Unsupervised Discovery of Object Landmarks as Structural Representations

Yuting Zhang[1], Yijie Guo[1], Yixin Jin[1], Yijun Luo[1], Zhiyuan He[1], Honglak Lee[2,1]
[1]University of Michigan, Ann Arbor [2]Google Brain
{yutingzh, guoyijie, jinyixin, lyjtour, zhiyuan, honglak}@umich.edu honglak@google.com

Abstract

Deep neural networks can model images with rich latent representations, but they cannot naturally conceptualize structures of object categories in a human-perceptible way. This paper addresses the problem of learning object structures in an image modeling process without supervision. We propose an autoencoding formulation to discover landmarks as explicit structural representations. The encoding module outputs landmark coordinates, whose validity is ensured by constraints that reflect the necessary properties for landmarks. The decoding module takes the landmarks as a part of the learnable input representations in an end-to-end differentiable framework. Our discovered landmarks are semantically meaningful and more predictive of manually annotated landmarks than those discovered by previous methods. The coordinates of our landmarks are also complementary features to pretrained deep-neural-network representations in recognizing visual attributes. In addition, the proposed method naturally creates an unsupervised, perceptible interface to manipulate object shapes and decode images with controllable structures.

1. Introduction

Computer vision seeks to understand object structures that reflect the physical states of objects and show invariance to individual appearance changes. Such intrinsic structures can serve as intermediate representations for high-level visual understanding. However, manual annotations or designs of object structures (e.g., skeleton, semantic parts) are costly and barely available for most object categories, making the automatic representation learning of object structure an attractive solution to this challenge.

Modern neural networks can learn latent representations to effectively solve various vision problems, including image classification [26, 53, 56, 20], segmentation [32, 40, 21], object detection [17, 80, 49], human pose estimation [39], 3D reconstruction [13, 67, 14], and image generation [25, 18, 43]. Several existing studies [17, 76, 1] observe that these representations naturally encode massive templates of particular visual patterns. However, little evidence suggests that deep neural networks can naturally conceptualize the intrinsic structures of an object category compactly and perceptibly.

We aim at learning the physical parameters of conceptualized object structures without supervision. As a typical representation of intrinsic structures, landmarks represent the spatial configuration of stable local semantics across different object instances of the same category. Thewlis et al. [59] proposed an unsupervised method to locate landmarks at the places where a convolutional neural network can detect stable visual patterns with high spatial equivariance to image transformations. However, this method did not explicitly encourage the landmarks to appear at critical locations for image modeling.

This paper addresses the problem of discovering landmarks in a generic image modeling process. In particular, we take landmark discovery as an intermediate step for image autoencoding. To leverage the training signals from the landmark-based image decoder, gradients need to go through the landmark coordinates, which makes Thewlis et al. [59]'s non-differentiable formulation infeasible. With a different way to calculate landmark coordinates, the image decoding module can make the landmark configuration informative regarding image reconstruction. We also introduce additional regularization terms to enforce the desirable properties of the detected landmarks and to prevent the landmark coordinates from encoding irrelevant or redundant latent information.

Our contributions in this paper are as follows.

1. We develop a differentiable autoencoder framework for object landmark discovery, which allows the image decoder to propagate training signals back to the landmark detection module. We introduce several soft constraints to reflect the properties of landmarks, forcing the discovered representations to be valid landmarks.
2. The proposed method discovers visually meaningful landmarks without supervision for a variety of objects. It outperforms the state-of-the-art method regarding the accuracy of predicting manually-annotated landmarks using discovered landmarks, and it performs comparably to fully supervised landmark detectors trained with a significant amount of labeled data.
3. The discovered landmarks show strong discriminative performance in recognizing visual attributes.
4. Our landmark-based image decoder is useful for controllable image decoding, such as object shape manipulation and structure-conditioned image generation.

