

PriSM: Parsing and Style-Mixed Consistency for Unsupervised Domain Adaptation in Facial Landmark Detection

Chieh-Yu Yang¹, Hou-Ning Hu², Sykai Chen², Yu-Lun Liu¹, and Yen-Yu Lin¹

¹ National Yang Ming Chiao Tung University, Hsinchu, Taiwan
jerrychild999922@gmail.com, {yulunliu, lin}@cs.nycu.edu.tw

² MediaTek, Hsinchu, Taiwan
{Hou-ning.Hu, Sy-kai.Chen}@mediatek.com

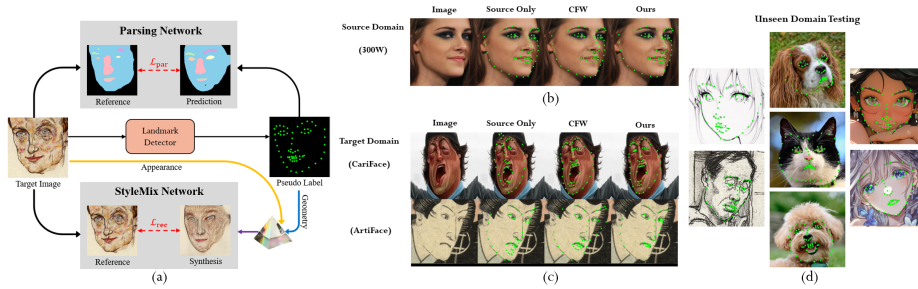


Fig. 1: (a) Overview of our PriSM framework. (b)-(c) Qualitative comparisons on the source (300W [42]) and target (CariFace [2] and ArtiFace [57]) domains. Our method preserves performance on the source domain while outperforming the SoTA method CFW [29] on the target domains. (d) Generalization capability of our method on unseen domains, including sketches, anime, and even animal faces.

Abstract. Facial landmark detectors trained on real human faces often fail to generalize effectively to stylized domains such as caricatures and artistic portraits, necessitating unsupervised domain adaptation (UDA). Although self-training is a widely used UDA strategy to bridge domain gaps, it frequently breaks down under large domain shifts as it is prone to amplifying confident yet erroneous pseudo-label predictions. To this end, we propose PriSM (**P**arsing and **S**tyl**e-M**ixed Consistency), a novel method for robust landmark pseudo-label validation. PriSM leverages two complementary signals: a Parsing Network enforces high-level structural alignment through face parsing consistency, while a StyleMix Network enforces fine-grained geometric constraints by reducing the reconstruction error between the input face and another face synthesized from the pseudo-label’s structure and the input face’s appearance. Extensive experiments on the challenging CariFace and ArtiFace bench-

marks under the UDA setting demonstrate that PriSM significantly outperforms existing state-of-the-art methods and exhibits strong generalizability to unseen domains. The source code is available at: <https://github.com/jerrychild999922/PriSM-UDA>.

Keywords: Facial Landmark Detection · Unsupervised Domain Adaptation · Face Parsing · Face Reconstruction

1 Introduction

Facial landmark detection aims to localize key facial points, such as the eyes, nose, mouth, and facial contour, in an image. It is essential to many computer vision and graphics applications, like 3D face reconstruction [30, 51], facial expression analysis [17, 20, 35], face recognition [1, 43], and AR/VR rendering [37, 41]. Modern facial landmark detectors [12, 23, 53, 62] have achieved strong performance on real human faces by utilizing supervised training on large-scale human-face datasets.

Recently, the rapid growth of AI-generated characters [39], stylized avatars [48] in social-media filters, and related applications has created a surge of demand for reliable landmark detection on non-photorealistic faces, such as caricatures [2], cartoons [14], anime [33], and artistic portraits [57]. However, as shown in Figure 1(c), detectors trained only on human faces in the source domain often fail on these stylized, generated faces in the target domain due to the severe domain gap in both appearance (e.g., texture, shading, and color palette) and geometry (e.g., exaggerated facial proportions). Furthermore, training a supervised detector on the target stylized domains by manual annotation is becoming infeasible due to the abundance and variety of stylized faces synthesized by modern generative models.

To address this issue, Unsupervised Domain Adaptation (UDA) offers a solution by adapting knowledge from a labeled source domain to an unlabeled target domain, thereby eliminating the need for manual annotation. Among existing UDA strategies, self-training allows a model to generate and trust highly confident “pseudo-labels” for target-domain data, and has recently shown great success across many tasks [4, 26, 54, 65, 66]. However, this intuitive strategy suffers from a key limitation, especially when a large domain gap is present, such as in facial landmark detection: A model trained on real faces often makes highly confident but inaccurate initial predictions on stylized faces. This yields a dangerous cycle where the model learns from and repeatedly reinforces its confident mistakes, leading to the collapse of self-training. Therefore, a more robust mechanism for validating pseudo-labels beyond simple confidence scores is essential for UDA in stylized domains.

To this end, we propose PriSM (Parsing and Style-Mixed Consistency), a novel dual-consistency framework for validating facial landmark pseudo-labels. PriSM adapts a landmark detector through two complementary auxiliary tasks: The first task, **face parsing**, is more tolerant of domain shifts and can enforce high-level structural alignment. The second task, **face reconstruction**, utilizes

the synthesis capabilities of a generative model as a validation tool to ensure fine-grained geometric correctness.

Specifically, face parsing captures coarser geometric layouts. Thus, it is relevant to landmark detection but is more robust to domain shifts. As shown in Figure 1(a), our Parsing Network leverages this stability by enforcing the parsing map generated from a pseudo-label to be consistent with the reference map produced by applying a pre-trained face parser to the input face. This process effectively enables high-level structural alignment for the pseudo-labels.

To complement the Parsing Network’s coarse guidance, our StyleMix Network in Figure 1(a) enforces fine-grained geometric constraints using an analysis-by-synthesis loop. Based on the principle that a face can be decomposed into its geometric structure and appearance and then reconstructed, this network uses the input face’s appearance and the pseudo-label’s geometry to reconstruct the input via a generative model. This way, the reconstruction quality is directly dependent on the geometric accuracy of the pseudo-label. The StyleMix Network hence provides a supervisory signal for refining intricate facial features. By integrating these two complementary networks, our method effectively corrects erroneous pseudo-labels and robustly adapts to target domains with significant domain shifts.

