performance of pre-trained CSE and StyleGAN2. Lastly, there are typical face understanding failure cases like occlusion or dark illumination. As the first work to tackle the difficult problem of interspecies face embedding, more follow-up studies are necessary to overcome current limitations. Nevertheless, we believe that this work has set the stage for a new research topic of cross-domain study for interspecies face understanding.

6 Broader Impact.

This study aims to understand the interspecies face, which can be potentially extended to various applications. For example, our method can assist to find lost pets, check the health status of farm animals through analyzing the facial expressions. Identifying wild animals for ecological experiments is another expected application of our method. Moreover, the proposed method can potentially be extended to the automatic motion capture, that can be used in the film production or content creation for AR/VR. We note that the studies using DIFE could potentially violate the personal privacy or animal rights. The studies extending DIFE should carefully consider the above potential misuses, and bring positive impacts to the society.

7 Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2022-0-00124, Development of Artificial Intelligence Technology for Self-Improving Competency-Aware Learning Capabilities), and No. 2020-0-01361, Artificial Intelligence Graduate School Program (Yonsei University).

References

[1] Abdal, R., Qin, Y., Wonka, P.: Image2stylegan++: How to edit the embedded images? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8296–8305 (2020)

[2] Abdal, R., Zhu, P., Mitra, N.J., Wonka, P.: Labels4free: Unsupervised segmentation using stylegan. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 13970–13979 (2021)

[3] Abdal, R., Zhu, P., Mitra, N.J., Wonka, P.: Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. ACM Transactions on Graphics (TOG) **40**(3), 1–21 (2021)

[4] Barsoum, E., Zhang, C., Canton Ferrer, C., Zhang, Z.: Training deep networks for facial expression recognition with crowd-sourced label distribution. In: ACM International Conference on Multimodal Interaction (ICMI) (2016)

[5] Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2017)

[6] Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: IEEE International Conference on Automatic Face & Gesture Recognition (F&G). pp. 67–74. IEEE (2018)

[7] Cheng, Z., Su, J.C., Maji, S.: On equivariant and invariant learning of object landmark representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9897–9906 (2021)

[8] Cho, S., Hong, S., Jeon, S., Lee, Y., Sohn, K., Kim, S.: Cats: Cost aggregation transformers for visual correspondence. Advances in Neural Information Processing Systems (NIPS) **34** (2021)

[9] Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)

[10] Feng, Z.H., Kittler, J., Awais, M., Huber, P., Wu, X.J.: Wing loss for robust facial landmark localisation with convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2235–2245 (2018)

[11] Güler, R.A., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7297–7306 (2018)

[12] Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)

[13] Hansen, M.F., Smith, M.L., Smith, L.N., Salter, M.G., Baxter, E.M., Farish, M., Grieve, B.: Towards on-farm pig face recognition using convolutional neural networks. Computers in Industry **98**, 145–152 (2018)

[14] Härkönen, E., Hertzmann, A., Lehtinen, J., Paris, S.: Ganspace: Discovering interpretable gan controls. Advances in Neural Information Processing Systems (NIPS) **33**, 9841–9850 (2020)

[15] Hummel, H.I., Pessanha, F., Salah, A.A., van Loon, T.J., Veltkamp, R.C.: Automatic pain detection on horse and donkey faces. In: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). pp. 793–800. IEEE (2020)

[16] Jakab, T., Gupta, A., Bilen, H., Vedaldi, A.: Unsupervised learning of object landmarks through conditional image generation. Advances in Neural Information Processing Systems (NIPS) **31** (2018)

[17] Jakab, T., Gupta, A., Bilen, H., Vedaldi, A.: Unsupervised learning of object landmarks through conditional image generation. Advances in Neural Information Processing Systems (NIPS) **31** (2018)

[18] Jakab, T., Gupta, A., Bilen, H., Vedaldi, A.: Self-supervised learning of interpretable keypoints from unlabelled videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8787–8797 (2020)

[19] Jakab, T., Gupta, A., Bilen, H., Vedaldi, A.: Self-supervised learning of interpretable keypoints from unlabelled videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8787–8797 (2020)

[20] Jiang, J., Ji, Y., Wang, X., Liu, Y., Wang, J., Long, M.: Regressive domain adaptation for unsupervised keypoint detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6780–6789 (2021)

