Pose-Guided Self-Training with Two-Stage Clustering for Unsupervised Landmark Discovery

Siddharth Tourani[1,2] Ahmed Alwheibi[1] Arif Mahmood[3] Muhammad Haris Khan[1]

[1]Mohamed bin Zayed University of Artificial Intelligence, [2]University of Heidelberg, [3] Information Technology University

tourani.siddharth@gmail.com, muhammad.haris@mbzuai.ac.ae, arif.mahmood@itu.edu.pk

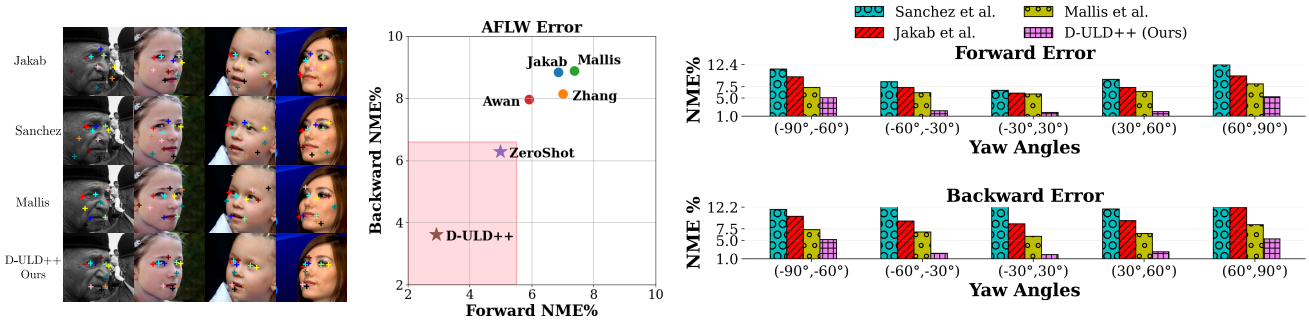

Figure 1. (Left) Visual comparison of proposed D-ULD++ with SOTA. (Middle) Mapping of various SOTA methods to NME space. (Right) D-ULD++ obtains minimum errors across yaw-angle ranges on AFLW Dataset. (Awan et al. [2], Jakab et al.[17], Mallis et al. [32], Sanchez et al.[42], Zhang et al. [64]). The NME metrics are explained in Sec. 4.

Abstract

Unsupervised landmarks discovery (ULD) for an object category is a challenging computer vision problem. In pursuit of developing a robust ULD framework, we explore the potential of a recent paradigm of self-supervised learning algorithms, known as diffusion models. Some recent works have shown that these models implicitly contain important correspondence cues. Towards harnessing the potential of diffusion models for the ULD task, we make the following core contributions. First, we propose a ZeroShot ULD baseline based on simple clustering of random pixel locations with nearest neighbour matching. It delivers better results than existing ULD methods. Second, motivated by the ZeroShot performance, we develop a ULD algorithm based on diffusion features using self-training and clustering which also outperforms prior methods by notable margins. Third, we introduce a new proxy task based on generating latent pose codes and also propose a two-stage clustering to facilitate effective pseudo-labeling, resulting in a significant performance improvement. Overall, our approach consistently outperforms state-of-the-art methods on four challenging benchmarks AFLW, MAFL, CatHeads and LS3D by significant margins. Code and models are available at: https://github.com/skt9/pose-proxy-uld/.

1. Introduction

Background: A large body of work approaches landmark detection for specific object categories in a fully-supervised setting [5, 12, 18, 24, 34, 55]. They assume the availability of sufficiently annotated ground truth data. Common object categories for landmark detection include human faces or bodies as they offer an adequate quantity of images annotated with landmarks [42]. However, akin to other computer vision problems, acquiring a large collection of annotated images for detecting landmarks in an arbitrary object category might be very costly [32]. Therefore, in this paper, we aim to discover object landmarks in an unsupervised way.

Challenges: Unsupervised learning of object landmarks poses significant challenges due to several reasons. Although landmarks are indexed with spatial coordinates, they convey high-level semantic information about object parts which is inherently difficult to learn without human supervision [2, 42]. Detected landmarks should be invariant to different viewpoints, occlusions, and other appearance variations and also capture the shape perception of non-

rigid objects, such as human faces [32, 42]. Some existing approaches to unsupervised landmark detection focus on learning strong representations, that can be mapped to manually annotated landmarks utilizing a few labeled images [50] or leverage proxy tasks, such as imposing equivariance constraints [49–51] or appending conditional image generation [2, 17, 42, 64]. Although obtaining promising performance in some cases, these methods struggle under different intra-class variations of an object, such as large changes in pose and expression (see Fig. 1).

Motivation: To deal with these challenges in unsupervised landmark discovery, we explore the potential of pre-trained diffusion-based generative models [16, 47]. Recent works have shown the utility of diffusion models beyond image synthesis, in tasks such as image editing [4, 33] and image-to-image translation [38]. As such, these models are capable of converting one object into another object without modifying the pose and context of the former [48]. This suggests the presence of implicit correspondence cues in the internal representations of diffusion models for different object classes [30, 48, 62].

Contributions:

- To explore diffusion models for unsupervised landmark detection, we begin with a simple clustering of their pre-trained internal representations indexed at randomly sampled pixel locations from object RoIs. For inference, we use nearest neighbour querying of randomly sampled pixel locations from unseen object image for discovering landmarks. Surprisingly, this *zero-shot baseline* surpasses most existing methods.
- Motivated by the zero-shot baseline, we develop an *unsupervised landmark detection algorithm built on diffusion features*, namely D-ULD, that uses clustering for pseudo-labelling which are then used for self-training. We note that, this simple algorithm produces superlative results and bypasses competing methods by visible margins.
- We capitalize on D-ULD and introduce *a new pose-guided proxy task* which reconstructs landmark heatmaps after producing latent pose codes. To better capture pose variations in landmark representations while clustering, we leverage these latent pose codes to develop *a two-stage clustering mechanism* for better pseudo-labels, resulting in further performance improvement. The inclusion of pose-guided proxy task and two-stage clustering in D-ULD results in a robust ULD algorithm, namely D-ULD++ (see Fig. 1).
- Extensive experiments on challenging datasets, featuring human and cat faces, reveal that D-ULD++ consistently achieves remarkable performance across all datasets.

