Generalizable Face Landmarking Guided by Conditional Face Warping

Jiayi Liang[1] Haotian Liu[1*] Hongteng Xu[2,3] Dixin Luo[1,4†]
[1]School of Computer Science and Technology, Beijing Institute of Technology, Beijing
[2]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing
[3]Beijing Key Laboratory of Big Data Management and Analysis Methods, Beijing
[4]Key Laboratory of Artificial Intelligence, Ministry of Education, Shanghai
{jiayi.liang, haotianliu, dixin.luo}@bit.edu.cn, hongtengxu@ruc.edu.cn

Abstract

As a significant step for human face modeling, editing, and generation, face landmarking aims at extracting facial keypoints from images. A generalizable face landmarker is required in practice because real-world facial images, e.g., the avatars in animations and games, are often stylized in various ways. However, achieving generalizable face landmarking is challenging due to the diversity of facial styles and the scarcity of labeled stylized faces. In this study, we propose a simple but effective paradigm to learn a generalizable face landmarker based on labeled real human faces and unlabeled stylized faces. Our method learns the face landmarker as the key module of a conditional face warper. Given a pair of real and stylized facial images, the conditional face warper predicts a warping field from the real face to the stylized one, in which the face landmarker predicts the ending points of the warping field and provides us with high-quality pseudo landmarks for the corresponding stylized facial images. Applying an alternating optimization strategy, we learn the face landmarker to minimize i) the discrepancy between the stylized faces and the warped real ones and ii) the prediction errors of both real and pseudo landmarks. Experiments on various datasets show that our method outperforms existing state-of-the-art domain adaptation methods in face landmarking tasks, leading to a face landmarker with better generalizability. Code is available at https://plustwo0.github.io/project-face-landmarker.

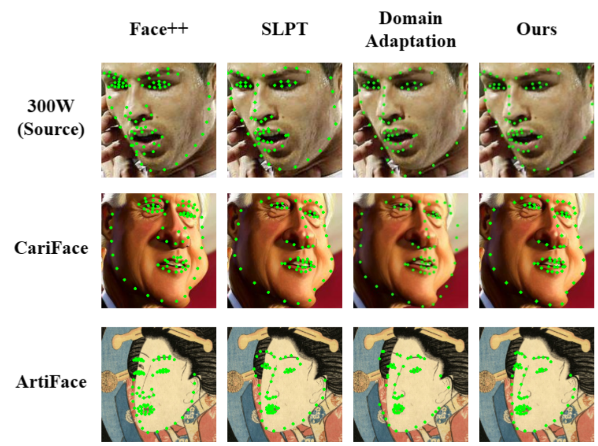


Figure 1. Both commercial software like Face++ and open-source method like SLPT [52] work well on landmarking real faces (e.g., those in 300W [38]) while achieving suboptimal performance when landmarking stylized faces (e.g., those in CariFace [4] and ArtiFace [55]). While existing domain adaptation method does not improve the performance significantly, our method achieves a generalizable face landmarker for various facial images.

1. Introduction

Face landmarking seeks to extract human facial keypoints (*e.g.*, eyes, nose, facial contour, and so on) from facial images. This task is important for many applications in the field of computer vision and graphics, such as face recognition [1, 28, 33, 43, 53], face stylization [5], and 3D face reconstruction [11, 26, 37]. Currently, many open-source and commercial face landmarkers [19, 51, 58] have been developed and achieved encouraging performance in this task.

Most existing face landmarkers are designed and trained for landmarking real human faces, while the rapid development of AIGC applications, such as artistic character creation and cartoon generation [2, 15], leads to a massive increase in demands for landmarking stylized facial images. Unfortunately, as shown in Fig. 1, existing face landmarkers often fail to landmark stylized facial images. Even if applying state-of-the-art domain adaptation strategies [12, 25, 27, 59, 61], the generalizability of the learned landmarkers in the stylized facial image domain is still unsatisfactory. Essentially, traditional face landmarkers work

*The first two authors contributed equally to this work.
†Corresponding author.

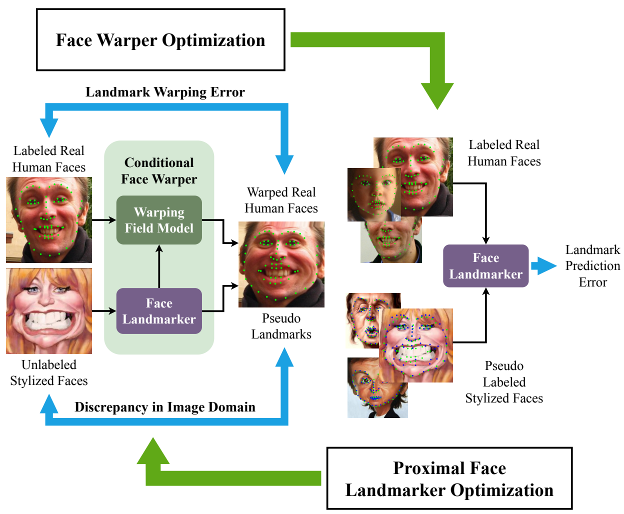


Figure 2. The scheme of our proposed method for learning a generalizable face landmarker.

well on real human faces because of the relatively stable geometry of real human faces and the sufficient labeled facial images. These two conditions, however, become questionable when landmarking stylized faces — the stylized facial images often have various facial styles, and manually landmarking such stylized faces is much more time-consuming than landmarking real faces. As a result, learning a generalizable face landmarker becomes a challenging task.

In this paper, we propose a simple but effective paradigm for learning a generalizable face landmarker, overcoming the challenges caused by the diversity of facial styles and the scarcity of labeled stylized faces. As illustrated in Fig. 2, given labeled real human facial images and unlabeled stylized facial images, we learn a face landmarker embedded in a conditional face warper. The face warper aims to deform real human faces according to stylized facial images, generating warped faces and corresponding warping fields. The face landmarker, as the key module of the warper, predicts the ending points of the warping fields and thus provides us with pseudo landmarks for the stylized facial images. The warping field is parametrized by a polyharmonic interpolation model. Under the guidance of the conditional face warping, we learn the face landmarker in an alternating optimization framework: The face landmarker is updated to i) minimize the discrepancy between the stylized faces and the corresponding warped real human faces and ii) minimize the prediction errors of both real and pseudo landmarks. In the first step, the face landmarker is learned associated with the warping field model, while in the second step, the face landmarker is updated with proximal regularization.