Extensive experiments conducted on two challenging stylized face datasets, CariFace [2] and ArtiFace [57], demonstrate that PriSM performs favorably against existing state-of-the-art methods under various UDA settings, as shown in Figure 1(c). While adapting to the target domain, our method effectively avoids catastrophic forgetting by maintaining superior performance on the source-domain 300W dataset [42], given in Figure 1(b). Furthermore, we notice PriSM’s strong generalizability to unseen domains. The results on a diverse range of unseen domains in Figure 1(d), such as sketches, anime, and even animal faces, showcase the remarkable robustness of our method.

The contributions of this paper are summarized below:

- We present PriSM, a novel dual-consistency framework for unsupervised facial landmark detector adaptation. It pioneers a new paradigm for robust pseudo-label validation by jointly enforcing high-level structural alignment and fine-grained face reconstruction.
- We propose a novel use of a fixed StyleGAN generator within an analysis-by-synthesis loop, which serves as a powerful validation tool, providing a clean, pixel-level reconstruction constraint on the geometric structure of pseudo-labels.
- Our proposed method achieves state-of-the-art performance for unsupervised domain adaptation on challenging stylized face benchmarks, CariFace and ArtiFace. It also demonstrates strong generalizability to various unseen domains without any fine-tuning, showcasing its remarkable robustness.

2 Related Work

2.1 Facial Landmark Detection

Facial landmark detection aims to localize a set of predefined keypoints on a face, primarily using one of two major formulations: **heatmap regression** and **coordinate regression**. In heatmap regression, each landmark is represented as a spatial probability map, where the peak corresponds to the predicted location [36, 47]. Methods based on heatmap regression have improved performance by introducing advanced architectures [21], novel loss functions [12, 49, 62], and geometry-aware representations [19] to handle real-world complexities.

Early methods for coordinate regression directly estimate landmark coordinates from CNN features using fully connected layers [7, 52, 60]. To better capture the inherent facial structure, subsequent research introduces explicit structural constraints using techniques like Graph Convolutional Networks (GCNs) [25, 63]. Recently, Transformer-based architectures [3, 23, 24] leverage self-attention to learn adaptive structural relationships between landmarks. Despite recent advances, supervised methods remain reliant on large-scale annotations and fail to generalize effectively under domain shifts, which motivates our exploration into unsupervised domain adaptation.

2.2 Unsupervised Domain Adaptation

Unsupervised Domain Adaptation (UDA) transfers knowledge from a labeled source domain to an unlabeled target domain to bridge the domain gap. Existing methods generally fall into three categories: **adversarial-based**, **image-translation-based**, and **self-training-based** methods.

Adversarial-based methods [8, 11, 38, 45, 46] train a feature extractor to learn domain-invariant features by competing against a domain discriminator. However, strict domain invariance may lose domain-specific cues, such as contour lines in cartoon faces, that are essential for accurate localization. Image-translation-based methods [9, 28, 31, 56, 64] generate paired training data by converting source images to the target style for supervised training. They can bridge the appearance gap, but inherently fail to capture critical geometric variations in facial landmark detection.

Self-training-based methods adapt models by generating pseudo-labels for unlabeled data and retraining on predictions above a confidence threshold. It is effective across a range of UDA tasks [4, 26, 54, 65, 66]. However, self-training is highly dependent on pseudo-label quality, as overconfident predictions can amplify early errors and degrade adaptation. While prior work attempts to refine pseudo-label generation or confidence estimation [13, 50, 59, 61], a promising research trend is to validate the primary task’s predictions using an auxiliary self-supervised task [6, 10]. Inspired by this, we introduce a dual-consistency validation framework. It jointly enforces structural alignment and pixel-level reconstruction and can provide a more stable and reliable supervision signal for evaluating pseudo-label quality under large domain shifts.

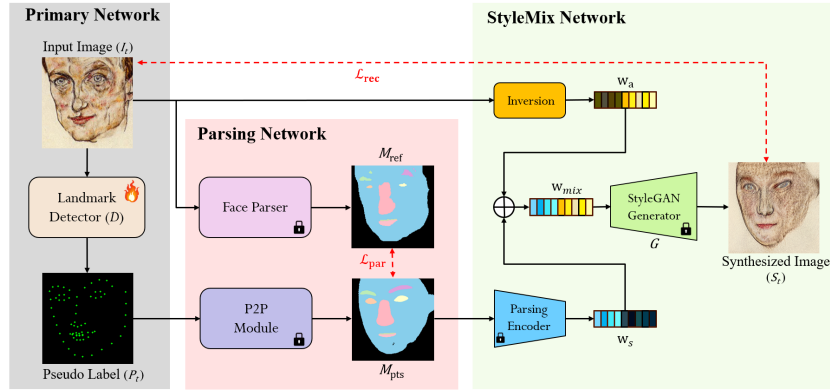


Fig. 2: Overview of our proposed PriSM framework. Our method is composed of three major components: **(Left)** The Primary Network employs a facial landmark detector to generate the pseudo-label for an input target-domain image. **(Middle)** The Parsing Network provides high-level structural validation by minimizing the parsing loss $\mathcal{L}_{\text{par}}$ between a reference parsing map M_{ref} from a pre-trained face parser and a predicted parsing map M_{pts} computed from the pseudo-label. **(Right)** The StyleMix Network synthesizes a face image by mixing the appearance latent code w_a , obtained via inversion of the input image, with the structure latent code w_s inferred from M_{pts} via a parsing encoder. The reconstruction loss $\mathcal{L}_{\text{rec}}$ between the synthesized and input faces serves as a supervisory signal to correct geometric inaccuracies in the pseudo-label.

2.3 Controllable Image Synthesis

Our PriSM framework formulates face reconstruction as a validation mechanism. It requires a synthesis process conditioned on both the geometric structure derived from pseudo-labels of facial landmarks and the appearance from the input face. Early geometric approaches like face warping [18, 29, 57] offer low-level structural control, but they are limited to simple transformations and lack a semantic understanding of facial structures. While recent diffusion-based models like ControlNet [27, 58] provide explicit structural control, we instead opt for a StyleGAN-based [15, 16] generator. We favor this choice specifically for its style-mixing capability, which allows for disentangled control over geometric structure (represented by facial landmarks in this work) and appearance attributes. This property enables our analysis-by-synthesis loop: We explicitly integrate the estimated landmarks with the input’s preserved appearance to reconstruct the face. It follows that the quality of the estimated landmarks can be evaluated based the resulting reconstruction fidelity.