2. Related Work

Unsupervised Landmark Detection (ULD): An early attempt at ULD involved enforcing equivariance constraints on a landmark detector. These constraints provide self-supervision by requiring the features produced by a detector to be equivariant to the geometric transformations of an image. For ULD, equivariance constraints on image and predefined deformations were used along with auxiliary losses, such as locality [50, 51], diversity [49] and others [8]. Albeit effective in discovering landmarks, such approaches struggle to produce semantically meaningful landmarks under large intra-class variations [32].

A promising class of methods proposed conditional image generation as a proxy task where the landmark detector is required to produce the geometry of an object [2, 17, 42]. The two distinct components in such pipelines are: a bottleneck that captures image geometry and a conditional image generator. Zhang et al. [64] considered landmark discovery as an intermediate step of image representation learning and proposed to predict landmark coordinates utilizing soft constraints. Others combined image generation and equivariance constraints to obtain landmark representations [8, 23, 28]. Lorenz et al. [28] disentangled shape and appearance leveraging equivariance and invariance constraints. Wiles et al. [56] proposed a self-supervised learning for facial attributes from unlabeled videos which are then utilized to predict landmarks by training a linear layer on top of learned embedding. Most of these methods are prone to detecting semantically irrelevant landmarks under large pose variations.

Clustering driven Self-Training: In self-training methods, a model's own predictions are utilized as pseudo-labels (PLs) for generating a training signal. Typical approaches utilize highly confident predictions as hard PLs [45, 58], or via model ensembling [35]. Self-training has been mostly leveraged for image classification [40, 45, 58], but also in other tasks [9, 19, 61].

Some other self-training methods use clustering to construct pseudo-labels [1, 6, 7, 37, 60, 65]. Typically, the clustering serves to assign PLs to the unlabelled training images which are then used to create supervision signal [6, 36]. Some other related methods are slot-attention mechanisms [21, 27]. A few mapped image patches into a categorical latent distribution of learnable embeddings [39, 53], and proposed routing mechanisms based on soft-clustering [25]. In ULD, the work of Mallis et al. [31, 32] forms landmark correspondence through clustering landmark representations. This clustering is used to select pseudo-labels for first stage self-training. We also cluster landmark representations to generate PLs, however, we use it to propose a landmark detection method based on stable diffusion.

Diffusion Model: Diffusion models [11, 16, 46, 47] generate better quality images on ImageNet [10] compared to GANs. Recently, latent diffusion models [41] facilitated their scaling to large scale data [44], also democratising high-resolution image synthesis by introducing the open-

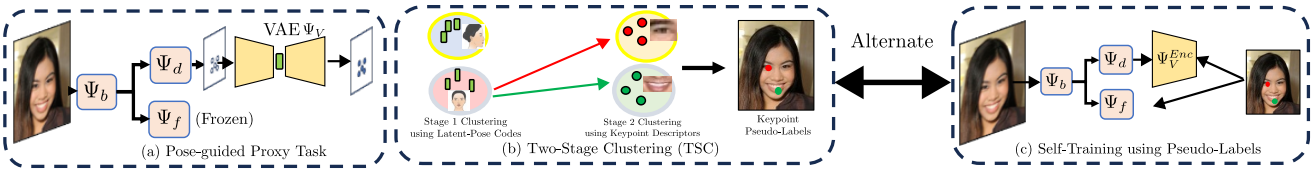

Figure 2. Proposed diffusion based unsupervised landmark detection algorithm D-ULD++: (a) Pose-guided proxy task to reduce noisy landmarks. (b) Two-stage clustering to improve pseudo-labels. (c) Self-training using pseudo-labels.

sourced text-to-image diffusion model, namely Stable Diffusion. Owing to its superlative generation capability, recent works explore the internal representations of diffusion models [3, 15, 52, 59]. For instance, [3, 59] investigate adapting pre-trained diffusion model for downstream recognition tasks. Different from these methods, we explore the potential of pre-trained Stable Diffusion for ULD.

3. Proposed Diffusion Based ULD Algorithm

Existing works targeting ULD struggle under intra-class variations. In pursuit of overcoming these challenges, we explore the potential of pre-trained diffusion-based generative models [16, 41, 47] for ULD. We first perform a simple clustering of the pre-trained internal representations of the diffusion model, taken at random positions within object RoI. At inference, nearest neighbour querying is used to discover zero-shot landmarks (Sec. 3.1). Motivated by the performance of this zero-shot baseline, we propose an unsupervised landmark detection algorithm (D-ULD) founded on diffusion features (Sec. 3.2). We further improve D-ULD by introducing a new pose-guided proxy task (Sec. 3.3.1). It reconstructs landmarks after projecting them to a latent pose space. To better capture the intra-class variations in landmark representations, we exploit this unsupervised latent pose space information to propose a two-stage clustering mechanism (Sec. 3.3.2). After incorporating pose-guided proxy task and two-stage clustering mechanism in D-ULD, we contribute a new algorithm for unsupervised landmark discovery, dubbed as D-ULD++. Fig. 2 provides an overview of D-ULD and D-ULD++ algorithms.

Problem statement: We assume the availability of a set of images $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^{H\times W\times 3}\}$ of a specific object category e.g., faces. After learning an initial set of keypoints on $\mathcal{X}$ via SILK [13], our training set becomes $\mathcal{X} = \{\mathbf{x_j}, \{\mathbf{p_i^j}\}_{i=1}^{N_j}\}$, where $\mathbf{p}_i^j \in \mathbb{R}^2$ is a keypoint in 2D space and N_j is the number of keypoints detected in image j. These learned keypoints don't necessarily correspond to ground truth landmarks. Relying only on $\mathcal{X}$, our aim is to train a model $\Psi : \mathcal{X} \rightarrow \mathcal{Y}$, where $\mathcal{Y} \in \mathbb{R}^{H^o\times W^o\times K}$ is the space of output heatmaps representing confidence maps for each of the K object landmarks we wish to detect [32]. We assume $[N]$ denotes the set $\{1, \ldots, N\}$.