[21] Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4401–4410 (2019)

[22] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8110–8119 (2020)

[23] Khan, M.H., McDonagh, J., Khan, S., Shahabuddin, M., Arora, A., Khan, F.S., Shao, L., Tzimiropoulos, G.: Animalweb: A large-scale hierarchical dataset of annotated animal faces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6939–6948 (2020)

[24] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. International Conference for Learning Representations (ICLR) (2015)

[25] Li, D., Yang, J., Kreis, K., Torralba, A., Fidler, S.: Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)

[26] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proceedings of Proceedings of European Conference on Computer Vision (ECCV) (2014)

[27] Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (December 2015)

[28] Lv, J., Shao, X., Xing, J., Cheng, C., Zhou, X.: A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3317–3326 (2017)

[29] Min, J., Lee, J., Ponce, J., Cho, M.: Spair-71k: A large-scale benchmark for semantic correspondence. arXiv prepreint arXiv:1908.10543 (2019)

[30] Neverova, N., Novotny, D., Khalidov, V., Szafraniec, M., Labatut, P., Vedaldi, A.: Continuous surface embeddings (2020)

[31] Neverova, N., Sanakoyeu, A., Novotny, D., Labatut, P., Vedaldi, A.: Discovering relationships between object categories via universal canonical maps (2021)

[32] Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Proceedings of European Conference on Computer Vision (ECCV). pp. 483–499. Springer (2016)

[33] Pakhomov, D., Hira, S., Wagle, N., Green, K.E., Navab, N.: Segmentation in style: Unsupervised semantic image segmentation with stylegan and clip. arXiv preprint arXiv:2107.12518 (2021)

[34] Pan, X., Dai, B., Liu, Z., Loy, C.C., Luo, P.: Do 2d gans know 3d shape? unsupervised 3d shape reconstruction from 2d image gans. In: International Conference for Learning Representations (ICLR) (2021)

[35] Peebles, W., Zhu, J.Y., Zhang, R., Torralba, A., Efros, A., Shechtman, E.: Gan-supervised dense visual alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)

[36] Rashid, M., Gu, X., Jae Lee, Y.: Interspecies knowledge transfer for facial keypoint detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6894–6903 (2017)

[37] Sanakoyeu, A., Khalidov, V., McCarthy, M.S., Vedaldi, A., Neverova, N.: Transferring dense pose to proximal animal classes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5233–5242 (2020)

[38] Schneider, S., Taylor, G.W., Kremer, S.C.: Similarity learning networks for animal individual re-identification-beyond the capabilities of a human observer. In: IEEE International Conference on Automatic Face & Gesture Recognition (F&G). pp. 44–52 (2020)

[39] Shen, Y., Yang, C., Tang, X., Zhou, B.: Interfacegan: Interpreting the disentangled face representation learned by gans. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2020)

[40] Shen, Y., Zhou, B.: Closed-form factorization of latent semantics in gans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1532–1540 (2021)

[41] Shoshan, A., Bhonker, N., Kviatkovsky, I., Medioni, G.: Gan-control: Explicitly controllable gans. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14083–14093 (2021)

[42] Shu, Z., Sahasrabudhe, M., Guler, R.A., Samaras, D., Paragios, N., Kokkinos, I.: Deforming autoencoders: Unsupervised disentangling of shape and appearance. In: Proceedings of Proceedings of European Conference on Computer Vision (ECCV). pp. 650–665 (2018)

[43] Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3476–3483 (2013)

[44] Thewlis, J., Albanie, S., Bilen, H., Vedaldi, A.: Unsupervised learning of landmarks by descriptor vector exchange. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6361–6371 (2019)

[45] Thewlis, J., Bilen, H., Vedaldi, A.: Unsupervised learning of object landmarks by factorized spatial embeddings. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5916–5925 (2017)

[46] Viazovetskyi, Y., Ivashkin, V., Kashin, E.: Stylegan2 distillation for feed-forward image manipulation. In: Proceedings of European Conference on Computer Vision (ECCV). pp. 170–186. Springer (2020)

[47] Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., Zhou, Q.: Look at boundary: A boundary-aware face alignment algorithm. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2129–2138 (2018)