3 Proposed Method

This section presents PriSM, a novel method for unsupervised domain adaptation in facial landmark detection. As shown in Figure 2, our method operates on a

self-training loop where the quality of pseudo-label is validated through a dual-consistency network. Initially, our Primary Landmark Detector generates the facial landmark pseudo-label for a target-domain face, detailed in Section 3.1. The pseudo-label is then fed into our Dual-Consistency Network, which uses two auxiliary tasks to produce robust supervisory signals that measure the geometric correctness of the pseudo-label, as described in Section 3.2. Finally, the training and optimization stage, developed in Section 3.3, combines these proposed losses to iteratively refine the detector, thereby enhancing its accuracy on the target domain.

3.1 Primary Landmark Detector

The core of our framework is a facial landmark detector to be adapted. The detector can be an arbitrary end-to-end differentiable detector. In this work, we adopt the SLPT architecture [53] for its great stability and performance, which is pre-trained on the labeled source domain. Our central goal is to adapt this detector to the unlabeled target domain. Therefore, it is designed as the sole learnable component throughout the entire adaptation phase.

As illustrated in Figure 2, the landmark detector D takes an input face image $I \in \mathbb{R}^{H \times W \times 3}$ and directly outputs the 2D coordinates for a set of $K = 68$ facial landmarks, represented as a point set $P = \{\mathbf{p}_k \in \mathbb{R}^2\}_{k=1}^K$. In our UDA pipeline, these landmark predictions serve as the pseudo-label P_t for each image I_t from the unlabeled target domain. These pseudo-labels are subsequently validated by our dual-consistency network to guide the adaptation process of the facial landmark detector.

3.2 Dual-Consistency Network

A key limitation of conventional self-training methods is their reliance on the detector’s internal confidence scores for pseudo-label quality assessment. Inspired by prior work [6, 10] leveraging external signals to objectively assess pseudo-label quality, we adopt a similar principle by formulating the validation signals as two auxiliary losses. These losses act as a reliable proxy for the ground-truth error: A low loss value signifies a geometrically accurate pseudo-label, while a high value indicates a significant error. These signals are generated by two networks that enforce consistency at two complementary levels: 1) A high-level structural alignment provided by our Parsing Network, and 2) a fine-grained pixel-wise constraint from the proposed StyleMix Network. The two networks are respectively elaborated in the following.

Parsing Network. As illustrated in Figure 2, our Parsing Network validates the structural integrity of the pseudo-label P_t for each target-domain face I_t by comparing two generated parsing maps. The first parsing map, $M_{\text{pts}} \in \mathbb{R}^{H' \times W' \times C}$ where $H' = W' = 256$ and C is the number of facial semantic classes, is produced by the Point-to-Parse (P2P) module, which takes the pseudo-label P_t as input. While this P2P module is pre-trained on a source domain dataset LaPa [32]

and employs a simple MLP-based encoder and a CNN-based decoder to produce the segmentation map, its key characteristic is its domain-agnostic nature. This robustness stems from the fact that the P2P module operates on the geometric coordinates of the landmarks. Its predictions are therefore based on the inherent spatial and structural relationships between facial keypoints, which remain largely consistent across different domains. The second parsing map M_{ref} serves as a reference and is produced by applying the face parser [34] pre-trained on the source domain [22] to the input image I_t .

To ensure a focused comparison, we constrain the validation on $C = 10$ core semantic classes (e.g., face skin, eyes, eyebrows) that can be reliably inferred from the $K = 68$ landmarks. We enforce structural consistency by minimizing the pixel-wise Cross-Entropy loss between the predicted parsing map and the reference parsing map. The resultant parsing loss, $\mathcal{L}_{\text{par}}$, is formulated as:

$$\mathcal{L}_{\text{par}} = \text{CrossEntropy}(M_{\text{pts}}, M_{\text{ref}}). \quad (1)$$

Our design is predicated on the insight that capturing broad semantic areas is inherently more resilient to domain gaps than precise landmark regression. By employing face parsing strictly as a relaxed constraint, minimizing the parsing loss $\mathcal{L}_{\text{par}}$ in Eq. 1 robustly facilitates the adaptation of the facial landmark detector. However, unlike our P2P module, the face parser is not fully domain-agnostic. Since it is trained on the source domain and operates on raw image pixels, it still suffers from the domain gap. Its output M_{ref} may degrade for fine-grained features in the target domain (see Figure 2). Consequently, the Parsing Network serves only as a coarse structural guide, relying on precise geometric refinement to our StyleMix Network. Furthermore, in cases of severe domain shifts, the parser may fail and produce a nearly empty reference map. To prevent learning from such erroneous signals, we introduce an filtering mechanism that dynamically discards the parsing supervision if the parsed face area is abnormally small (*i.e.*, $\text{Area}(M_{\text{ref}}) < \tau \times \text{Area}(M_{\text{pts}})$, where τ is a predefined threshold).

StyleMix Network. While the Parsing Network provides coarse structural guidance, it lacks the precision to correct fine-grained geometric errors. To address this, we introduce the StyleMix Network, which imposes a robust, pixel-wise constraint on the pseudo-label through an analysis-by-synthesis loop. The core intuition is straightforward: if a pseudo-label is geometrically accurate, a face image synthesized by combining the geometric structure from the pseudo-label with the appearance from the original input image should be nearly identical to the input itself. Any resulting discrepancy in the reconstruction would thus primarily result from geometric inaccuracies in the pseudo-label. Crucially, since our GAN inversion adapts to out-of-distribution textures, the reconstruction error specifically penalizes these geometric misalignments regardless of the target domain’s artistic style. This property yields valid gradients that provide stable guidance for the detector.

As detailed in Figure 2, the StyleMix Network begins by disentangling the geometric structure and the appearance attributes of the input face I_t into two

latent codes within StyleGAN’s w^+ space [16]. First, an appearance latent code $w_a \in \mathbb{R}^{18 \times 512}$ is obtained for the face using the GAN inversion technique [55]. Simultaneously, a structure latent code $w_s \in \mathbb{R}^{18 \times 512}$, is generated by feeding the parsing map M_{pts} into a pre-trained Parsing Encoder [55].

With both latent codes acquired, we perform style-mixing [16] to create a composite latent code $w_{\text{mix}} \in \mathbb{R}^{18 \times 512}$, which is yielded by combining the coarse-level styles (*i.e.* layers 0 $\sim$ 7) from the structure code w_s , which governs pose and shape, with the fine-level styles (*i.e.* layers 8 $\sim$ 17) from the appearance code w_a , which controls color and texture. The mixed code w_{mix} is fed into a fixed StyleGAN generator G to synthesize the image S_t . We enforce pixel-wise consistency by minimizing the perceptual LPIPS loss between the synthesized image S_t and the original input I_t . The reconstruction loss $\mathcal{L}_{\text{rec}}$ is formulated as:

$$\mathcal{L}_{\text{rec}} = \text{LPIPS}(S_t, I_t). \quad (2)$$

Note that the entire StyleMix Network is frozen during the adaptation process, ensuring that the gradient from $\mathcal{L}_{\text{rec}}$ flows back through the synthesis loop to update only the parameters of the facial landmark detector. This provides a clean, self-supervised signal that directly penalizes fine-grained geometric errors of the pseudo-label.