Diffusion model overview: Diffusion models are generative models that approximate the data distribution by denoising a base data distribution (assumed Gaussian). In the forward diffusion process, the input image $\mathbf{I}$ is gradually transformed into a Gaussian noise over a series of T timesteps. Then a sequence of denoising iterations $\epsilon_\theta(\mathbf{I}_t, t)$, parameterized by θ, and $t = 1 \ldots T$ take as input the noisy image $\mathbf{I}_t$ at each timestep and predict the noise ϵ added at that iteration [16]. Latent Diffusion models (LDM) [41], instead of operating on images directly encode images as latent codes $\mathbf{z}$ and perform diffusion process. A decoder maps the latent representation to the image again.

$$\mathcal{L}_{LDM} = \mathbb{E}_{\mathbf{I},t,\epsilon\sim\mathcal{N}(0,1)}\left[||\epsilon - \epsilon_\theta(\mathbf{z}_t, t)||_2^2\right]. \quad (1)$$

The denoiser for LDM consists of self- and cross-attention layers [54]. We make use of the features of a pre-trained Stable Diffusion LDM for our method [41].

3.1. Proposed Zero-Shot Baseline

We propose ZeroShot, a zero-shot baseline to measure the efficacy of diffusion features for the ULD task, comprised of feature aggregation, clustering and exemplar assignment.

Feature aggregation: Feature maps in Stable Diffusion are spread over network layers and diffusion time-steps. So, extracting useful feature descriptors from them is a non-trivial task. Similar to [29], we employ a network that aggregates features over layers and time-steps to obtain a single feature map. The aggregator network, takes as input the features maps $\mathbf{r}_{l,t}$ obtained from Stable Diffusion (SD) for an input image $\mathbf{x}_j$. We upscale $\mathbf{r}_{l,t}$ to image resolution and pass it through a bottleneck layer $\mathbf{B}_l$ [14, 59] to obtain a fixed channel count, and weight it with a mixing weight $\mathbf{w}$. The final aggregated feature map is:

$$\mathbf{F}_{ag} = \sum_{t=1}^{T}\sum_{l=1}^{L}\mathbf{w}_{l,t}\mathbf{B}_l(\mathbf{r}_{l,t}), \{\mathbf{r}_{l,t}\}_{l=1,t=1}^{L,T} = \text{SD}(\mathbf{x}) \quad (2)$$

where L is the number of layers, T is the number of timesteps. There are $L \times T$ combinations of layer indices and timesteps. We learn unique mixing weights $\mathbf{w}_{l,t}$ across all layer and timestep combinations via backpropagation. The overall process of getting feature map $\mathbf{F}_{ag}$ may be defined as: $\mathbf{F}_{ag} = \Psi_b(\mathbf{x})$, where Ψ_b includes both the stable diffusion process and the feature aggregation.

Clustering: For every image in the training set, we randomly sample pixels within the ROI region in the output

feature map $\mathbf{F}_{ag}$. For each sampled pixel, we get a descriptor which is then clustered using K-means and the cluster centroids from the training set are retained. Detectron2 [57] is employed to detect the ROI. On the test set, again random pixels are sampled for each image within ROI, and after their corresponding descriptors are extracted, they are assigned the label of their closest cluster. Each image may then consist of multiple pixel locations belonging to a particular cluster. As some of these can be noisy assignments, we prune the locations so that per image each cluster has a single best location via exemplar assignment.

Exemplar assignment: We assign to each image at most K pixel locations by choosing for each cluster the locations whose descriptor is closest to the cluster centroid in feature space. We discard the remaining locations, leaving for each image at most K pixel locations. These K pixel locations after assignment are considered as discovered landmarks.

Discussion: Fig. 3a shows some qualitative results of ZeroShot on the AFLW dataset. While ZeroShot is able to reliably detect keypoints for most front-oriented faces, it occasionally tends to confuse keypoints on the left and right side of side-oriented faces for some examples. Further, it tends to have poor keypoint localization on side profile faces. Fig. 3b shows the forward and backward NME% for ZeroShot amongst other methods for various yaw angles on the AFLW dataset. Errors for the front-facing angles $(-30^\circ$ to $30^\circ)$ are significantly lower than those of the more side-oriented ones. Compared to [32], an error reduction of 14.6% in front-oriented faces is seen. Whereas a more modest improvement of 4.5% is seen for side-oriented faces (yaw angles $(-60^\circ$ to $-30^\circ)$ and $(30^\circ$ to $60^\circ)$) and 2.1% for more extreme view points (yaw angles $(-90^\circ$ to $-60^\circ)$ and $(60^\circ$ to $90^\circ)$). The limitations of ZeroShot motivate our next contribution.

3.2. Proposed Diffusion-based D-ULD Algorithm

In this section, we describe our unsupervised landmark detection algorithm (D-ULD) based on diffusion features that uses clustering to obtain pseudo-labels (PLs) and employs these PLs for self-training.

Network Structure: We append a descriptor head Ψ_f and a landmark detector head Ψ_d after Ψ_b. The detector head Ψ_d will produce a single-channel heatmap $\mathbf{H}_j = \Psi_d(\Psi_b(\mathbf{x}_j)) \in \mathbb{R}^{H\times W\times 1}$ for image $\mathbf{x}_j$. This heatmap reveals the confidence of the model for an object landmark at a given location. Non-maximum suppression is used to obtain landmarks from $\mathbf{H}_j$ [32]. The descriptor Ψ_f produces for image $\mathbf{x}_j$ a descriptor volume $\mathbf{F}_j = \Psi_f(\Psi_b(\mathbf{x}_j)) \in \mathbb{R}^{H\times W\times D}$, where D is the feature dimension. From $\mathbf{F}_j$, we extract a feature descriptor $f_i^j \in \mathbb{R}^D$ corresponding to the landmark location found by Ψ_d. We denote the network consisting of Ψ_b, Ψ_f and Ψ_d as Ψ.

Bootstrapping: We initially train the network Ψ (except the stable diffusion network) via a self-supervised keypoint training scheme [13]. Specifically, during bootstrapping the network Ψ is trained by learning correspondences between an image and it's augmentation, by minimizing a BCE loss for the detector head Ψ_d and a negative-log-likelihood loss for the descriptor head Ψ_f. We use flips and random rotations as augmentations. After the initial keypoint learning, the training set consists of images, learned keypoints and the corresponding descriptors, *i.e.* $\mathcal{X} = \{\mathbf{x}_j, \{\mathbf{p}_i^j, \mathbf{f}_i^j\} \mid i \in$ keypoints for image $\mathbf{x}_j\}$.