The extensive experiments in various face landmarking tasks demonstrate the effectiveness of learning method. The impacts of different loss functions and data settings on our learning method are analyzed through detailed ablation studies. Experimental results show that our learning method results in a face landmarker that is generalizable to the facial images with different styles, which outperforms representative domain adaptation methods consistently in various stylized face landmarking tasks.

2. Related Work

2.1. Face Landmarking

Given facial images, the early face landmarking methods learn regression models to predict landmark coordinates directly [6, 32]. The models are often parameterized by neural networks like Transformer [46], capturing facial attributes that have been proven to be crucial for landmark prediction [24]. The coordinate regression is achieved in a coarse-to-fine framework [58], leading to a cascading landmarking pipeline. Recently, to make the landmark prediction robust to the variations in pose, scale, and occlusion, DAN [22] introduces the novel use of heatmaps and extracts features from the entire face rather than local patches around landmarks. SBR [9] utilizes the registration of synthesized images to provide supervisory signals for training. Adaptive Wing Loss [48] is proposed to address the imbalance between foreground pixels and background pixels by analysis of the main drawbacks of different loss functions. HRNet [47] produces high-resolution maps by connecting and exchanging information via merging multi-scale picture features across many branches.

As aforementioned, most existing face landmarkers are learned for real human faces, which can not be adapted directly to stylized faces (e.g., cartoon and artistic faces). Essentially, treating labeled real human faces as a source domain and unlabeled stylized faces as a target domain, we can learn a generalizable face landmark by solving a domain adaptation (DA) problem [30, 31]. Accordingly, many DA techniques [27, 49] have potential in our problem, including the classic metric learning-based methods (e.g., CORAL [41], contrastive domain discrepancy [17], and maximum mean discrepancy [16]) and the recent adversarial learning-based methods [12, 35, 56, 59]. However, in the following content, we will show that directly applying these DA techniques often fails to achieve generalizable face landmarking because of the significant gap between the source and target face domains.

2.2. Face Warping

Face warping is a technique that involves geometrically deforming source facial images to specified target shapes. The key step of this task is predicting a warping field between the source and target images that captures the shifts of image pixels. To achieve this aim, DST [20] and FoA [55] find matching keypoints between source and target images and then generate a dense warping field through data interpolation [8]. Instead of matching keypoints, some methods learn neural networks to predict dense warping fields directly based on paired images, e.g., Flownet [10], AutoToon [15], RAFT [44], and their variants [29]. However, these methods require the paired images to be similar to each other, which cannot capture significant deformations between the faces in different domains.

Compared to face stylization [13, 18, 39, 40], face warping is a relatively easier task because it only considers the deformation of shapes while ignoring the transfer of textures. However, it should be noted that this task is more relevant to face landmarking, in which the warping field provides us with strong evidence to shift face landmarks of source faces to target ones [50]. Inspired by such a strong correlation, we develop the proposed learning paradigm.

3. Proposed Method

Denote $\mathcal{X}$ as the image space and $\mathcal{Y}$ as the landmark space, respectively. In this work, we observe a set of labeled real human faces, i.e., $\mathcal{D}^{(L)} = \{\boldsymbol{X}_i^{(L)}, \boldsymbol{Y}_i^{(L)}\}_{i=1}^{N_L} \subset \mathcal{X}_R \times \mathcal{Y}$ and a set of unlabeled stylized faces, i.e., $\mathcal{D}^{(U)} = \{\boldsymbol{X}_i^{(U)}\}_{i=1}^{N_U} \subset \mathcal{X}_S$, where $\mathcal{X}_R, \mathcal{X}_S \subset \mathcal{X}$ correspond to the real and stylized face domains, respectively. Each $\boldsymbol{X}_i \in \mathbb{R}^{H\times W\times 3}$ represents an image, and each $\boldsymbol{Y}_i = [\boldsymbol{y}_{i,k}] \in \mathbb{R}^{2\times K}$ records K face landmark coordinates, where $\boldsymbol{y}_{i,k} \in \mathbb{R}^2$. We aim to learn a face landmarker, denoted as $f_\theta : \mathcal{X} \mapsto \mathcal{Y}$, where θ is the model parameter. The model should be able to predict face landmarks from facial images and moreover, generalize to both $\mathcal{X}_R$ and $\mathcal{X}_S$. To achieve this aim, we embed the face landmarker into a conditional face warper and learn it associated with a parametric warping field predictor in an alternating optimization framework, as illustrated in Fig. 2.

3.1. Face Landmarking Guided by Face Warping

In this study, we take the SLPT model [52] as the backbone of our face landmarker. Given a stylized face $\boldsymbol{X}_i^{(U)}$, the face landmarker predicts its landmarks as $\widehat{\boldsymbol{Y}}_i^{(U)} = [\hat{\boldsymbol{y}}_{i,k}^{(U)}] = f_\theta(\boldsymbol{X}_i^{(U)})$. At the same time, we can sample a labeled real face $(\boldsymbol{X}_j^{(L)}, \boldsymbol{Y}_j^{(L)}) \sim \mathcal{D}^L$. Treating the labeled and predicted landmarks as keypoints, we can model a warping field from the real face to the stylized one by the following polyharmonic interpolation model [14]:

$$w_{i,\gamma}(\boldsymbol{y}) = \sum\nolimits_{k=1}^{K} \boldsymbol{\omega}_k \phi(\|\boldsymbol{y} - \hat{\boldsymbol{y}}_{i,k}^{(U)}\|_2) + \boldsymbol{V}\boldsymbol{y} + \boldsymbol{b}, \tag{1}$$