3.3 Training and Optimization

The overall training process is designed to progressively adapt the primary landmark detector D to the unlabeled target domain. This is achieved by fine-tuning the detector with our dual-consistency losses and by integrating this process into a self-training framework. The total objective for adapting the detector on the target domain, which we refer to as the adaptation loss $\mathcal{L}_{\text{adapt}}$, is a weighted sum of the parsing loss $\mathcal{L}_{\text{par}}$ in Eq. 1 and the reconstruction loss $\mathcal{L}_{\text{rec}}$ in Eq. 2, *i.e.*,

$$\mathcal{L}_{\text{adapt}} = \lambda_{\text{par}} \mathcal{L}_{\text{par}} + \lambda_{\text{rec}} \mathcal{L}_{\text{rec}}, \quad (3)$$

where λ_{par} and λ_{rec} are the balancing hyperparameters. Inspired by prior work [29], we employ an alternating optimization strategy that consists of two iterative steps: 1) A dual-consistency guided adaptation step and 2) a supervised fine-tuning step with refined pseudo-labels, detailed below.

Dual-Consistency Guided Adaptation. In the first step of each iteration, we aim at refining the parameters θ of the detector D using our proposed validation mechanism. We optimize these parameters by minimizing the adaptation loss $\mathcal{L}_{\text{adapt}}$ on the unlabeled target dataset, $\mathcal{X}_T$. This produces an intermediate, adapted set of parameters θ' via

$$\theta' = \arg \min_{\theta} \mathbb{E}_{I_t \sim \mathcal{X}_T} [\mathcal{L}_{\text{adapt}}(D_{\theta}(I_t), I_t)]. \quad (4)$$

This step leverages our consistency checks to enhance the detector’s geometric accuracy in the target domain without requiring ground-truth labels.

Supervised Fine-tuning with Refined Pseudo-Labels. In the second step, we use the improved detector, now parameterized by θ' , to generate a new, higher-quality set of pseudo-labels $\{P'_t = D_{\theta'}(I_t)\}$ for the entire target dataset $\mathcal{X}_T$. These refined pseudo-labels, along with the labeled source-domain dataset $\mathcal{X}_S = \{(I_s, Y_s)\}$, form a combined training set. Starting from the adapted weights θ' as a warm start, we perform supervised fine-tuning on this combined set. This step involves the landmark detector’s loss $\mathcal{L}_{\text{det}}$ and yields the updated parameters θ via

$$\theta = \arg \min_{\theta} \left(\mathbb{E}_{(I_s, Y_s) \sim \mathcal{X}_S} [\mathcal{L}_{\text{det}}(D_{\theta}(I_s), Y_s)] + \mathbb{E}_{(I_t, P'_t) \sim \mathcal{X}_T} [\mathcal{L}_{\text{det}}(D_{\theta}(I_t), P'_t)] \right). \quad (5)$$

This fine-tuning step serves a crucial dual purpose: it consolidates the knowledge gained on the target domain by treating the refined pseudo-labels as ground truth, while the inclusion of source data ensures that the model’s original knowledge is preserved, effectively preventing catastrophic forgetting. These two steps are repeated until convergence, resulting in a detector robustly adapted to the target domain.

4 Experiments

This section evaluates the proposed method by detailing the experimental settings and presenting the results of both comparative and ablation studies.

4.1 Experimental Setup

Implementation Details. We adopt the SLPT [53] architecture as our primary landmark detector, initialized with weights pre-trained on the 300W dataset [42]. All input images are resized to 256×256 . During training, all auxiliary network components are frozen. The reference parsing map in the Parsing Network is generated by an EHANet [34] pre-trained on CelebAMask-HQ [22]. The StyleMix Network employs a fixed StyleGAN2 [16] generator pre-trained on FFHQ [15]. We obtain the appearance and structure latent codes using the latent optimization technique and the pre-trained parsing encoder, respectively, both provided by the official StyleGANEX project [55]. Our alternating optimization strategy iterates between a two-epoch adaptation step and a two-epoch fine-tuning step, both of which use the Adam optimizer with learning rates of 1×10^{-7} and 1×10^{-6} , respectively. The entire process runs for 300 epochs with a batch size of 5. The loss weights in Eq. 3 are empirically set to $\lambda_{\text{par}} = 0.0025$ and $\lambda_{\text{rec}} = 1.0$. The area filter threshold τ is empirically set to 0.1. Our framework is implemented in PyTorch, and all experiments are conducted on a single NVIDIA RTX A5000 GPU.

Table 1: Quantitative comparison of UDA methods for the 300W to CariFace adaptation, evaluated using NME. Generalization to the unseen ArtiFace domain is also reported. The best result in each column is **bolded**. The † symbol indicates results reproduced by us.

Adaptation Setting	Learning Method	Source domain (300W)			Target domain	Unseen domain	Average NME
		Common	Challenge	Full	CariFace	Artiface	
UDA (300W→CariFace)	BDL [28]	3.28	6.17	3.84	13.63	5.65	9.10
	ASN [44]	2.92	5.26	3.38	12.21	4.75	8.12
	RevGrad [8]	2.84	5.58	3.38	12.19	5.16	8.11
	†FDA [56]	2.87	5.22	3.29	11.18	4.54	7.24
	†CycleGAN [64]	2.86	5.30	3.34	10.67	4.64	7.02
	CFW [29]	2.79	4.91	3.20	7.70	3.95	5.46
	Ours	2.78	4.82	3.18	7.35	3.80	5.26

Table 2: Quantitative comparison of UDA methods for the 300W to ArtiFace adaptation, evaluated using NME. Generalization to the unseen CariFace domain is also reported. The best result in each column is **bolded**. The † symbol indicates results reproduced by us.