[48] Xiao, S., Feng, J., Xing, J., Lai, H., Yan, S., Kassim, A.: Robust facial landmark detection via recurrent attentive-refinement networks. In: Proceedings of Proceedings of European Conference on Computer Vision (ECCV). pp. 57–72. Springer (2016)

[49] Yang, H., Zhang, R., Robinson, P.: Human and sheep facial landmarks localisation by triplet interpolated features. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1–8. IEEE (2016)

[50] Yu, H., Xu, Y., Zhang, J., Zhao, W., Guan, Z., Tao, D.: Ap-10k: A benchmark for animal pose estimation in the wild. Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NIPS Datasets and Benchmarks) (2021)

[51] Zhang, J., Shan, S., Kan, M., Chen, X.: Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In: Proceedings of European Conference on Computer Vision (ECCV). pp. 1–16. Springer (2014)

[52] Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (2016)

[53] Zhang, Y., Guo, Y., Jin, Y., Luo, Y., He, Z., Lee, H.: Unsupervised discovery of object landmarks as structural representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2694–2703 (2018)

[54] Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Facial landmark detection by deep multi-task learning. In: Proceedings of European Conference on Computer Vision (ECCV). pp. 94–108. Springer (2014)

[55] Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Facial landmark detection by deep multi-task learning. In: Proceedings of Proceedings of European Conference on Computer Vision (ECCV). pp. 94–108. Springer (2014)

[56] Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Learning deep representation for face alignment with auxiliary attributes. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2015)

[57] Zhang, Z., Song, Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5810–5818 (2017)

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Abstract and Section 1.
   (b) Did you describe the limitations of your work? [Yes] See Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 6.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] See Section 6.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Visit `https://github.com/kingsj0405/dife`
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4.2
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Appendix B.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4.2
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Appendix F.
   (b) Did you mention the license of the assets? [Yes] See Appendix F.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Visit `https://github.com/kingsj0405/dife`
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Appendix F.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Section 6.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

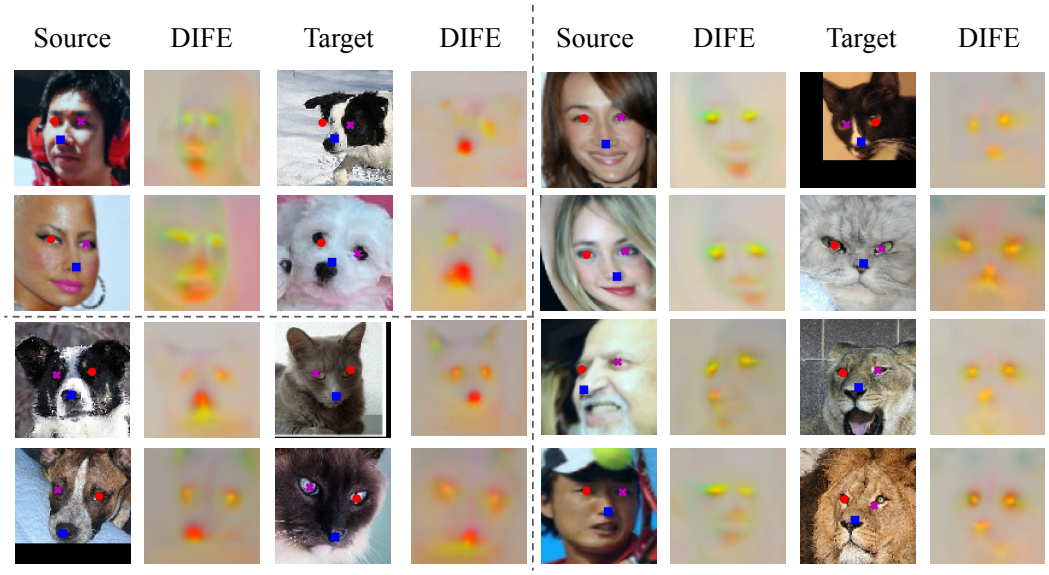


Figure 10: More Qualitative Results with Embedding Visualizations

Table 5: The mean and standard deviation of multiple training.