where $\gamma = \{\{\boldsymbol{\omega}_k \in \mathbb{R}^2\}_{k=1}^K, \boldsymbol{V} \in \mathbb{R}^{2\times 2}, \boldsymbol{b} \in \mathbb{R}^2\}$ correspond to the parameters of the warping field. As shown in (1), the vector $\boldsymbol{y}$ denotes the u-v coordinate of a pixel in the stylized facial image, and $w_{i,\gamma}(\boldsymbol{y})$ gives the inverse mapping from the pixel $\boldsymbol{y}$ to a coordinate in the real human facial image, conditioned on $\boldsymbol{X}_i^{(U)}$. The first term $\sum_{k=1}^K \omega_k \phi(\|\boldsymbol{y} - \hat{\boldsymbol{y}}_{i,k}^{(U)}\|_2)$ achieves nonparametric regression for modeling nonrigid deformations, in which $\phi(r)$ is a predefined thin-plate spline function. The second term $\boldsymbol{V}\boldsymbol{y} + \boldsymbol{b}$ is a linear parametric model capturing the rigid transformation of $\boldsymbol{y}$.

For each pixel coordinate $\boldsymbol{y} \in \{1, ..., H\} \times \{1, ..., W\}$, we can trace it back to the real human facial image based on $w_{i,\gamma}(\boldsymbol{y})$ and obtain the pixel color as $\boldsymbol{X}_j^{(L)}(w_{i,\gamma}(\boldsymbol{y}))$. Accordingly, with the grid sampler constructed via inverse mapping function $w_{i,\gamma}$, we obtain the warped real human facial image conditioned on $\boldsymbol{X}_i^{(U)}$, denoted as $\widehat{\boldsymbol{X}}_{j|i}^{(L)}$. For $\boldsymbol{y} \in \{1, ..., H\} \times \{1, ..., W\}$ and $j = 1, ..., N_L$, we have

$$\widehat{\boldsymbol{X}}_{j|i}^{(L)}(\boldsymbol{y}) = \boldsymbol{X}_j^{(L)}(w_{i,\gamma}(\boldsymbol{y})). \tag{2}$$

Unlike WarpGAN [40], which generates the warped face by predicting dense keypoints and their displacements by two fully-connected layers during training, we directly use the predicted and observed landmarks to define sparse displacements and estimate other pixels' displacements by spline-based interpolation, which improves computational efficiency significantly. Moreover, by applying the warping field model with limited degree-of-freedom (i.e., few learnable parameters), we can focus more on the learning of the face landmarker in the training phase.

Specifically, the warped face together with the warping field provides a useful guidance for the learning of the face landmarker. In particular, we formulate the learning problem of the face landmarker as follows:

$$\begin{aligned}\min_{\theta,\gamma}\; & \underbrace{\sum\nolimits_{j=1}^{N_L} \|f_\theta(\boldsymbol{X}_j^{(L)}) - \boldsymbol{Y}_j^{(L)}\|_F^2}_{\text{Landmarking error in the source domain}} \\ & + \underbrace{\sum\nolimits_{i=1}^{N_U} \sum\nolimits_{j=1}^{N_L} \|\nabla \widehat{\boldsymbol{X}}_{j|i}^{(L)} - \nabla \boldsymbol{X}_i^{(U)}\|_F^2}_{\text{Discrepancy of image gradient}} \\ & + \underbrace{\sum\nolimits_{i=1}^{N_U} \sum\nolimits_{j=1}^{N_L} \|w_{i,\gamma}(\widehat{\boldsymbol{Y}}_i^{(U)}) - \boldsymbol{Y}_j^{(L)}\|_F^2}_{\text{Landmark warping error}},\end{aligned} \tag{3}$$

where $\|\cdot\|_F$ represents the Frobenius norm of matrix. In (3), the first term is the landmarking error for real faces, which corresponds to the data fidelity loss in the source domain. The second term measures the discrepancy between the stylized face and the warped real face in the gradient field, in which the gradient operation ∇ is implemented by the Sobel operator. The third term is the landmark warping error. Both the second and third terms are determined jointly by the landmarker f_θ and warping field model $w_{i,\gamma}$.

3.2. Alternating Optimization Strategy

The optimization problem in (3) is non-convex because the landmarker is implemented by a neural network and is coupled to the warping field model. As a result, learning θ and

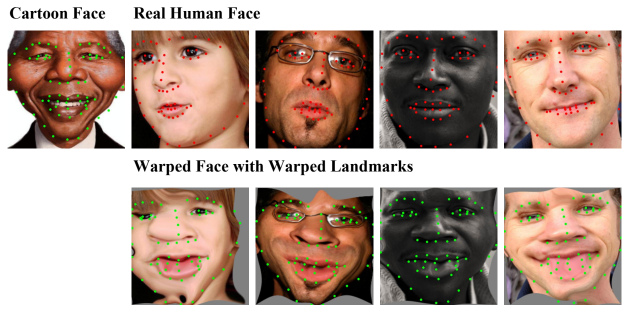


Figure 3. Illustrations of conditional face warping results. Taking a cartoon face as the target, our model warps real human faces accordingly. The red dots indicate real human face landmarks, and green dots indicate cartoon and warped face landmarks.

γ jointly often falls into an undesired local optimum even an unstable saddle point. To mitigate this issue, we propose an alternating optimization framework. In principle, we can decompose the optimization problem in (3) into the following two subproblems and solve them iteratively.

- **Face Warper Optimization:** The first subproblem corresponds to the optimization of the face warper, i.e.,

$$\theta^{(1)},\gamma^{(1)} = \arg\,\min_{\theta,\gamma}\sum\nolimits_{i,j}\|\nabla\widehat{\boldsymbol{X}}_{j|i}^{(L)}-\nabla\boldsymbol{X}_i^{(U)}\|_F^2 + \sum\nolimits_{i,j}\|w_{i,\gamma}(\widehat{\boldsymbol{Y}}_i^{(U)})-\boldsymbol{Y}_j^{(L)}\|_F^2. \tag{4}$$

  In this subproblem, we only care about whether the real human faces can be warped as the stylized faces with high accuracy, so the term of landmarking error is ignored. We solve this problem by Adam [21]: in each step, we update θ and γ based on a batch of randomly-sampled face pairs.