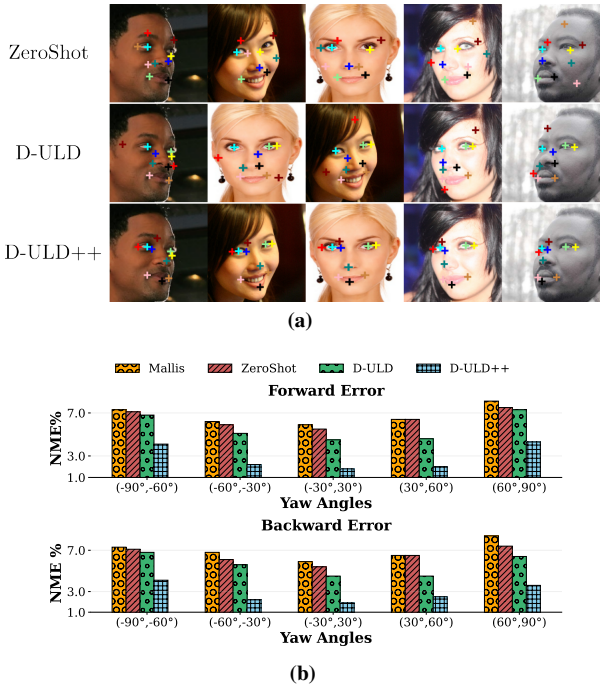

Figure 3. Comparisons of ZeroShot, D-ULD and D-ULD++. (a) Visual results on exemplar images showing different colored keypoints. (b) Yaw angle split of fwd. and bwd. errors (NME%) for AFLW dataset. Mallis [32] is shown for additional comparison.

Self-Training: To improve landmark detection across a larger spectrum of viewpoint changes and symmetric view pairs, we resort to a self-training scheme that uses keypoint pseudo-labelling [31]. We initialize this step by clustering all keypoint descriptors $\{\mathbf{f}_i^j \mid \forall i \in$ keypoints in $\mathbf{x}_j, \forall j \in$ images$\}$ into K clusters using K-means. Then, we perform exemplar assignment and remove redundant keypoints, leaving for each image $\mathbf{x}_j$ a set of K keypoints and their corresponding descriptors $\{\mathbf{p}_i^j, \mathbf{f}_i^j \mid \forall i \in [K]\}$. This provides us with a pseudo-label $\mathbf{c}_i^j \in [K]$ for keypoint $\mathbf{p}_i^j$. Updating the training dataset with pseudo-labels, our dataset is now $\mathcal{X} = \{\mathbf{x}_j, \{\mathbf{p}_i^j, \mathbf{f}_i^j, \mathbf{c}_i^j\}_{i=1}^K\}$.

Training objectives: We use the pseudo-labels $\mathbf{c}_i^j$ contained in the training set $\mathcal{X}$ to train the network Ψ. The loss corresponding to the detector head is the mean-squared error (MSE) loss between the pseudo-label heatmaps and

detector head output. To train the descriptor head Ψ_f, we use a contrastive loss that pulls feature descriptors $\mathbf{f}_i^j$ from the same cluster together and pushes features from different clusters away from each other [32]. Once the training is done for a fixed number of epochs, the training set is updated with improved keypoints, descriptors, and new pseudo-labels. Thus, $\mathbf{p}$, $\mathbf{f}$ and $\mathbf{c}$ are functions of the epoch t. We refer to the training set at epoch t as $\mathcal{X}_t = \{\mathbf{x}_j, \{\mathbf{p}_i^j(t), \mathbf{f}_i^j(t), \mathbf{c}_i^j(t)\}_{i=1}^K\}$.

Discussion: Fig. 3b shows that our D-ULD algorithm improves significantly over ZeroShot across all yaw-angles, In Fig. 3a for D-ULD, landmarks on the eye pupils, nose and mouth are accurately localized while the landmarks around the eyes require further improvement.

3.3. Proposed D-ULD++ Algorithm

While D-ULD algorithm improves upon the ZeroShot baseline across pose variations, it still has a tendency to output semantically non-meaningful landmarks for extreme poses. It is because the landmark descriptors inherently contain local image information, and there is no explicit mechanism in the D-ULD to capture the collective spatial configuration of landmarks. To this end, we design a simple proxy task that takes the predicted landmark heatmaps from D-ULD and aims to reconstruct them after projecting into a low dimensional latent pose space. Next, to better capture the variations caused by large viewpoint changes, we propose a two-stage clustering mechanism.

3.3.1 Pose-guided Proxy Task

A relationship between the facial landmarks and the corresponding pose of that face is quite intuitive. For example, the visibility of particular landmarks can be easily associated with a particular pose. Thus each pose constraints the landmark visibility and location to a relatively smaller space while each set of landmarks constraint the facial pose as well. Given accurate pose and landmark positions, a mapping may be learned in both directions. However, in our case, the pose supervision is not available while the landmark positions are noisy. We therefore propose a proxy task to project the noisy landmarks to a latent pose space and then back project to the original space with the aim of removing the localization noise in landmarks. This proxy task constrains the spatial configuration of landmarks with respect to each other.

For this purpose we append an variational autoencoder (VAE) Ψ_V to the detector head Ψ_d. The input to Ψ_V is the predicted landmark heatmap $\mathbf{H}_j$. Ψ_V is trained to construct a pseudo-ground truth heatmap $\mathbf{G}_j$ formed by placing K 2D-Gaussians corresponding to predicted landmark locations $\{\mathbf{p}_i^j \mid i \in [K]\}$ on a single channel. The output of the encoder produces a latent pose code φ_j corresponding to $\mathbf{H}_j$. This latent code is then used by the decoder to reconstruct the $\mathbf{G}_j$. The dimension of the latent code vector φ_j is fixed to 64.

Discussion: The Ψ_V is trained on only landmark information without any pose supervision. We observe in Section 4, in-spite this limitation, the latent codes obtained after training in-fact capture relevant pose information. As such, a k-Means clustering of the latent codes is performed and the clusters thus obtained contain images corresponding to the same pose-range. The additional supervision by Ψ_V improves both forward and backward NME error by 7.8% and 12.4% on AFLW and 4.9% and 6.4% on MAFL in Table 2.