Adaptation Setting	Learning Method	Source domain (300W)			Target domain	Unseen domain	Average NME
		Common	Challenge	Full	ArtiFace	Cariface	
UDA (300W→ArtiFace)	BDL [28]	2.99	5.32	3.44	5.90	13.40	8.73
	ASN [44]	2.89	5.81	3.46	5.65	16.58	10.31
	RevGrad [8]	2.99	5.81	3.55	4.74	12.46	8.26
	†FDA [56]	2.79	5.18	3.26	4.51	11.23	7.48
	†CycleGAN [64]	2.93	5.38	3.41	4.90	11.08	7.48
	CFW [29]	2.90	5.14	3.34	3.93	10.93	7.34
	Ours	2.72	5.00	3.17	3.61	10.92	7.25

Computational Complexity. Training involves auxiliary inference for the Parser and StyleGAN, which increases the batch time by roughly $3.1\times$ compared to standard self-training (0.25s to 0.78s) with 22GB VRAM, remaining within the same order of magnitude. Crucially, these auxiliary networks are discarded during inference, keeping inference speed identical to standard detectors.

Datasets. We conduct experiments on three datasets with 68-point landmark annotations: 300W [42], CariFace [2], and ArtiFace [57], along with an unannotated Diverse Generalization Set for qualitative evaluation. These datasets are used to serve as the source, target, and unseen domains.

- **300W.** Our labeled source domain is the 300W dataset, a standard benchmark for in-the-wild real faces. It consists of 3,148 training images and a test set of 689 images, which is further divided into a common subset of 554 images and a challenging subset of 135 images.
- **CariFace.** This dataset serves as our first unlabeled target domain, containing celebrity caricatures with large geometric exaggerations that create a significant domain gap. For domain adaptation experiments, we use its 3,372 training images and 800 test images, the latter of which are also used for unseen domain evaluation.
- **ArtiFace.** This dataset acts as our second unlabeled target domain. It contains artist portraits in diverse styles, presenting a severe appearance-based

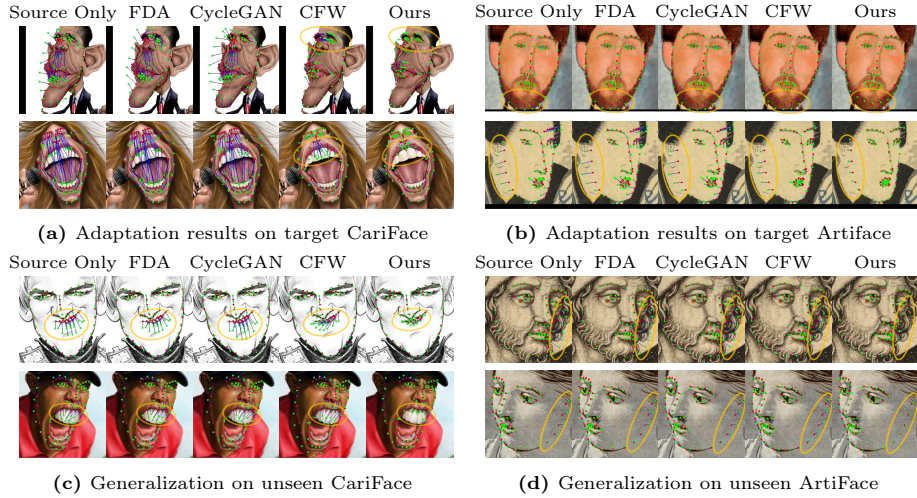


Fig.3: Qualitative comparisons of various UDA methods on stylized domains. Predicted landmarks are shown in green, ground-truth in red, and blue lines indicate the error distances. Orange ellipses highlight challenging regions where our method, PriSM, excels while competitors struggle. (a) & (b) The adaptation results on target datasets, CariFace and Artiface, respectively. (c) & (d) The generalization results on unseen datasets, CariFace and Artiface, respectively. The notation ($X \rightarrow Y$) indicates UDA from the source-domain dataset X to the target-domain dataset Y .

domain shift. In our primary domain adaptation task, the dataset is split into 128 training and 32 test images, while the full set of 160 images is used for unseen domain evaluation.

- **Diverse Generalization Set.** In addition to these benchmarks, we collect a diverse set of unannotated images utilized purely for qualitative testing of generalizability. This collection contains 2,000 sketches [14], 1,000 anime faces [33], 5,153 cat faces, and 4,739 dog faces [5].

Evaluation Metrics. We follow the standard protocol and use the Normalized Mean Error (NME) to evaluate the performance of our landmark detector. The NME is calculated as the average Euclidean distance between the predicted and ground-truth landmarks, normalized by the inter-ocular distance, which is the distance between the outer corners of the eyes. A lower value of NME indicates better performance.

4.2 Comparisons with State-of-the-Art Methods

To demonstrate the effectiveness of our proposed method, we conduct extensive comparisons to the state-of-the-art UDA approaches and report both quantitative metrics and qualitative results.

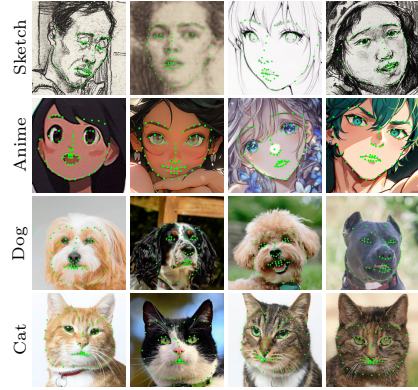


Fig. 4: Qualitative results of our method on the Diverse Generalization Set, including sketches, anime, dogs, and cats.

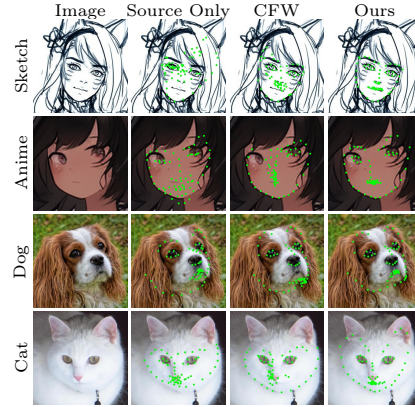


Fig. 5: Qualitative comparison with baselines on the Diverse Generalization Set. Our method shows superior robustness.

Quantitative Comparison. We compare our PriSM against several state-of-the-art UDA methods, all adapting the SLPT backbone and following the protocol of CFW [29]. Methods marked with a dagger ($\dagger$) were reproduced by us, while others are cited from [29].

We perform unsupervised domain adaptation from the source-domain 300W dataset to target-domain CariFace and ArtiFace datasets, and report the results in Table 1 and Table 2, respectively. It can be observed that our proposed PriSM establishes a new state-of-the-art across both adaptation tasks. Specifically, on the 300W-to-CariFace task, PriSM achieves a target domain NME of 7.35, surpassing the SoTA competing method CFW [29]. A similar advantage is observed in the 300W-to-ArtiFace task, where PriSM’s NME of 3.61 again outperforms all competitors. The superior results on the target domains validate that PriSM can accurately evaluate pseudo-label qualities by jointly enforcing high-level parsing map alignment and fine-grained face reconstruction.