- **Proximal Face Landmarker Optimization:** Given $\theta^{(1)}$ and the predicted landmarks (i.e., $\widehat{\boldsymbol{Y}}_i^{(U)} = f_{\theta^{(1)}}(\boldsymbol{X}_i^{(U)})$ for $i=1,...,N_L$), we can treat $\theta^{(1)}$ as the initial variable and optimize it with a proximal regularizer:

$$\theta^{(2)} = \arg\,\min_{\theta}\sum\nolimits_{j=1}^{N_L}\|f_\theta(\boldsymbol{X}_j^{(L)})-\boldsymbol{Y}_j^{(L)}\|_F^2 + \underbrace{\sum\nolimits_{i=1}^{N_U}\|f_\theta(\boldsymbol{X}_i^{(U)})-\widehat{\boldsymbol{Y}}_i^{(U)}\|_F^2}_{\text{Pseudo landmarking error in the target domain}}. \tag{5}$$

  Here, the second term in (5) measures the estimation errors of the pseudo landmarks achieved in the previous step. Essentially, it works as a proximal regularizer, ensuring that the optimized landmarks $f_{\theta^{(2)}}(\boldsymbol{X}_i^{(U)})$ is not too far away from the previous estimation $\widehat{\boldsymbol{Y}}_i^{(U)}$. Similarly, we can solve this problem by Adam [21] as well.

Fig. 3 shows the warping effect on real human faces achieved by solving (4). In Fig. 3, the first row shows the real human faces with landmarks and the target stylized face, and the second row shows the warping results, in which the green dots are predicted landmarks. These results empirically demonstrate the rationality of our alternating optimization framework. In particular, we can find that solving (4) leads to reasonable warping results, which are similar to the target stylized face on shape. The similarity on face shape indicates that the predicted landmarks can be treated as reliable pseudo labels of the stylized face, which can be used to construct the proximal regularizer that penalizing the pesudo landmarking errors in the target domain. Repeating the above two steps till converge, we obtain the target face landmarker that is generalizable for both real and stylized faces. Algorithm 1 shows the learning scheme.

Algorithm 1 Proposed learning scheme of face landmarker

Require: Labeled real faces $\mathcal{D}^{(L)}$ and unlabeled stylized faces $\mathcal{D}^{(U)}$. The number of iterations (i.e., M). Epochs for the subproblems (i.e., L_1 and L_2).

```
Initialize {γ^(0)} with a pretrained model on D^(L) and {θ^(0)} randomly.
for m = 0, ..., M − 1 do
    Sample a batch {X_i^(U), X_i^(L), Y_i^(L)}_{i=1}^N.
    Face warper optimization:
    Take θ^(2m), γ^(m) as the initialization, then solve (4) by Adam with L_1 epochs and obtain θ^(2m+1).
    Proximal face landmarker optimization:
    Take θ^(2m+1) as the initialization, then solve (5) by Adam with L_2 epochs and obtain θ^(2m+2).
end for
return Output a generalizable face landmarker f_θ(2M).
```

4. Experiment

We apply our learning method to learn a face landmarker and test it on landmarking faces with various styles. Extensive experiments, including comparisons with baselines and analytic ablation studies, demonstrate the effectiveness of our learning method and the generalizability of the corresponding face landmarker. All the experiments are conducted on a single NVIDIA 3090 GPU. Representative experimental results are shown below. **More experimental results and implementation details are given in the supplementary material.**

4.1. Dataset

In this study, we conduct experiments based on the following three commonly-used face datasets.

- **300W Dataset.** 300W [38] is comprised of five well-known real human face datasets including LFPW [3], AFW [60], HELEN [23], XM2VTS [34], and IBUG [38].
- **CariFace Dataset.** CariFace [4] is created by searching and selecting thousands of various caricatures from different celebrities on the Internet.

Table 1. Data settings for the three learning paradigms.

Learning Paradigm	#Training Images and Label Information						#Testing Images				
	300W		CariFace		ArtiFace		300W			CariFace	ArtiFace
							Common	Challenge	Full		
DA (300W→CariFace)	3,148	Labeled	3,372	Unlabeled	—	—	554	135	689	800	—
DA (300W→ArtiFace)	3,148	Labeled	—	—	128	Unlabeled	554	135	689	—	32
GZSL (Unseen ArtiFace)	3,148	Labeled	3,372	Unlabeled	—	—	554	135	689	800	160
GZSL (Unseen CariFace)	3,148	Labeled	—	—	128	Unlabeled	554	135	689	800	32
Oracle	3,148	Labeled	3,372	Labeled	128	Labeled	554	135	689	800	32

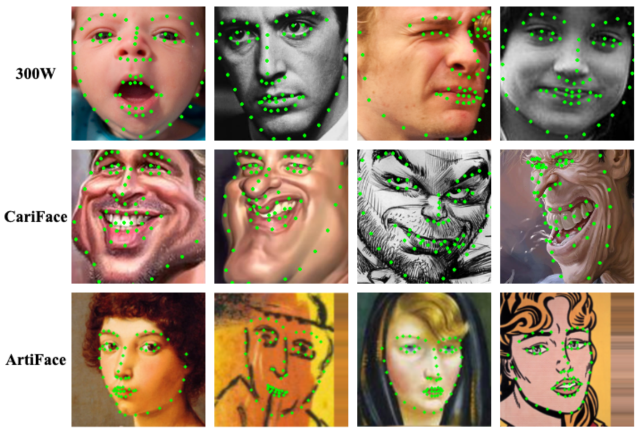


Figure 4. Illustrations of typical samples in the 300W, CariFace, and ArtiFace datasets, each of which is annotated with landmarks.

- **ArtiFace Dataset.** ArtiFace [55] contains 160 artistic portraits of 16 artists, which covers diverse artwork styles ranging from Renaissance to Comics.