3.3.2 Two-stage Clustering

In realistic settings, deformable objects like faces can undergo large variations like extreme 3D rotations. A simple clustering of keypoint descriptors is restrictive in capturing such variations due to the inherently local nature of keypoint information [2, 32]. This poses a limitation in achieving a robust ULD method. To overcome this, we devise a two-stage clustering (TSC) mechanism which exploits the pseudo-pose generated by our pose-guided proxy task. In particular, TSC leverages the latent code to perform stage-1 pose-based clustering. Next, in each pose-based cluster, we perform stage-2 clustering of keypoint descriptors from images that belong to the same stage-1 cluster. Thus, TSC essentially decomposes the problem of recovering landmark correspondence under large intra-class variations into two stages. Moreover, it allows us to leverage both local and image-level information by using the learned keypoint descriptors and latent codes. Such a scheme facilitates reduced within-cluster variations compared to the simple clustering.

In TSC, in the first stage, latent codes $\boldsymbol{\varphi}_j$ are partitioned into Q clusters using k-means, and latent code labels $\boldsymbol{u}_j \in [Q]$ are assigned. In the second stage, we perform a further K-means clustering of a collection of keypoint descriptors $\mathbf{f}$ from images belonging to the same pose-based clusters. The number of clusters in the second stage is the same as the number of landmarks K to be detected in a particular dataset. Thus, the two-stage clustering provides us with a total of $Q \times K$ clusters. $[R] = [Q] \times [K]$ denotes the new label set. With this TSC, for given image $\mathbf{x}_j$, we find the pseudo-label $\mathbf{c}_i^j$, by retaining for each cluster $r \in [R]$, the keypoint whose descriptor is nearest to the cluster centroid. We then update the training set with the latent codes $\boldsymbol{\varphi}_j$, their labels $\boldsymbol{u}_j$, as well as the more fine-grained keypoint labels $\mathbf{c}_i^j$ obtained from TSC. $\mathcal{X} = \{\mathbf{x}_j, \{\mathbf{p}_i^j, \mathbf{f}_i^j, \mathbf{c}_i^j\}_{i\in[R]}, \boldsymbol{\varphi}_j, \boldsymbol{u}_j\}$ is the updated dataset. Now, similar to D-ULD, we alternate between the clustering and using the resulting pseudo-labels for supervision.

Training objectives: D-ULD++ is trained in two steps. In

first step, we backprop through Ψ_v using ELBO loss. Similar to D-ULD, the detector head Ψ_d is trained via the MSE loss with the pseudo-labels obtained using TSC, and the the descriptor head Ψ_f is trained using the contrastive loss on keypoint descriptors $\mathbf{f}_i^j$. In the second step, we do not use the decoder and so the ELBO loss. The detector head is trained same as first step, while the descriptor head is kept frozen. Instead, we just use the encoder in Ψ_v and backprop through it using the contrastive loss which encourages latent codes with the same labels to be close to each other:

$$\mathcal{L}_d(\boldsymbol{\varphi}_j, \boldsymbol{\varphi}_{j'}) = \mathbf{1}_{[\boldsymbol{u}_j = \boldsymbol{u}_{j'}]} ||\boldsymbol{\varphi}_j - \boldsymbol{\varphi}_{j'}|| + \mathbf{1}_{[\boldsymbol{u}_j \neq \boldsymbol{u}_{j''}]} \max(0, m - ||\boldsymbol{\varphi}_j - \boldsymbol{\varphi}_{j''}||) \quad (3)$$

where $\boldsymbol{\varphi}_j$ is the output of the VAE Encoder Ψ_V^{Enc}, *i.e.* $\boldsymbol{\varphi}_j = \Psi_V^{Enc}(\Psi_d(\Psi_b(\mathbf{x}_j))$. We select any $\boldsymbol{\varphi}_{j'}$ and $\boldsymbol{\varphi}_{j''}$ such that $\boldsymbol{u}_j = \boldsymbol{u}_{j'}$ and $\boldsymbol{u}_j \neq \boldsymbol{u}_{j''}$ respectively. We provide a pseudo-code for D-ULD++ in the supplementary.

Analysis: As can be seen in Fig. 3b our D-ULD++ brings an improvement compared to D-ULD over all pose variations in aggregate. Fig. 3a shows the variations and the learned landmarks are indeed better localized and more accurately matched across poses.

4. Experiments

Datasets, Evaluation metrics & Baselines: We perform experiments on four diverse datasets: AFLW [22], LS3D [5], CatHeads [63] and MAFL [26]. We follow the same dataset splits and evaluation protocol used in [32]. See supplementary material for details on datasets. For evaluation, we use the standard metrics of Forward and Backward Normalized Mean-squared Error NME% widely used for ULD [31, 32, 42]. In Forward NME, we train a regressor to map the discovered landmarks to the ground truth landmarks and compute NME%. In the Backward NME, we train a regressor to map the ground truth landmarks to discovered landmarks and compute NME%. The mappings are learned using a random subset of images from the dataset. We use the same subsets as [32]. We benchmark against the following baselines: Lorenz et al. [28], Shu et al. [43], Jakab et al. [17], Zhang et al. [64], Sanchez et al. [42], Mallis et al. [32] and Awan et al. [2]. Where possible, we used pre-trained models, otherwise we re-trained these methods. Additionally, as all these methods do not make use of diffusion models, so we also retrain [32] with Stable Diffusion features via the same aggregator network that we use instead of its FAN [5] backbone. We refer to this method as Mallis (D). We use $K = 10$ and $M = 100$ as in the original method using the same training protocol.

Training details: Training is performed in the following sequence. **Bootstrapping:** The network Ψ is initially trained with the Adam [20] optimizer with learning rate 1×10^{-4} and betas $(0.9, 0.999)$; trained for $50K$ iterations, with a batch size of 12 per GPU. **D-ULD:** After bootstrapping, we alternate between clustering and training the network using pseudo-ground truth every 5000 iterations. We train for a total of $100k$ iterations for all four datasets. The margin m used in the contrastive loss is set to 0.8. **Pose-guided Proxy-Task:** For this stage, we initialize the model Ψ with the training of D-ULD, append the VAE Ψ_V to the descriptor head Ψ_d. For this stage, Ψ_f is frozen and Ψ_V is minimized by the ELBO loss. We reduce the learning rate for the Adam optimizer to 5×10^{-5} and train for $50k$ iterations. **D-ULD++:** We discard the decoder of Ψ_V and initialize the rest of the network weights with those obtained after proxy-task training. We train this network for $100k$ iterations using the Adam with the same parameters as before, but with a learning rate of 5×10^{-4}. Every 5000 iterations, we perform clustering and update the pseudo-ground truth. We report results for ZeroShot, D-ULD and D-ULD++ averaged over 5 evaluations and report relative gain throughout.