DIFE	Pixel Error		
	Human+Dog	Human+Cat	Dog+Cat
Mean	11.73	11.00	6.51
Standard Deviation	0.21	0.55	0.55

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 $\mathbf{E}^s$ and target embedding $\mathbf{E}^t$ 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

$$\sigma(\mathbf{E}^s, \mathbf{E}^t, u_i, u_j) = \frac{\mathbf{E}^s_{u_i} \cdot \mathbf{E}^t_{u_j}}{\left\|\mathbf{E}^s_{u_i}\right\| \left\|\mathbf{E}^t_{u_j}\right\|}, \tag{7}$$

$$\hat{u}_t = \arg\min_{u \in \Omega} \sigma(\mathbf{E}^s, \mathbf{E}^t, u_s, u), \tag{8}$$

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:

$$e_p = \left\|u_t - \hat{u}_t\right\|. \tag{9}$$

With lower pixel error, we can assure the learned embedding works well as visual descriptors as matching reliably different face geometry of various species [44].

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 5.

Table 6: The human landmark detection results.

Category	Method	Unsup.	Cross-Domain	MAFL	AFLW$_M$	AFLW$_R$	300W
Supervised learning	TCDCN [55]	X	X	**7.95**	7.65	-	5.54
	MTCNN [52]	X	X	5.39	6.90	-	-
	Wing Loss [10]	X	X	-	-	-	4.04
Generative modeling based	Deforming AE [42]	O	X	**5.45**	-	-	-
	ImGen. [17]	O	X	2.54	-	6.31	-
	ImGEN.++ [19]	O	X	-	-	-	5.12
Equivariance based	Dense 3D [42]	O	X	**4.02**	**10.99**	**10.14**	**8.23**
	DVE [44]	O	X	2.86	7.53	6.54	4.65
	On Equvariant [7]	O	X	2.44	6.99	6.27	5.22
	DIFE	O	O	**3.40**	**10.11**	**8.68**	**7.57**

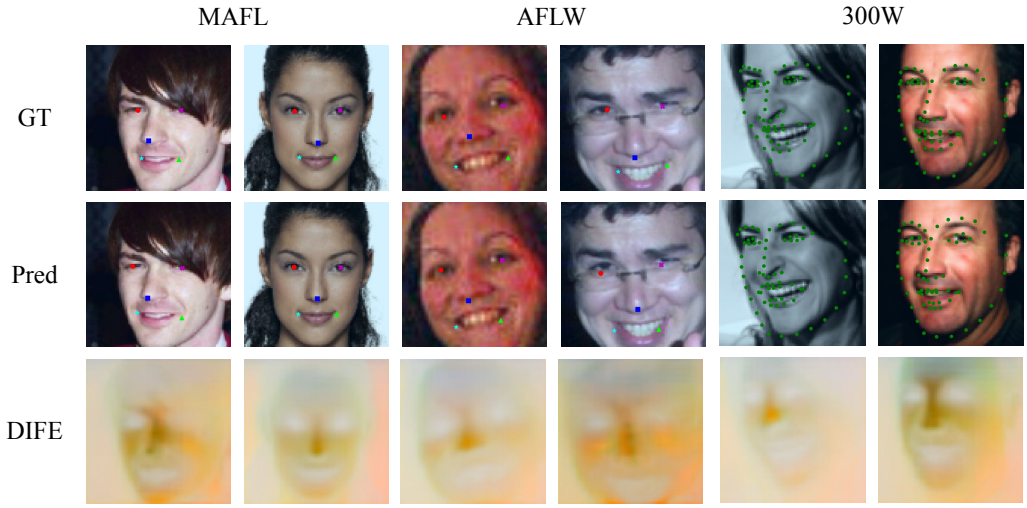


Figure 11: The qualitative results of the human keypoint regression.

C The human landmark detection

We carry out the keypoint regression experiment setup following DVE [44] and On Equivariant And Invariant [7] 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. 11.

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 $\mathbf{x}^{d1}$, $\mathbf{x}^{d2}$ and the feature of StyleGAN2 $\mathbf{F}^{d1}$, $\mathbf{F}^{d2}$ is generated with the StyleGAN2 $\mathcal{G}^{d1}$, $\mathcal{G}^{d2}$. Here $d1$ and $d2$ means the the domain of instance. The continuous surface embedding $\mathbf{C}^{d1}$, $\mathbf{C}^{d2}$ is extracted with the CSE $\mathcal{E}$. Therfore, DIFE $\mathbf{E}^{d1}$, $\mathbf{E}^{d2}$ and domain-specific embedding $\mathbf{D}^{d1\to d1}$, $\mathbf{D}^{d2\to d2}$ are extracted with the encoder Φ and the domain converter Ψ^{d1}, Ψ^{d2}. We utilize superscript $d \to d'$ 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 $\mathcal{L}_{K1}$ and StyleGAN2 knowledge distillation loss $\mathcal{L}_{K2}$ are calculated with