Notably, PriSM’s superiority is comprehensive. It maintains the best performance on the source 300W domain by incorporating source data in its alternating optimization to effectively prevent catastrophic forgetting. It also achieves the lowest error on unseen domains, demonstrating a strong generalizability that comes from making fewer assumptions about target-specific styles and instead focusing on fundamental geometric principles.

Qualitative Comparison. We provide visual comparisons to illustrate the advantages of our method. Figures 3a and 3b showcase the adaptation results on the CariFace and ArtiFace datasets, respectively. PriSM demonstrates superior performance, particularly on heavily distorted facial features and ambiguous contours where other methods struggle. This robustness is further validated in our cross-domain generalization experiments, where a model adapted to one stylized domain is tested on the unseen domain. As shown in Figures 3c and 3d, PriSM

Table 3: Parser Reliability Analysis. We evaluate the mean IoU (%) against manual masks across 50 target images per dataset. We compare: (1) **Face Parser**: The raw parsing map M_{ref} generated by the source-trained EHANet; and (2) **Source LM** & (3) **PriSM LM**: The parsing maps M_{pts} converted via our P2P module from unadapted and our adapted landmarks, respectively. Abbreviated classes: BG (Background), Br (Eyebrows), Ey (Eyes), U/L-Lip (Upper/Lower Lips), In-M (Inner Mouth).

CariFace	BG	Skin	L-Br	R-Br	L-Ey	R-Ey	Nose	U-Lip	In-M	L-Lip	mIoU
Face Parser	81.76	71.27	35.95	29.97	34.61	29.26	54.83	33.76	34.99	27.81	43.42
Source LM	66.02	63.47	34.44	30.87	28.33	23.63	53.01	14.47	33.26	20.78	36.83
PriSM LM	66.61	64.18	36.55	33.14	28.76	26.13	56.58	17.87	31.96	21.90	38.37

ArtiFace	BG	Skin	L-Br	R-Br	L-Ey	R-Ey	Nose	U-Lip	In-M	L-Lip	mIoU
Face Parser	83.92	72.09	34.81	34.62	48.50	47.04	55.16	31.82	1.49	35.07	44.45
Source LM	72.10	67.82	37.90	36.46	29.04	30.61	57.94	19.69	5.95	44.90	40.24
PriSM LM	73.22	68.83	39.42	36.57	30.37	32.00	57.79	25.61	7.39	48.80	42.00

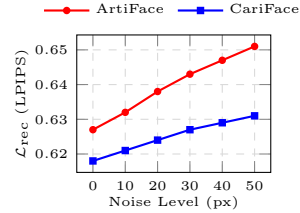


Table 4: Sensitivity analysis of StyleGAN supervision. We perturb the ground-truth landmarks of target images ($N_{\text{Arti}}=50$, $N_{\text{Cari}}=200$) with random noise ranging from 0 to 50 pixels and measure the corresponding reconstruction loss $\mathcal{L}_{\text{rec}}$.

maintains strong performance even on unseen styles, highlighting its remarkable generalizability.

Generalization to Diverse Styles. We further probe PriSM’s limits on a diverse set of completely unseen images in Figures 4 and 5. Even on sketches, anime, and non-human faces, PriSM generates plausible and structurally coherent landmarks, far surpassing the source-only [53] and CFW [29] competitors. This suggests our method learns a more generalizable understanding of facial geometry.

4.3 Reliability of Pre-trained Priors

A core premise of PriSM is the utilization of pre-trained networks (*i.e.*, the face parser and the StyleGAN generator) as supervisory tools. Since these auxiliary networks are trained exclusively on real human faces, it is crucial to verify their reliability when applied to different target domains.

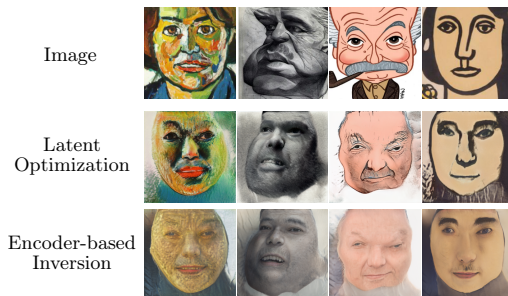
Reliability of the Face Parser. Our Parsing Network is predicated on the insight that capturing broad semantic areas is inherently more resilient to domain gaps than precise landmark regression. To validate this, we evaluate the structural accuracy (mIoU) of three configurations on CariFace and ArtiFace in Table 3, employing our adaptive Area Filter to simulate actual training conditions. As shown in Table 3, the Face Parser achieves the highest overall mIoU, consistently outperforming the unadapted Source LM. This confirms that despite severe domain shifts, the coarse parser remains robust, serving as a reliable structural anchor. Notably, PriSM LM successfully improves parsing consistency over the Source LM across almost all facial components, validating that our Parsing Network guides the geometric coordinates toward structurally plausible layouts.

Table 5: Ablation study on PriSM components. We report the NME on CariFace and ArtiFace after domain adaptation, with the source-only model as a baseline.

Type of loss	Source Only	$\mathcal{L}_{\text{par}}$	$\mathcal{L}_{\text{rec}}$			$\mathcal{L}_{\text{par}} + \mathcal{L}_{\text{rec}}$
		CE	DINO	Grad	LPIPS	CE & LPIPS
CariFace	11.14	7.41	8.09	7.71	8.07	7.35
ArtiFace	4.62	3.86	3.93	3.76	3.70	3.61

Table 6: Quantitative comparison of GAN inversion methods.

	Encoder-based Inversion	Latent Optimization
CariFace	7.37	7.35
ArtiFace	3.73	3.61

**Fig. 6: Qualitative comparison of GAN inversion methods.** While the encoder-based approach tends to generate human-like faces, the latent-optimization method better preserves the inputs’ styles and aligns well with the requirements of this work.

Reliability of the StyleGAN Prior. To verify that the StyleGAN reconstruction loss ($\mathcal{L}_{\text{rec}}$) provides valid gradients for geometry refinement on stylized inputs, we perform a sensitivity analysis. As illustrated in Figure 4, the results reveal a strict monotonic correlation between geometric deviation and reconstruction error. This demonstrates that the human-centric StyleGAN prior effectively generalizes and provides valid supervision, yielding a smooth loss landscape that can stably guide the detector toward the ground truth even on stylized inputs.