2. Related work

Discriminative part learning. Parts are commonly used object structures in computer vision. The deformable part-based model [15] learns object part configurations to optimize the object detection accuracy, where similar ideas are rooted in earlier constellation approaches [16, 66, 6]. A recent method [72] based on the deep neural network performs end-to-end learning of deformable mixture of parts for pose estimation. The recurrent architecture [19] and spatial transformer network [23] are also used to discover and refine object parts for fine-grained image classification [27]. In addition, discriminative mid-level patches can be also discovered without explicit supervision [54]. Object-part discovery based on subspace analysis and clustering techniques is also shown to improve neural-network-based image recognition [52]. Unlike the approaches specific to discriminative tasks, our work focuses on learning landmarks for generic image modeling.

Learning structural representations. To capture the intrinsic structures of objects, existing studies [44, 45, 37] disentangle visual content into multiple factors of variations, like the camera viewpoint, motion, and identity. The physical parameters of these factors are, however, still embedded in non-perceptible latent representations. Methods based on multi-task learning [78, 21, 65, 81] can take conceptualized structures (e.g., landmarks, masks, depth) as additional outputs. These structures in this setting are designed by humans and require supervision to learn.

Learning explicit structures for image correspondence. Object structures create correspondence among object instances. Colocalization [57, 9] realizes the coarsest level of object correspondence. In a finer granularity, AnchorNet [41] learns object parts and their correspondence across different objects and categories. WarpNet [24] corresponds images in the same class by estimating the parameter of a thin plate spline (TPS) transformation [4], and it can roughly reconstruct 3D point cloud using a single-view image. The 3D interpreter network [67] utilizes 2D landmark annotations to discover 3D skeletons as the explicit structures of objects. Our discovered landmarks are denser than object parts and sparser than 3D points. These landmark representations are also more sensitive to precise locations and obtained without supervision.

Landmark discovery with equivariance. Object structures like landmarks should be equivariant to image transformation, including object and camera motions. Using this property in 2D image domain, Rocco et al. [50] proposed to discover TPS control points to match pairs of object images densely. Thewlis et al. [58] tried to densely map different objects to a canonical coordinate that reflects object structures. Instead of learning dense correspondence, Thewlis et al. [59] took the same equivariance property as the guidance to train deep neural networks for object landmark discovery without manual supervision. A similar idea was formulated differently using hand-crafted features in early work [30]. In comparison, our method not only takes the equivariance as a constraint to ensure the validity of the landmarks, but also use a differentiable formulation to incorporate the landmark coordinates into a generic image modeling process. Moreover, our discovered landmarks are more predictive of manually annotated landmarks than those obtained by Thewlis et al. [59], and our method works on a broader range of object categories.

Image modeling with landmarks. Many unsupervised deep learning techniques exist to model visual content, including stacked autoencoders (SAE) [2, 36], variational autoencoders [25], generative adversarial networks (GAN) [18, 43], and auto-regressive networks [63] (e.g., PixelCNN [62]). The GAN- and PixelCNN-based image generators conditioned on given object landmarks are proposed in [46, 47]. In contrast, our method uses the SAE framework to automatically discover landmarks that are informative for unsupervised image modeling.

Landmark detection. A vast amount of supervised landmark detection methods exist in the literature. For human faces, there are active appearance models [10, 38, 11], template-based methods [42, 83], regression-based methods [61, 12, 7, 48], and more recent methods based on deep neural networks [55, 77, 81, 82, 75, 70, 68, 33, 71]. Landmark detection methods are also available for human bodies [73, 60, 39], and birds [75]. We use our discovered landmarks to predict manually annotated landmarks and compare our method with some recent supervised models.

3. Autoencoding-based landmark discovery

We aim at automatically discovering landmarks as an explicit representation of visual content. We propose an autoencoder that encodes landmark coordinates as (a part of) the encoder outputs (Section 3.1). Without supervision from hand-crafted labels, we introduce several constraints to encourage the discovered landmark coordinates to reflect the visual concept that agrees with human perception (Section 3.2). The proposed constraints prevent landmark-based representations from degenerating to non-perceptible latent representations. Another pathway of the encoder extracts the local latent descriptor for each discovered landmark (Section 3.3). We use both the landmarks and the latent descriptors to reconstruct the input image (Section 3.4). This section presents the fully differentiable neural network architecture (Figure 1) and training objectives (Section 3.5) for landmark discovery and unsupervised image modeling.