Algorithm 1 The Pipeline of Training For Two Domains

1: **(0) Prepare DIFE training**
2: Set the learning rate η
3: Load pre-trained weights For parameters of CSE $\mathcal{E}$, StyleGAN2 $\mathcal{G}^{d1}$, $\mathcal{G}^{d2}$
4: Initialize network parameters θ_{Φ}, $\theta_{\Psi^{d1}}$, $\theta_{\Psi^{d2}}$
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: $\mathbf{x}^{d1}, \mathbf{F}^{d1} \leftarrow \mathcal{G}^{d1}(z)$
9: $\mathbf{x}^{d2}, \mathbf{F}^{d2} \leftarrow \mathcal{G}^{d2}(z)$
10: $\mathbf{C}^{d1}, \mathbf{C}^{d2} \leftarrow \mathcal{E}(\mathbf{x}^{d1}, \mathbf{x}^{d2})$
11: **(1-2) Extract embedding**
12: $\mathbf{E}^{d1}, \mathbf{E}^{d2} \leftarrow \Phi(\mathbf{x}^{d1}, \mathbf{x}^{d2})$
13: $\mathbf{D}^{d1\to d1} \leftarrow \Psi^{d1}(\mathbf{E}^{d1})$
14: $\mathbf{D}^{d2\to d2} \leftarrow \Psi^{d2}(\mathbf{E}^{d2})$
15: **(1-3) Calculate losses For multi-teacher knowledge distillation**
16: $\mathcal{L}_{K1} = \mathcal{L}_K(\mathbf{E}^{d1}, \mathbf{C}^{d1}) + \mathcal{L}_K(\mathbf{E}^{d2}, \mathbf{C}^{d2})$
17: $\mathcal{L}_{K2}^{d1} = \mathcal{L}_K(\mathbf{D}^{d1\to d1}, \mathbf{F}^{d1})$
18: $\mathcal{L}_{K2}^{d2} = \mathcal{L}_K(\mathbf{D}^{d2\to d2}, \mathbf{F}^{d2})$
19: **if** not initialization period **then**
20: **(2-1) Intra pseudo-paired data synthesis**
21: $\mathbf{x}^{d1\to d1} \leftarrow \mathcal{G}^{d1}(\mathbf{D}^{d1\to d1})$ with latent space exploration
22: $\mathbf{x}^{d2\to d2} \leftarrow \mathcal{G}^{d2}(\mathbf{D}^{d2\to d2})$ with latent space exploration
23: **(2-2) Inter pseudo-paired data synthesis**
24: $\mathbf{D}^{d2\to d1} \leftarrow \Psi^{d1}(\mathbf{E}^{d2})$
25: $\mathbf{D}^{d1\to d2} \leftarrow \Psi^{d2}(\mathbf{E}^{d1})$
26: $\mathbf{x}^{d2\to d1} \leftarrow \mathcal{G}^{d1}(\mathbf{D}^{d2\to d1})$ with latent space exploration
27: $\mathbf{x}^{d1\to d2} \leftarrow \mathcal{G}^{d2}(\mathbf{D}^{d1\to d2})$ with latent space exploration
28: **(2-3) Extract embedding**
29: $\mathbf{E}^{d1\to d1}, \mathbf{E}^{d1\to d2}, \mathbf{E}^{d2\to d1}, \mathbf{E}^{d2\to d2} \leftarrow \Phi(\mathbf{x}^{d1\to d1}, \mathbf{x}^{d1\to d2}, \mathbf{x}^{d2\to d1}, \mathbf{x}^{d2\to d2})$
30: **(2-3) Calculate losses For semantic matching**
31: $\mathcal{L}_{M1} = \mathcal{L}_M(\mathbf{E}^{d1}, \mathbf{E}^{d1\to d1}) + \mathcal{L}_M(\mathbf{E}^{d2}, \mathbf{E}^{d2\to d2})$
32: $\mathcal{L}_{M2} = \mathcal{L}_M(\mathbf{E}^{d1}, \mathbf{E}^{d1\to d2}) + \mathcal{L}_M(\mathbf{E}^{d2}, \mathbf{E}^{d2\to d1})$
33: **else**
34: $\mathcal{L}_{M1} = 0$
35: $\mathcal{L}_{M2} = 0$
36: **end if**
37: **(3) Update network parameters**
38: $\theta_{\Phi} = \theta_{\Phi} + \eta \nabla_{\theta_{\Phi}}(\mathcal{L}_{K1} + \mathcal{L}_{K2} + \mathcal{L}_{M1} + \mathcal{L}_{M2})$
39: $\theta_{\Psi^{d1}} = \theta_{\Psi^{d1}} + \eta \nabla_{\Psi^{d1}}(\mathcal{L}_{K2}^{d1})$
40: $\theta_{\Psi^{d2}} = \theta_{\Psi^{d2}} + \eta \nabla_{\Psi^{d2}}(\mathcal{L}_{K2}^{d2})$
41: **end for**