4.4 Ablation Studies

Effectiveness of PriSM Components. We conduct ablation studies to validate our dual-consistency framework, with results in Table 5. We first evaluate each auxiliary network independently. Employing only the Parsing Network ($\mathcal{L}_{\text{par}}$), using a Cross-Entropy (CE) loss, already yields a significant improvement over the baseline by enforcing high-level structural alignment. For the StyleMix Network ($\mathcal{L}_{\text{rec}}$), we find that fine-grained losses like the Sobel-based gradient loss (Grad) and perceptual loss (LPIPS) consistently outperform the high-level, semantic-focused DINO loss, as precise landmark validation requires low-level details. While Grad performs well by emphasizing edges, LPIPS shows superior robust-

ness, particularly on the stylistically diverse ArtiFace dataset. Our full PriSM model, which combines the complementary signals of CE parsing and LPIPS reconstruction, achieves the best performance on both datasets. This confirms our hypothesis that the coarse structural guidance from the Parsing Network and the fine-grained geometric supervision from the StyleMix Network work in synergy for robust adaptation.

Effectiveness of PriSM Components. The effectiveness of our StyleMix Network depends on high-fidelity image inversion, which ensures that the reconstruction loss $\mathcal{L}_{\text{rec}}$ accurately reflects geometric errors rather than appearance mismatches. We compare two GAN inversion strategies: an encoder-based approach [40] and our chosen method, per-image latent optimization. The optimization-based technique treats the latent code as a trainable parameter, iteratively refining it to minimize reconstruction error. As shown in Figure 6, the encoder, pre-trained on real faces, generates an appearance latent code biased towards realistic textures instead of the target’s style, which pollutes the reconstruction loss with appearance errors, violating our core principle. In contrast, the latent optimization method accurately captures the appearance of both artistic and caricature inputs, providing the clean signal needed for geometric supervision. The quantitative results in Table 6 confirm that this improved reconstruction quality translates to better landmark accuracy, validating our choice.

Limitations. PriSM is limited to the 68-point landmark definition. This restriction stems from our Point-to-Parse module trained for this point configuration.

5 Conclusion

We introduce PriSM, a novel UDA framework that tackles self-training failure by a dual-consistency validation mechanism. It leverages a Parsing Network for coarse structural alignment and a novel StyleMix Network for fine-grained geometric constraint through an analysis-by-synthesis loop. In extensive experiments, PriSM demonstrates state-of-the-art adaptation performance on two challenging UDA benchmarks, and exhibits strong capabilities in mitigating catastrophic forgetting on the source domain and promoting generalization to diverse unseen domains.

Acknowledgement

This work was supported in part by the National Science and Technology Council (NSTC) under grants 114-2221-E-A49-038-MY3 and 115-2634-F-A49-011, and by MediaTek. This work was supported by H100 GPU computing resources donated by Wistron.

References

1. Agbolade, O., Nazri, A., Yaakob, R., Abd Ghani, A.A., Cheah, Y.K.: Landmark-based homologous multi-point warping approach to 3D facial recognition using multiple datasets. *PeerJ Computer Science* (2020)
2. Cai, H., Guo, Y., Peng, Z., Zhang, J.: Landmark detection and 3D face reconstruction for caricature using a nonlinear parametric model. *Graphical Models* (2021)
3. Chiang, J.C., Hu, H.N., Hou, B.S., Tseng, C.Y., Liu, Y.L., Chen, M.H., Lin, Y.Y.: ORFormer: Occlusion-robust transformer for accurate facial landmark detection. In: *WACV* (2025)
4. Choi, J., Jeong, M., Kim, T., Kim, C.: Pseudo-labeling curriculum for unsupervised domain adaptation. In: *BMVC* (2019)
5. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: StarGAN v2: Diverse image synthesis for multiple domains. In: *CVPR* (2020)
6. Dutta, A., Lal, R., Garg, Y., Ta, C.K., Raychaudhuri, D.S., Cruz, H.D., Roy-Chowdhury, A.K.: POSTURE: Pose guided unsupervised domain adaptation for human body part segmentation. *arXiv* (2024)
7. Feng, Z.H., Kittler, J., Awais, M., Huber, P., Wu, X.J.: Wing loss for robust facial landmark localisation with convolutional neural networks. In: *CVPR* (2018)
8. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: *ICML* (2015)
9. Gong, H., Wang, Y., Wang, Y., Xiao, J., Wan, X., Li, H.: Diffuse-UDA: Addressing unsupervised domain adaptation in medical image segmentation with appearance and structure aligned diffusion models. *arXiv* (2024)
10. liang Hu, S., Sun, H., Li, B., Wei, D., Li, W., Lu, J.Z.: Fast adaptation for human pose estimation via meta-optimization. In: *CVPR* (2024)
11. Huang, J., Guan, D., Xiao, A., Lu, S.: RDA: Robust domain adaptation via fourier adversarial attacking. In: *ICCV* (2021)
12. Huang, Y., Yang, H., Li, C., Kim, J., Wei, F.: ADNet: Leveraging error-bias towards normal direction in face alignment. In: *ICCV* (2021)
13. Jin, H., Che, H., Chen, H.: Unsupervised domain adaptation for anatomical landmark detection. In: *MICCAI* (2023)
14. Ju, X., Zeng, A., Wang, J., Xu, Q., Zhang, L.: Human-art: A versatile human-centric dataset bridging natural and artificial scenes. In: *CVPR* (2023)
15. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: *CVPR* (2019)
16. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. In: *CVPR* (2020)
17. Khan, F.: Facial expression recognition using facial landmark detection and feature extraction via neural networks. *arXiv* (2018)
18. Kim, S.S.Y., Kolkin, N.I., Salavon, J., Shakhnarovich, G.: Deformable style transfer. In: *ECCV* (2020)
19. Kumar, A., Marks, T.K., Mou, W., Wang, Y., Jones, M., Cherian, A., Koike-Akino, T., Liu, X., Feng, C.: LUVLi face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In: *CVPR* (2020)
20. Kumar, A.J.R., Bhanu, B.: Micro-expression classification based on landmark relations with graph attention convolutional network. In: *CVPR* (2021)
21. Lan, X., Hu, Q., Chen, Q., Xue, J., Cheng, J.: HH: Towards more accurate face alignment via heatmap in heatmap. *arXiv* (2021)