Each face in the datasets is annotated with 68 landmarks. Typical faces in the datasets and their landmarks are shown in Fig. 4. We can find that the faces in the three datasets have distinguished styles, which correspond to three different domains. In particular, compared to 300W [38], CariFace [4] exhibits abstract and exaggerated patterns, leading to large representation variations. ArtiFace [55] not only has larger variations across different artistic categories but also differs greatly in terms of the aspect of facial scales, orientations, locations, and so on.

4.2. Learning Paradigms and Baselines

Given the above datasets, we consider the following three learning paradigms:

- **Domain Adaptation (DA).** Given labeled 300W faces and unlabeled stylized faces from CariFace or ArtiFace, we learn a face landmarker based on various domain adaptation methods.
- **Generalized Zero-shot Learning (GZSL).** In the challenging GZSL setting, we learn a face landmarker based on the above DA-based methods and test it in an unseen face domain (e.g., learning the landmarker on labeled 300W and unlabeled CariFace and testing on ArtiFace).
- **Oracle.** In this setting, the labeled faces of all three datasets are accessible, and we can learn the face landmarker by classic supervised learning.

For a fair comparison, in each learning paradigm, we set the architecture of the face landmarker based on the SLPT in [52]. Ideally, we would like to learn landmarkers in the DA and GZSL settings, making its performance comparable to the oracle. In the oracle setting, we can learn the face landmarker directly via classic supervised learning (SL), i.e., $\min_\theta \sum_{(\boldsymbol{X},\boldsymbol{Y})\sim\mathcal{D}} \|f_\theta(\boldsymbol{X}) - \boldsymbol{Y}\|_F^2$. In the DA and GZSL settings, besides minimizing the landmark estimation errors, we can apply various image style transfer and domain adaptation methods, e.g., RevGrad [12], CycleGAN [59], BDL [25], AdaptSegNet [45] and FDA [54], to impose domain adaptation regularization during training. These methods work as the baselines of our method.

Given the landmarkers learned by various methods, we evaluate them with the standard metric, Normalized Mean Error (NME). In Tab. 1, we show the training and testing data settings in the above three learning paradigms. Following existing work [9, 42, 52], we further split the 689 testing faces in 300W into 554 faces in common scenarios and 135 faces in challenging scenarios. The NMEs for the common, challenge, and full scenarios are recorded.

4.3. Numerical and Visual Comparisons

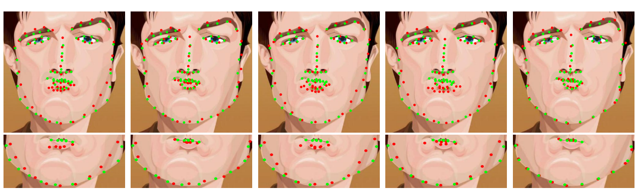
In Tab. 2, we show the performance of various learning methods in different settings, demonstrating the effectiveness and superiority of our method. In particular, existing DA methods often fail to improve the generalization power of model a lot in face landmarking tasks — their performance in target and unseen domains is inferior to that in the oracle setting, with a significant gap on NME. A potential reason for this phenomenon is that these methods focus on the adaptation of image domain and the landmark-related loss is not dominant in their learning processes. As a result, instead of learning the face landmarker, they make more efforts to optimize the parameters of other modules (e.g., the neural network-based face stylization modules and discriminators) during training.

Table 2. Comparisons for various methods on their NMEs. In DA and GZSL settings, the best results are bold.

Learning Paradigm	Learning Method	300W			CariFace	ArtiFace	Average NME
		Common	Challenge	Full			
DA (300W→CariFace) and GZSL (Unseen ArtiFace)	SL+RevGrad [12]	2.84	5.58	3.38	12.19	5.16	7.83
	SL+CycleGAN [59]	**2.74**	5.43	3.27	12.11	4.70	7.70
	SL+BDL [25]	3.28	6.17	3.84	13.63	5.65	8.77
	SL+ASN [45]	2.92	5.26	3.38	12.21	4.75	7.80
	SL+FDA [54]	2.89	5.18	3.34	12.66	4.60	8.07
	Ours	2.79	**4.91**	**3.20**	**7.70**	**3.95**	**5.46**
DA (300W→ArtiFace) and GZSL (Unseen CariFace)	SL+RevGrad [12]	2.99	5.81	3.55	12.46	4.74	8.26
	SL+CycleGAN [59]	3.00	5.65	3.52	12.64	5.34	8.36
	SL+BDL [25]	2.99	5.32	3.44	13.40	5.90	8.73
	SL+ASN [45]	**2.89**	5.81	3.46	16.58	5.65	10.31
	SL+FDA [54]	3.05	6.21	3.68	12.33	5.87	8.27
	Ours	2.90	**5.14**	**3.34**	**10.93**	**3.93**	**7.34**
Oracle	SL	2.68	4.86	3.10	5.48	3.31	4.36

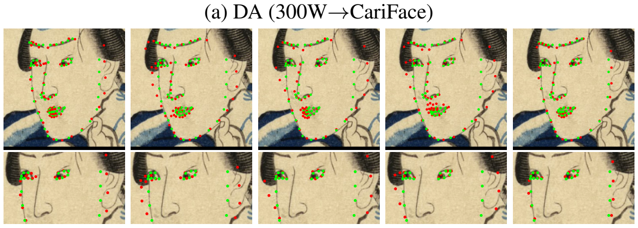
Table 3. The impacts of different model architectures on the NME performance of our method. In each setting, the best results are bold.