4.1. Results

Table 1 summarizes the results of our approach and baselines. The simple ZeroShot baseline outperforms previously published methods across all datasets, notably surpassing them by over 30% on LS3D and 10% on AFLW for both forward and backward NME respectively. This emphasizes the efficacy of Stable Diffusion features. The interesting comparison is between Mallis (D) and D-ULD++. Mallis (D) simply oversegments the initial clusters into more fine-grained clusters (the initial K clusters are segmented into $M >> K$ clusters), while D-ULD++ proposes a pose-guided proxy task and a two-stage clustering. D-ULD++ outperforms Mallis (D) by notable margins. It also shows a relative improvement of more than 20% over published methods on all four datasets.

Figure 5 displays the Cumulative Error Distribution (CED) curves. Our D-ULD++ exhibits notably lower errors than others. Compared to others, our method's curve displays a gradual slope, suggesting a gradual decline in keypoint localization accuracy. In contrast, the curves for other methods start with higher base errors and exhibit steeper slopes, indicating a failure in keypoint detection/localization accuracy beyond a certain number of keypoints. Figure 6 provides a visual comparison of landmarks discovered by D-ULD++ and other baselines.Figure 4 gives further insight into the localization accuracy and reliability of the landmarks detected by D-ULD++. In MAFL, that has 68 facial landmarks configuration, each landmark is aligned with the best-matching unsupervised landmarks using the Hungarian algorithm. A value of $K = 30$ is employed across all methods. Notably, our detected landmarks exhibit top accuracy in tracking semantically relevant facial landmarks. D-ULD++ has almost perfect accuracy in detecting keypoints in the eyebrow, nose, eyes and mouth region.

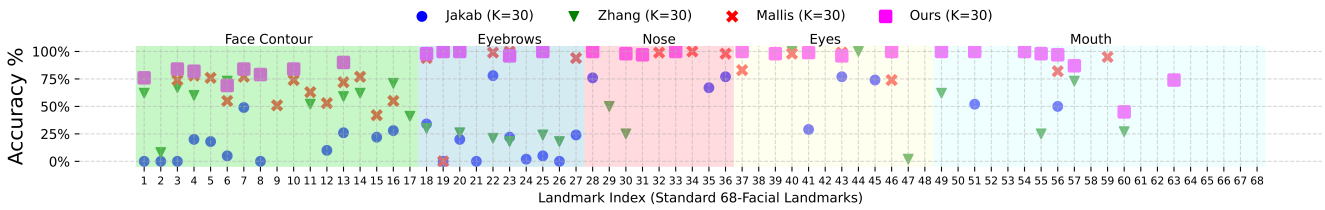

Figure 4. Evaluation of the ability of raw unsupervised landmarks to capture supervised landmark locations on MAFL. Each unsupervised landmark is mapped to the best corresponding supervised landmark using the Hungarian Algorithm. Then accuracy is calculated for a distance threshold of $0.2 \cdot d_{iod}$ to a landmark location, where d_{iod} is the interocular distance. Accuracy is shown for each of the 68-facial landmarks sorted by ascending order of index. Different landmark areas are highlighted with different colours and labelled as such (1-17 face contour, 18-27 eyebrows etc.

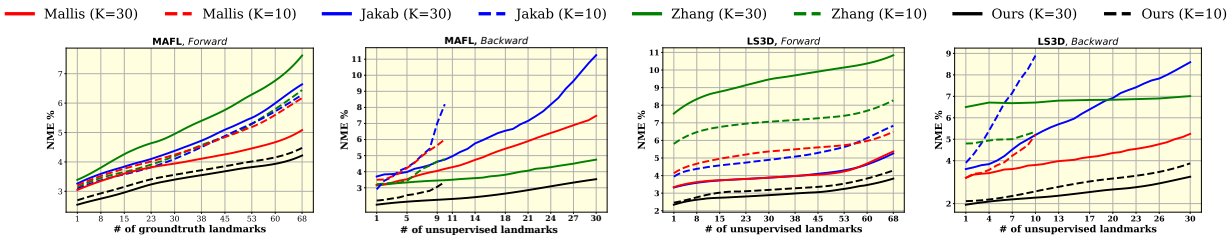

Figure 5. Cumulative Error Distribution (CED) Curves of forward and backward NME for MAFL and LS3D.

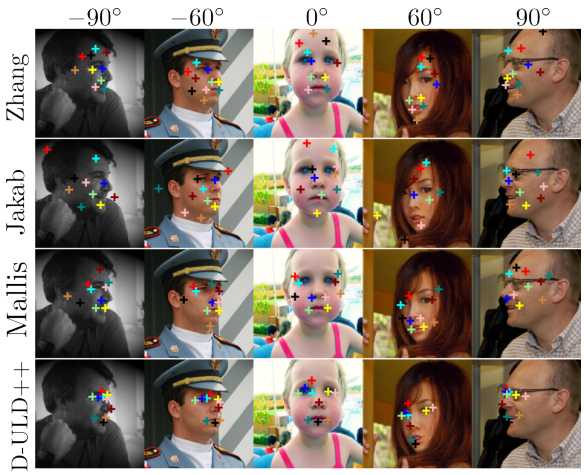

Figure 6. Landmarks discovered across poses in LS3D.

5. Ablation and Analysis

Performance contribution of our methods: Over ZeroShot baseline, our D-ULD provides consistent improvement in both forward and backward NME (Table 2). Upon introducing the pose-guided proxy task (PP) in D-ULD, we see a further improvement in all instances. After including our two-stage clustering (TSC) with PP in D-ULD, we note a consistent notable improvement over each of the variants. To assess the contribution of Equation (3) we compare it to training the network with the full VAE. The VAE output is supervised by the ELBO loss. D-ULD++ is superior to full VAE alternative, justifying use of Equation (3).