22. Lee, C.H., Liu, Z., Wu, L., Luo, P.: MaskGAN: Towards diverse and interactive facial image manipulation. In: CVPR (2020)
23. Li, H., Guo, Z., Rhee, S.M., Han, S., Han, J.J.: Towards accurate facial landmark detection via cascaded transformers. In: CVPR (2022)
24. Li, J., Jin, H., Liao, S., Shao, L., Heng, P.A.: RepFormer: Refinement pyramid transformer for robust facial landmark detection. arXiv (2022)
25. Li, W., Lu, Y., Zheng, K., Liao, H., Lin, C., Luo, J., Cheng, C.T., Xiao, J., Lu, L., Kuo, C.F., et al.: Structured landmark detection via topology-adapting deep graph learning. In: ECCV (2020)
26. Li, Y.J., Dai, X., Ma, C.Y., Liu, Y.C., Chen, K., Wu, B., He, Z., Kitani, K., Vajda, P.: Cross-domain adaptive teacher for object detection. In: CVPR (2022)
27. Li, Y., Kim, G., gi Kwak, J., Ku, B., Ko, H.: Towards multi-domain face landmark detection with synthetic data from diffusion model. In: ICASSP (2024)
28. Li, Y., Yuan, L., Vasconcelos, N.: Bidirectional learning for domain adaptation of semantic segmentation. In: CVPR (2019)
29. Liang, J., Liu, H., Xu, H., Luo, D.: Generalizable face landmarking guided by conditional face warping. In: CVPR (2024)
30. Liu, F., Zhao, Q., Liu, X., Zeng, D.: Joint face alignment and 3D face reconstruction with application to face recognition. IEEE TPAMI (2018)
31. Liu, M.Y., Breuel, T.M., Kautz, J.: Unsupervised image-to-image translation networks. In: NeurIPS (2017)
32. Liu, Y., Shi, H., Si, Y., Shen, H., Wang, X., Mei, T.: A high-efficiency framework for constructing large-scale face parsing benchmark. arXiv (2019)
33. lowres: Anime synthetics. <https://huggingface.co/datasets/lowres/anime-synthetics> (2024), hugging Face Dataset
34. Luo, L., Xue, D., Feng, X.: EHANet: An effective hierarchical aggregation network for face parsing. Applied Sciences (2020)
35. Meng, H., Yuan, F., Tian, Y., Yan, T.: Facial expression recognition based on landmark-guided graph convolutional neural network. Journal of Electronic Imaging (2022)
36. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: ECCV (2016)
37. Onizuka, H., Thomas, D., Uchiyama, H., Taniguchi, R.i.: Landmark-guided deformation transfer of template facial expressions for automatic generation of avatar blendshapes. In: ICCV (2019)
38. Pei, Z., Cao, Z., Long, M., Wang, J.: Multi-adversarial domain adaptation. In: AAAI (2018)
39. Peng, H.Y., Zhang, J.P., Guo, M.H., Cao, Y.P., Hu, S.M.: CharacterGen: Efficient 3D character generation from single images with multi-view pose canonicalization. ACM TOG (2024)
40. Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D.: Encoding in style: a stylegan encoder for image-to-image translation. In: CVPR (2021)
41. Rochow, A., Schwarz, M., Schreiber, M., Behnke, S.: VR facial animation for immersive telepresence avatars. In: IROS (2022)
42. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: ICCV (2013)
43. Sun, Z., Feng, C., Patras, I., Tzimiropoulos, G.: LAFS: Landmark-based facial self-supervised learning for face recognition. In: CVPR (2024)

44. Tsai, Y.H., Hung, W.C., Schuler, S., Sohn, K., Yang, M.H., Chandraker, M.: Learning to adapt structured output space for semantic segmentation. In: CVPR (2018)
45. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: CVPR (2017)
46. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: CVPR (2019)
47. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al.: Deep high-resolution representation learning for visual recognition. IEEE TPAMI (2020)
48. Wang, T., Zhang, B., Zhang, T., Gu, S., Bao, J., Baltrusaitis, T., Shen, J., Chen, D., Wen, F., Chen, Q., et al.: Rodin: A generative model for sculpting 3D digital avatars using diffusion. In: CVPR (2023)
49. Wang, X., Bo, L., Fuxin, L.: Adaptive wing loss for robust face alignment via heatmap regression. In: ICCV (2019)
50. Wang, Y., Peng, J., Zhang, Z.: Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation. In: ICCV (2021)
51. Wood, E., Baltrušaitis, T., Hewitt, C., Johnson, M., Shen, J., Milosavljević, N., Wilde, D., Garbin, S., Sharp, T., Stojiljković, I., et al.: 3D face reconstruction with dense landmarks. In: ECCV (2022)
52. Wu, W., Yang, S.: Leveraging intra and inter-dataset variations for robust face alignment. In: CVPRW (2017)
53. Xia, J., Qu, W., Huang, W., Zhang, J., Wang, X., Xu, M.: Sparse local patch transformer for robust face alignment and landmarks inherent relation learning. In: CVPR (2022)
54. Yang, J., Shi, S., Wang, Z., Li, H., Qi, X.: ST3D: Self-training for unsupervised domain adaptation on 3D object detection. In: CVPR (2021)
55. Yang, S., Jiang, L., Liu, Z., Loy, C.C.: StyleGANEX: StyleGAN-based manipulation beyond cropped aligned faces. In: ICCV (2023)
56. Yang, Y., Soatto, S.: FDA: Fourier domain adaptation for semantic segmentation. In: CVPR (2020)
57. Yaniv, J., Newman, Y., Shamir, A.: The face of art: landmark detection and geometric style in portraits. ACM TOG (2019)
58. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023)
59. Zhang, P., Zhang, B., Zhang, T., Chen, D., Wang, Y., Wen, F.: Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In: CVPR (2021)
60. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Facial landmark detection by deep multi-task learning. In: ECCV (2014)
61. Zhao, X., Mithun, N.C., Rajvanshi, A., Chiu, H.P., Samarasekera, S.: Unsupervised domain adaptation for semantic segmentation with pseudo label self-refinement. In: WACV (2024)
62. Zhou, Z., Li, H., Liu, H., Wang, N., Yu, G., Ji, R.: STAR loss: Reducing semantic ambiguity in facial landmark detection. In: CVPR (2023)
63. Zhu, B., Lin, C., Wang, Q., Liao, R., Qian, C.: Fast and accurate: Structure coherence component for face alignment. arXiv (2020)
64. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)
65. Zou, Y., Yu, Z., Kumar, B.V.K.V., Wang, J.: Domain adaptation for semantic segmentation via class-balanced self-training. In: ECCV (2018)

66. Zou, Y., Yu, Z., Liu, X., Kumar, B.V.K.V., Wang, J.: Confidence regularized self-training. In: ICCV (2019)