Learning Paradigm	Learning Method	300W			CariFace	ArtiFace	Average NME
		Common	Challenge	Full			
DA (300W→CariFace) and GZSL (Unseen ArtiFace)	SBR [9]	3.61	6.36	4.15	8.00	5.09	6.11
	HRNet [47]	2.93	5.29	3.40	7.73	4.33	5.60
	SLPT [52] (Default)	**2.79**	**4.91**	**3.20**	**7.70**	**3.95**	**5.46**
DA (300W→ArtiFace) and GZSL (Unseen CariFace)	SBR [9]	3.70	6.82	4.32	11.35	5.32	8.04
	HRNet [47]	3.04	5.44	3.51	**9.85**	4.30	**6.86**
	SLPT [52] (Default)	**2.90**	**5.14**	**3.34**	10.93	**3.93**	7.34



Source-only RevGrad CycleGAN BDL Ours

(a) DA (300W→CariFace)



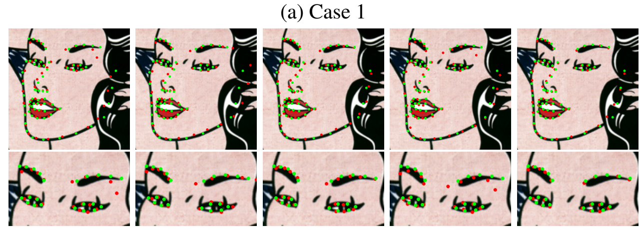
Source-only RevGrad CycleGAN BDL Ours

(b) DA (300W→ArtiFace)

Figure 5. Visual comparisons for various methods in the two DA settings. We only highlight points on the inner lips in the enlarged region of the mouth in (a), as well as the eyes and the sides of the cheeks, excluding points on the eyebrows in (b).

Source-only RevGrad CycleGAN BDL Ours

(a) Case 1

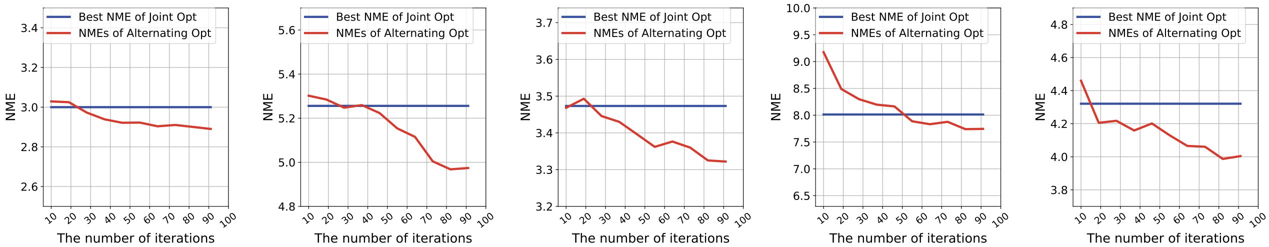


Source-only RevGrad CycleGAN BDL Ours

(b) Case 2

Figure 6. Visual comparisons for various methods in the GZSL (Unseen ArtiFace) setting. The zoomed-in area in (a) highlights the noses and facial contours, while that in (b) concentrates on the upper part of faces.

Different from the baselines, the performance of our model in the target domain (e.g., CariFace and ArtiFace) is improved consistently, while the performance in the source domain does not degrade a lot. In both DA and GZSL settings, our method outperforms the baselines in most situations. Especially in the GZSL settings, the landmarkers



(a) 300W (Common) (b) 300W (Challenge) (c) 300W (Full) (d) CariFace (Target) (e) ArtiFace (Unseen)

Figure 7. Comparisons for different optimization strategies on NME performance in the GZSL (Unseen ArtiFace) setting.

Table 4. Comparison with SOTA landmark detectors on NME.

Learning Method	Train on 300W				Our method	
Backbone	OP	SPIGA	STAR	SLPT	STAR	SLPT
DA on Cariface	10.46	11.23	10.97	11.05	7.62	7.70
GZSL on Artiface	5.55	5.19	5.30	4.56	4.86	3.93

obtained by our method show encouraging generalization power, which achieves lower NME than the baselines in the unseen domain (i.e., ArtiFace). Overall, the NME of our method in the source domain (i.e., 300W) is comparable to that achieved in the oracle setting. For the target even unseen domains, our method reduces the gap to the oracle.

Besides the numerical comparisons, we provide some visualization results obtained by different methods in Figs. 5 and 6. For a complete comparison, the results of the SLPT trained only on 300W are shown in the figures as well, which corresponds to learning a face landmarker only based on the data in the source domain. We can find that by applying our learning method, the face landmarker obtains better landmarks, which aligns with the ground truth with smaller errors, particularly for the landmarks of face contour, nose, mouth, and eyebrow.

To further verify the generalization of our method, we select three more state-of-the-art landmarkers, including OpenPose (OP) [7], SPIGA [36] and STAR [57] for numerical comparisons, and incorporate STAR as the backbone model of our method. Tab. 4 demonstrates that our learning method is applicable for various backbone models (e.g., STAR and SLPT), and integrating existing landmarkers with our learning method is an effective and competitive approach to enhance their generalizability.

4.4. Analytic Experiments

4.4.1 Joint v.s. Alternating Optimization

As aforementioned, learning the face landmarker and the warping field model jointly makes it much easier to fall into undesired local optimum or unstable saddle points. To verify this claim, we apply the joint optimization strategy, i.e., solving (3) by optimizing θ and γ jointly in each gradient descent step and compare its performance with ours. In Fig. 7, we show the best NME achieved by this joint optimization strategy and the NME achieved by our alternating optimization method in each iteration. We can find that the joint optimization strategy tends to overfit the source domain (300W), leading to power generlizability in the target and unseen domains. On the contrary, with the increase of iterations, the NMEs of our method on the target and unseen domains decrease consistently and become lower than those of the joint optimization after several iterations.

4.4.2 Impacts of Model Architectures

Besides the optimization strategy, the model architecture also has an impact on the model performance. By default, we implement the face landmarker as the SLPT model. In this experiment, we further explore the performance of other model architectures, including SBR [9] and HRNet [47]. Tab. 3 shows the performance of different model architectures in DA and GZSL settings. We can find that the SLPT-based face landmarker works better than the two competitors in this experiment. A potential reason for this phenomenon is that both SBR and HRNet are heatmap-based face landmarkers. Given an input facial image, they output a heatmap indicating the distribution of landmarks rather than a set of deterministic landmark coordinates. We have to first detect the landmarks from the heatmap and then pass them through the warping field model. Accordingly, in the face warper optimization step (i.e., solving (4)), the landmarker and the warping field model cannot be trained in an end-to-end way because the backpropagation of the gradient becomes inapplicable. As a result, we have to update θ and γ alternatively, leading to suboptimal performance.