Effect of TSC-guided Self-Training: To get a better understanding of whether the pseudo-labels from TSC do indeed capture pose variations, we do a simple cluster analysis. Recall D-ULD++ generates a latent pose code for each image landmarks, *i.e.* there is a *1-to-1* relation between the latent code and image landmarks. These latent codes are then clustered into Q clusters. As the LS3D images come with the 5 yaw-angle ranges (the ranges are shown in Figure 7) they belong to, we cluster the latent codes into $Q = 5$ clusters. For the latent codes within each cluster, we find which yaw-angle range the majority of the latent codes belong to (by checking the range of their corresponding images) and map each cluster to a yaw-angle range. We find that the $Q = 5$ clusters neatly map to each of the 5 ranges. In the process, we also measure the percentage of latent codes that lie within the clusters assigned range. We refer to this percentage as clustering accuracy %. Figure 7 shows the cluster accuracy for $K = 10$ clusters as a function of iterations for the various pose ranges. As can be clearly seen, clustering accuracy increases up to iteration $50k$ before levelling off. Table 2 shows TSC outperforms the simple clusterinig in D-ULD, achieving a higher Silhouette coefficient and Calinski-Harabasz (CH) Index. This indicates TSC's superiority in forming compact clusters.

Hyper-Parameter Study: Figure 8 shows the effect of varying the number of clusters for k-Means on Forward and Backward NME for both keypoint and pseudo-pose clustering. K and Q are the number of clusters for keypoint and

	Method	MAFL		AFLW		LS3D		CatHeads	
		F	B	F	B	F	B	F	B
Published	Jakab [17] (K=10)	3.19	4.53	6.86	8.84	5.38	7.06	4.53	4.06
	Zhang [64] (K=10)	3.46	4.91	7.01	8.14	6.74	7.21	4.62	4.15
	Sanchez [42] (K=10)	3.99	14.74	6.69	25.84	26.41	5.44	4.42	4.17
	Mallis [32] (K=10)	3.19	4.23	7.37	8.89	6.53	6.57	9.31	10.08
	Awan [32]	3.50	5.18	5.91	7.96	5.21	4.69	3.76	3.94
	Mallis (D) (K=10)	2.74	3.11	3.38	3.75	2.89	3.76	3.14	3.62
	Zero Shot (K=10)	3.14	3.27	4.98	6.29	3.53	4.14	3.47	3.59
		+1.19%	+6.14%	+15.74%	+20.97%	+34.35%	+11.72%	+7.7%	+8.88%
	D-ULD++ (K=10)	**2.19**	**2.78**	**2.92**	**3.62**	**2.12**	**2.85**	**2.89**	**3.12**
		+31.34%	+34.28%	+50.91%	+54.52%	+59.31%	+39.23%	+23.13%	+20.81%

Table 1. Error comparison on MAFL, AFLW, LS3D and CatHeads, in Forward and Backward NME (denoted as F and B). The results of other methods are taken directly from the papers (for the case where all MAFL training images are used to train the regressor and the error is measured w.r.t. to 5 annotated points). A set of 300 training images is used to train the regressors. Error is measured w.r.t. the 68-landmark configuration typically used in face alignment. The best performance is in **bold**. Below ZeroShot and D-ULD++ are the (relative) percentage improvements shown in blue over the best published results (shown underlined).

Method	AFLW		MAFL	
	F	B	F	B
ZeroShot	4.98	6.29	3.14	3.98
D-ULD	3.42	4.89	2.82	3.21
+ PP	3.15	4.46	2.68	3.12
+ (TSC w/ Full VAE)	3.07	4.12	2.59	3.03
+ TSC (D-ULD++)	**2.92**	**3.62**	**2.19**	**2.78**

Table 2. Ablation Evaluation on AFLW and MAFL. PP stands for pose-guided proxy task and TSC stands for two-stage clustering.

Methods	MAFL		AFLW	
	Sil.	CH	Sil.	CH
ZeroShot ($K = 30$)	0.74	357.4	0.72	142.9
D-ULD ($K10$)	0.79	372.1	0.81	182.3
D-ULD ($K = 30$)	0.82	382.6	0.83	190.7
D-ULD++ ($Q = 10, K = 10$)	0.86	377.4	0.86	192.3
D-ULD++ ($Q = 10, K = 30$)	0.87	392.6	086	194.7

Table 3. Quality of clustered landmark representations using Silhouette coefficient (Sil.) and Calinski-Harabasz (CH) Index. Q and K are the number of pose and keypoint clusters.

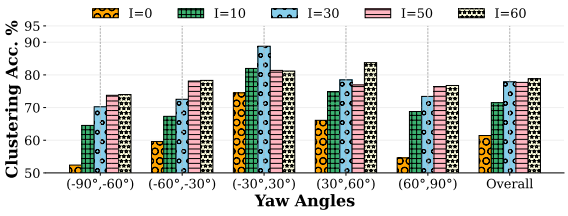

Figure 7. Clustering accuracy percentages are reported across different yaw-angle ranges and overall for D-ULD++ on LS3D. The variable I denotes the iteration number, as a multiple of 1000.

pseudo-pose clustering respectively (Section 3). Increasing Q and K has the effect of lowering both NMEs. As K increases the keypoints captured with each cluster are more specific and localized. Similarly, increasing Q would have the effect of capturing finer pose variations. This saturates at $Q = 30$ and $K = 30$, but this maybe due to the limitation of the dataset size.

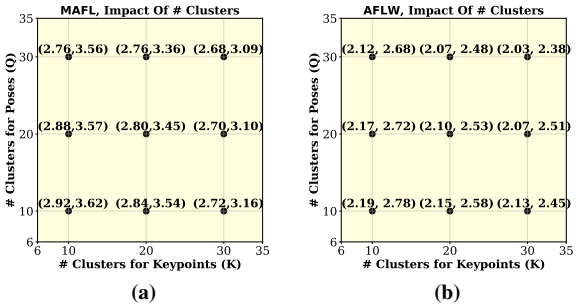

Figure 8. Impact of number of clusters on performance. Along the X-axis and Y-axis number of clusters for keypoints and poses are shown respectively. For each permutation of number of clusters the forward and backward NME is shown as a tuple.

6. Conclusion

We explored the effectiveness of Stable Diffusion to tackle unsupervised landmark detection (ULD) problem. We first proposed a ZeroShot baseline that is only based on clustering of diffusion features and nearest neighbour querying that excels in performance than SOTA methods. This motivated us to develop ULD algorithms that involves fine-tuning diffusion features. We propose D-ULD which shows the effectiveness of fine-tuning and D-ULD++ which proposes a novel proxy task and two-stage clustering mechanism based on this proxy task. Our methods consistently achieves significant improvement over existing baselines across all datasets.