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

$$\mathcal{L}_K(\mathbf{A}, \mathbf{B}) = \|\mathbf{A} - \mathbf{B}\|_2 . \tag{10}$$

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 $\mathbf{x}^{d1\to d1}$, $\mathbf{x}^{d2\to d2}$ and inter pseudo-paired data $\mathbf{x}^{d1\to d2}$, $\mathbf{x}^{d2\to d1}$ are generated with latent space exploration described at Section 3.2. With encoder Φ, each DIFE for pseudo-paired data is extracted. At last semantic matching loss $\mathcal{L}_M$ are utilized to get intra semantic matching loss $\mathcal{L}_{M1}$ and inter semantic matching loss $\mathcal{L}_{M2}$.

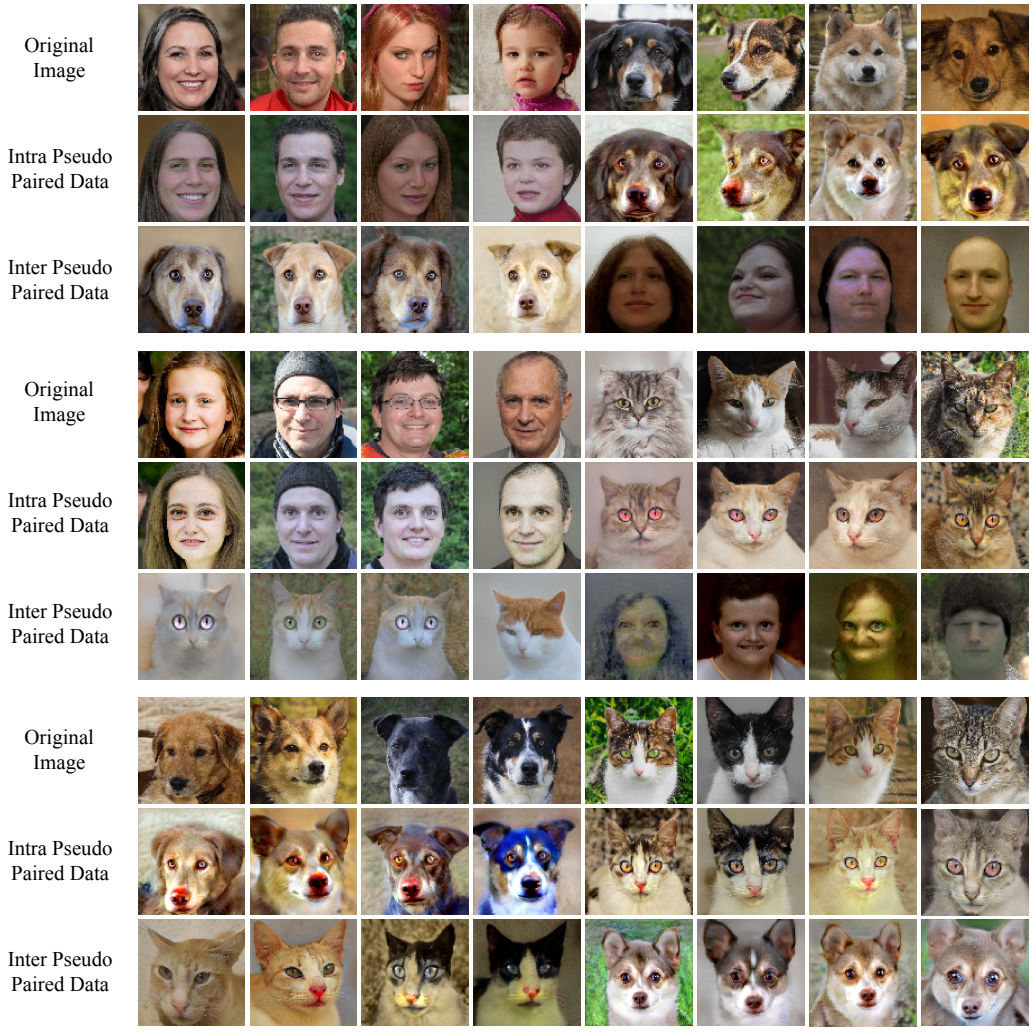


Figure 12: More Examples of Pseudo Paired Data

Here the equation for semantic matching loss with cosine similarity σ defined at Appendix B:

$$\mathcal{L}_M(\mathbf{A}, \mathbf{B}) = \sum_{u_i \in \Omega} \left\| u_i - \sum_{u_j \in \Omega} \sigma(\mathbf{A}, \mathbf{B}, u_i, u_j) u_j \right\|_2 . \tag{11}$$