4.4.3 Effects on Loss Term

In the context of our model's loss function, the image gradient field serves as its input. It has been observed through experiments that utilizing distinct gradient operators leads to the generation of varying image gradient fields, which in

Table 5. Quantitative comparison on CariFace dataset.

Operator	Gray	Spatial	Laplacian	Canny	Sobel
NME	✓	8.282	7.870	8.005	**7.695**
	×	7.958	7.921	8.011	7.863

Table 6. Comparisons for different losses on NME.

Type of Loss		MSE	Perceptual	w.o. Grad MSE	Grad MSE
300W	Common	3.08	2.89	2.96	**2.79**
	Challenge	5.43	5.13	5.00	**4.91**
	Full	3.54	3.33	3.36	**3.20**
DA on Caricature		8.26	8.23	9.48	**7.70**
GZSL on ArtiFace		4.12	4.07	4.06	**3.95**

turn influences the model's performance. Consequently, in this section, we assess and compare the training outcomes derived from image gradient fields obtained through the application of diverse gradient operators.

Moreover, it is crucial to recognize that the choice of the input image, specifically whether it is a grayscale image or not, can also exert a certain influence on the results. As a result, we have incorporated a comparison of the outcomes when the input is a grayscale image. It is important to emphasize that all experiment settings are maintained consistently. The results are shown in Tab. 5. Utilizing grayscale images yields superior effects compared to not using grayscale images. This could be attributed to the fact that converting an image to grayscale eliminates the trivial effects of color and complex textures. The most optimal outcome is achieved by employing the Sobel operator. This may be due to the Sobel operator's emphasis on edge information within the image, enabling it to capture more geometric details. Additionally, the Sobel operator incorporates a mild smoothing effect during gradient computation, which mitigates the impact of noise.

Furthermore, we consider i) replacing the gradient MSE loss in (3) with the pixel MSE or perceptual loss and ii) removing the gradient MSE loss. Tab. 6 shows the rationality of the gradient MSE for the reason that i) landmarks are distributed on edges; and ii) the gradient field filters out unnecessary color information, simplifying the task. These findings form the basis for our decision to adopt this particular method in our experimental process.

4.4.4 Rationality of Proposed Warping Field Model

Besides the model architecture of the face landmarker, the warping field model impacts our face landmarker as well — when the predicted warping field is inaccurate, we cannot obtain reliable pseudo landmarks for stylized faces. Therefore, we investigate different face warping models and

(a) Source (b) Target (c) AutoToon (d) CariGANs (e) RAFT (f) Ours

Figure 8. Comparisons for different face warpers.

demonstrate the rationality of the warping field model implemented in our work. In particular, applying different warping field models in our training process, we visualize their warping results in Fig. 8. We can find that although the face landmarker with the polyharmonic interpolation model leads to a very simple face warper, it outperforms many existing neural network-based image warping methods, such as AutoToon [15] and CariGANs [5] for facial manipulation and the optical flow methods like RAFT [44]. These methods either require one-to-one correspondence information between source and target images or assume the deformation between the two images to be slight, making them unsuitable for our problem, especially for the stylized faces with significant nonrigid deformations.

5. Conclusion

In this paper, we propose a simple but effective method for learning a generalizable face landmarker applicable to facial images with different styles. Given labeled real human faces and unlabeled stylized faces, our method learns the face landmarker under the guidance of conditional face warping, demonstrating the usefulness of the warping information. An alternating optimization framework is proposed to learn the face landmarker together with the warping field model. Experiments demonstrate the effectiveness of our method. Especially in the generalized zero-shot learning scenarios, our method achieves encouraging landmarking accuracy in unseen face domains.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (62102031), the foundation of Key Laboratory of Artificial Intelligence, Ministry of Education, P.R. China, and the Young Scholar Program (XSQD-202107001) from Beijing Institute of Technology.

References

[1] Olalekan Agbolade, Azree Nazri, Razali Yaakob, Abdul Azim Abd Ghani, and Yoke Kqueen Cheah. Landmark-based homologous multi-point warping approach to 3d facial recognition using multiple datasets. *PeerJ Computer Science*, 6:e249, 2020. 1

[2] Sadegh Aliakbarian, Pashmina Cameron, Federica Bogo, Andrew Fitzgibbon, and Thomas J. Cashman. Flag: Flow-based 3d avatar generation from sparse observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13253–13262, 2022. 1

[3] Peter N. Belhumeur, David W. Jacobs, David J. Kriegman, and Neeraj Kumar. Localizing parts of faces using a consensus of exemplars. In *CVPR 2011*, 2011. 4

[4] Hongrui Cai, Yudong Guo, Zhuang Peng, and Juyong Zhang. Landmark detection and 3d face reconstruction for caricature using a nonlinear parametric model, 2020. 1, 4, 5

[5] Kaidi Cao, Jing Liao, and Lu Yuan. Carigans: Unpaired photo-to-caricature translation. *arXiv preprint arXiv:1811.00222*, 2018. 1, 8

[6] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. *International journal of computer vision*, 107:177–190, 2014. 2

[7] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 7

[8] Forrester Cole, David Belanger, Dilip Krishnan, Aaron Sarna, Inbar Mosseri, and William T Freeman. Synthesizing normalized faces from facial identity features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3703–3712, 2017. 3

[9] Xuanyi Dong, Shoou-I Yu, Xinshuo Weng, Shih-En Wei, Yi Yang, and Yaser Sheikh. Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 360–368, 2018. 2, 5, 6, 7

[10] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015. 3

[11] Pengfei Dou, Shishir K Shah, and Ioannis A Kakadiaris. End-to-end 3d face reconstruction with deep neural networks. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5908–5917, 2017. 1

[12] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 1, 2, 5, 6

[13] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015. 3

[14] C. A. Glasbey and K. V. Mardia. A review of image-warping methods. *Journal of Applied Statistics*, 25(2):155–171, 2002. 3

[15] Julia Gong, Yannick Hold-Geoffroy, and Jingwan Lu. Autotoon: Automatic geometric warping for face cartoon generation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 360–369, 2020. 1, 3, 8