References

[1] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. *arXiv preprint arXiv:1911.05371*, 2019.

[2] Mamona Awan, Muhammad Haris Khan, Sanoojan Baliah, Muhammad Ahmad Waseem, Salman Khan, Fahad Shahbaz Khan, and Arif Mahmood. Unsupervised landmark discovery using consistency guided bottleneck. *arXiv preprint arXiv:2309.10518*, 2023.

[3] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021.

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023.

[5] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE ICCV*, pages 1021–1030, 2017.

[6] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149, 2018.

[7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

[8] Zezhou Cheng, Jong-Chyi Su, and Subhransu Maji. On equivariant and invariant learning of object landmark representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9897–9906, 2021.

[9] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1635–1643, 2015.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[12] Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Style aggregated network for facial landmark detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 379–388, 2018.

[13] Pierre Gleize, Weiyao Wang, and Matt Feiszli. Silk–simple learned keypoints. *arXiv preprint arXiv:2304.06194*, 2023.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[17] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. *NeurIPS*, pages 13520–13531, 2018.

[18] Muhammad Haris Khan, John McDonagh, and Georgios Tzimiropoulos. Synergy between face alignment and tracking via discriminative global consensus optimization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3811–3819. IEEE, 2017.

[19] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 876–885, 2017.

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

[21] Thomas Kipf, Gamaleldin F Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional object-centric learning from video. *arXiv preprint arXiv:2111.12594*, 2021.

[22] Martin Koestinger, Paul Wohlhart, Peter M Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *2011 IEEE ICCV workshops*, pages 2144–2151. IEEE, 2011.

[23] Tejas D Kulkarni, Ankush Gupta, Catalin Ionescu, Sebastian Borgeaud, Malcolm Reynolds, Andrew Zisserman, and Volodymyr Mnih. Unsupervised learning of object keypoints for perception and control. *Advances in neural information processing systems*, 32, 2019.

[24] Abhinav Kumar, Tim K Marks, Wenxuan Mou, Ye Wang, Michael Jones, Anoop Cherian, Toshiaki Koike-Akino, Xiaoming Liu, and Chen Feng. Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8236–8246, 2020.

[25] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pages 3744–3753. PMLR, 2019.

[26] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[27] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob

Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in Neural Information Processing Systems*, 33:11525–11538, 2020.

[28] Dominik Lorenz, Leonard Bereska, Timo Milbich, and Bjorn Ommer. Unsupervised part-based disentangling of object shape and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10955–10964, 2019.

[29] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. In *Advances in Neural Information Processing Systems*, 2023.

[30] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *arXiv preprint arXiv:2305.14334*, 2023.

[31] Dimitrios Mallis, Enrique Sanchez, Matthew Bell, and Georgios Tzimiropoulos. Unsupervised learning of object landmarks via self-training correspondence. *Advances in Neural Information Processing Systems*, 33, 2020.

[32] Dimitrios Mallis, Enrique Sanchez, Matt Bell, and Georgios Tzimiropoulos. From keypoints to object landmarks via self-training correspondence: A novel approach to unsupervised landmark discovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[33] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

[34] Xin Miao, Xiantong Zhen, Xianglong Liu, Cheng Deng, Vassilis Athitsos, and Heng Huang. Direct shape regression networks for end-to-end face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5040–5049, 2018.

[35] Duc Tam Nguyen, Chaithanya Kumar Mummadi, Thi Phuong Nhung Ngo, Thi Hoai Phuong Nguyen, Laura Beggel, and Thomas Brox. Self: Learning to filter noisy labels with self-ensembling. In *International Conference on Learning Representations*, 2019.

[36] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016.

[37] Mehdi Noroozi, Ananth Vinjimoor, Paolo Favaro, and Hamed Pirsiavash. Boosting self-supervised learning via knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9359–9367, 2018.

[38] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023.

[39] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

[40] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *International Conference on Learning Representations*, 2020.

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[42] Enrique Sanchez and Georgios Tzimiropoulos. Object landmark discovery through unsupervised adaptation. *NeurIPS*, 32:13520–13531, 2019.

[43] Enrique Sánchez-Lozano, Georgios Tzimiropoulos, Brais Martinez, Fernando De la Torre, and Michel Valstar. A functional regression approach to facial landmark tracking. *IEEE transactions on pattern analysis and machine intelligence*, 40(9):2037–2050, 2017.

[44] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[45] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

[46] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[47] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[48] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *arXiv preprint arXiv:2306.03881*, 2023.

[49] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *Proceedings of the IEEE international conference on computer vision*, pages 5916–5925, 2017.

[50] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object frames by dense equivariant image labelling. *arXiv preprint arXiv:1706.02932*, 2017.

[51] James Thewlis, Samuel Albanie, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of landmarks by descriptor vector exchange. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6361–6371, 2019.

[52] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023.

[53] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[54] 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.

[55] 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*, pages 6971–6981, 2019.

[56] Olivia Wiles, A Koepke, and Andrew Zisserman. Self-supervised learning of a facial attribute embedding from video. *arXiv preprint arXiv:1808.06882*, 2018.

[57] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[58] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020.

[59] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023.

[60] Xueting Yan, Ishan Misra, Abhinav Gupta, Deepti Ghadiyaram, and Dhruv Mahajan. Clusterfit: Improving generalization of visual representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6509–6518, 2020.

[61] Dingwen Zhang, Junwei Han, and Yu Zhang. Supervision by fusion: Towards unsupervised learning of deep salient object detector. In *Proceedings of the IEEE international conference on computer vision*, pages 4048–4056, 2017.

[62] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *arXiv preprint arXiv:2305.15347*, 2023.

[63] Weiwei Zhang, Jian Sun, and Xiaoou Tang. Cat head detection-how to effectively exploit shape and texture features. In *ECCV*, pages 802–816. Springer, 2008.

[64] Yuting Zhang, Yijie Guo, Yixin Jin, Yijun Luo, Zhiyuan He, and Honglak Lee. Unsupervised discovery of object landmarks as structural representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2694–2703, 2018.

[65] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6002–6012, 2019.