At the end of iteration, the parameters of the encoder θ_Φ 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. 12. 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.

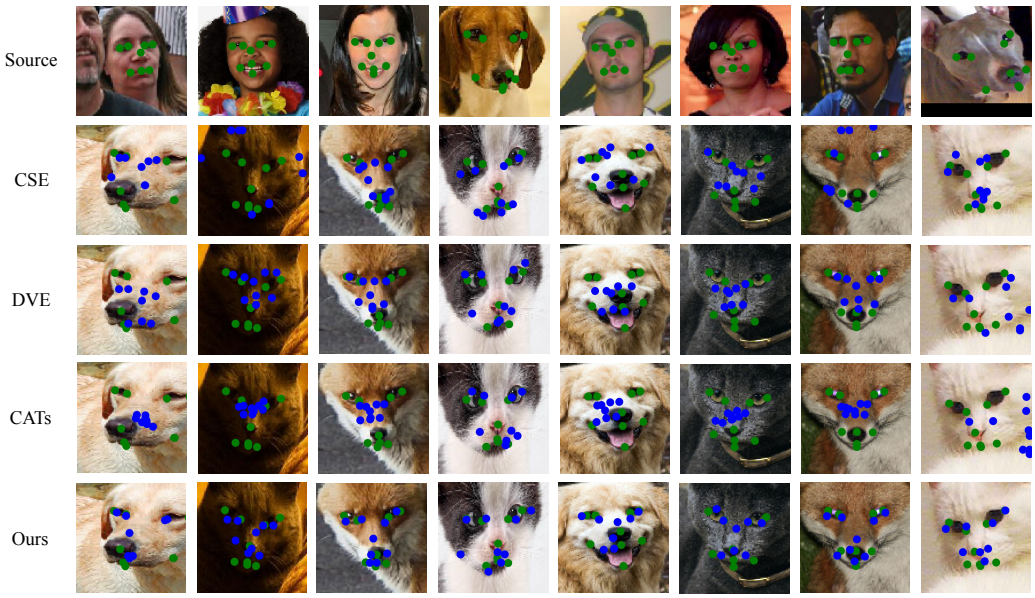


Figure 13: The qualitative results of the interspecies keypoint transfer on WFLW and AnimalWeb. The blue landmarks are predicted and the green landmarks are the ground truth.

F Intellectual property rights of assets

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

- Detectron2, DensePose-CSE [31]: The license of code is Aphache 2.0 License which is an open source license.
- StyleGAN, FFHQ [21]: 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 [9]: The dataset is available under Creative Commons BY-NC 4.0 license by NAVER Corporation. There is no problem to use, copy, tranform 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 [50]: The dataset is available under Creative Commons BY-NC 4.0 license which is same with AFHQ.
- CelebA [27], MAFL [56]: The name of license is unknown but datasets are available for non-commercial research purposes.