[16] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19, 2006. 2

[17] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4893–4902, 2019. 2

[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3

[19] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014. 1

[20] SunnieS. Kim, NicholasI. Kolkin, Jason Salavon, and Gregory Shakhnarovich. Deformable style transfer, 2020. 2

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[22] Marek Kowalski, Jacek Naruniec, and Tomasz Trzcinski. Deep alignment network: A convolutional neural network for robust face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017. 2

[23] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S. Huang. *Interactive Facial Feature Localization*, page 679–692. 2012. 4

[24] Hui Li, Zidong Guo, Seon-Min Rhee, Seungju Han, and Jae-Joon Han. Towards accurate facial landmark detection via cascaded transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4176–4185, 2022. 2

[25] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6936–6945, 2019. 1, 5, 6

[26] Feng Liu, Dan Zeng, Qijun Zhao, and Xiaoming Liu. Joint face alignment and 3d face reconstruction. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 545–560. Springer, 2016. 1

[27] Shaolei Liu, Xiaoyuan Luo, Kexue Fu, Manning Wang, and Zhijian Song. A learnable self-supervised task for unsupervised domain adaptation on point cloud classification and segmentation. *Frontiers of Computer Science*, 17(6):176708, 2023. 1, 2

[28] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017. 1
[29] Xiao-Chang Liu, Yong-Liang Yang, and Peter Hall. Learning to warp for style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3702–3711, 2021. 3
[30] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015. 2
[31] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. *Advances in neural information processing systems*, 29, 2016. 2
[32] Jiangjing Lv, Xiaohu Shao, Junliang Xing, Cheng Cheng, and Xi Zhou. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3317–3326, 2017. 2
[33] Iacopo Masi, Stephen Rawls, Gérard Medioni, and Prem Natarajan. Pose-aware face recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4838–4846, 2016. 1
[34] Kieron Messer, Jiri Matas, Josef Kittler, Juergen Luettin, Gilbert Maitre, et al. Xm2vtsdb: The extended m2vts database. In *Second international conference on audio and video-based biometric person authentication*, pages 965–966. Citeseer, 1999. 4
[35] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Multi-adversarial domain adaptation. In *Thirty-second AAAI conference on artificial intelligence*, 2018. 2
[36] Andrés Prados-Torreblanca, José M Buenaposada, and Luis Baumela. Shape preserving facial landmarks with graph attention networks. *arXiv preprint arXiv:2210.07233*, 2022. 7
[37] Joseph Roth, Yiying Tong, and Xiaoming Liu. Unconstrained 3d face reconstruction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2606–2615, 2015. 1
[38] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *2013 IEEE International Conference on Computer Vision Workshops*, 2014. 1, 4, 5
[39] Ahmed Selim, Mohamed Elgharib, and Linda Doyle. Painting style transfer for head portraits using convolutional neural networks. *ACM Transactions on Graphics (ToG)*, 35(4): 1–18, 2016. 3
[40] Yichun Shi, Debayan Deb, and Anil K Jain. Warpgan: Automatic caricature generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10762–10771, 2019. 3
[41] Baochen Sun, Jiashi Feng, and Kate Saenko. Correlation alignment for unsupervised domain adaptation. *Domain adaptation in computer vision applications*, pages 153–171, 2017. 2
[42] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5
[43] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014. 1
[44] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 3, 8
[45] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7472–7481, 2018. 5, 6
[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2
[47] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 2, 6, 7
[48] Xinyao Wang, Liefeng Bo, and Li Fuxin. Adaptive wing loss for robust face alignment via heatmap regression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2
[49] Kunhong Wu, Fan Jia, and Yahong Han. Domain-specific feature elimination: multi-source domain adaptation for image classification. *Frontiers of Computer Science*, 17(4): 174705, 2023. 2
[50] Ruizheng Wu, Xiaodong Gu, Xin Tao, Xiaoyong Shen, Yu-Wing Tai, et al. Landmark assisted cyclegan for cartoon face generation. *arXiv preprint arXiv:1907.01424*, 2019. 3
[51] Wayne Wu, Chen Qian, Shuo Yang, Quan Wang, Yici Cai, and Qiang Zhou. Look at boundary: A boundary-aware face alignment algorithm. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2129–2138, 2018. 1
[52] Jiahao Xia, Weiwei Qu, Wenjian Huang, Jianguo Zhang, Xi Wang, and Min Xu. Sparse local patch transformer for robust face alignment and landmarks inherent relation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4052–4061, 2022. 1, 3, 5, 6
[53] Jiaolong Yang, Peiran Ren, Dongqing Zhang, Dong Chen, Fang Wen, Hongdong Li, and Gang Hua. Neural aggregation network for video face recognition. In *Proceedings of*

the IEEE conference on computer vision and pattern recognition, pages 4362–4371, 2017. 1

[54] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4085–4095, 2020. 5, 6

[55] Jordan Yaniv, Yael Newman, and Ariel Shamir. The face of art: landmark detection and geometric style in portraits. *ACM Transactions on graphics (TOG)*, 38(4):1–15, 2019. 1, 2, 5

[56] Jing Zhang, Zewei Ding, Wanqing Li, and Philip Ogunbona. Importance weighted adversarial nets for partial domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8156–8164, 2018. 2

[57] Zhenglin Zhou, Huaxia Li, Hong Liu, Nanyang Wang, Gang Yu, and Rongrong Ji. Star loss: Reducing semantic ambiguity in facial landmark detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15475–15484, 2023. 7

[58] Zhou, Erjin and Fan, Haoqiang and Cao, Zhimin and Jiang, Yuning and Yin, Qi. Extensive Facial Landmark Localization with Coarse-to-Fine Convolutional Network Cascade. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, 2013. 1, 2

[59] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 2, 5, 6

[60] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2879–2886. IEEE, 2012. 4

[61] Yi Zhu, Xindong Wu, Jipeng Qiang, Yunhao Yuan, and Yun Li. Representation learning via an integrated autoencoder for unsupervised domain adaptation. *Frontiers of Computer Science*, 17(5):175334, 2023. 1