3.1. Architecture of landmark detector

We formulate landmark localization as the problem of detecting particular keypoints in the image [39]. Specifically, each landmark has a corresponding detector, which convolutionally outputs a detection score map with the de-

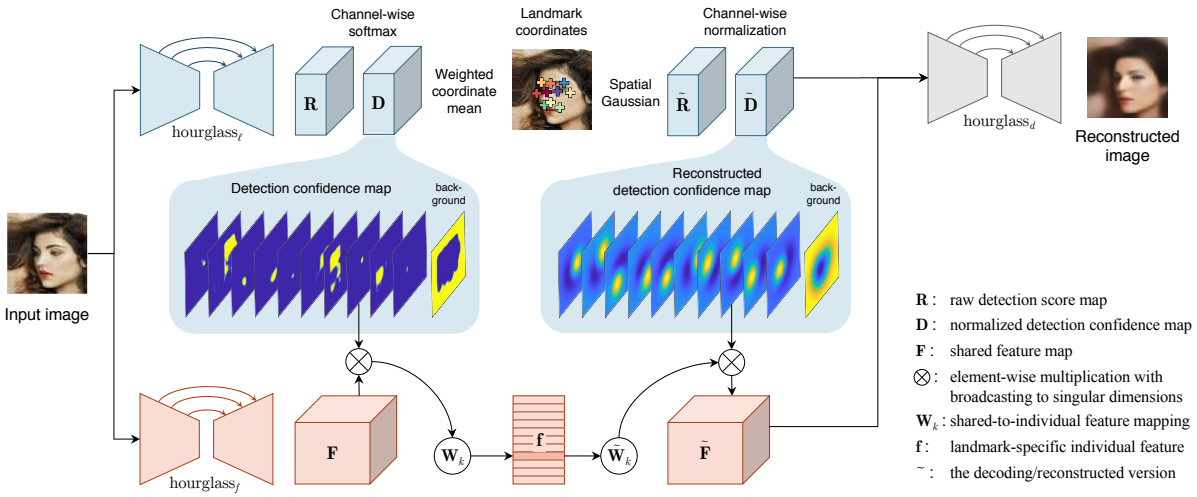

Figure 1: Neural network architectures of our autoencoding framework for unsupervised landmark discovery. See text for the details.

tected landmark located at the maximum. In this framework, we use a deep neural network to transform an image $\mathbf{I}$ to a $(K+1)$-channel detection confidence map $\mathbf{D} \in [0,1]^{W\times H\times(K+1)}$. This map detects K landmarks, and the $(K+1)$-th channel represents background. $\mathbf{D}$'s resolution $W \times H$ can be either equal to or less than that of $\mathbf{I}$, but they should have the same aspect ratio.

Inspired by the success of the stacked hourglass network in human pose estimation [39], we propose a light-weighted hourglass-style network to get the raw detection score map

$$\mathbf{R} = \text{hourglass}_\ell(\mathbf{I}; \theta_\ell) \in \mathbb{R}^{W\times H\times(K+1)}, \quad (1)$$

where θ_ℓ denotes the parameters. The hourglass-style architecture (Appendix G.2) allows detectors to focus on the critical local patterns at landmark locations while utilizing higher-level context. Then, we transform the unbounded raw scores to probabilities and encourage each channel to detect a different pattern. To this end, we normalize $\mathbf{R}$ across the channels (including the background) using softmax and obtain the detection confidence map

$$\mathbf{D}_k(u,v) = \frac{\exp(\mathbf{R}_k(u,v))}{\sum_{k'=1}^{K+1}\exp\left(\mathbf{R}_{k'}(u,v)\right)}, \quad (2)$$

where the matrix $\mathbf{D}_k$ is the k-th channel of $\mathbf{D}$, and the scalar $\mathbf{D}_k(u,v)$ is the value of $\mathbf{D}_k$ at the pixel (u,v). Later, we also use the vector $\mathbf{D}(u,v) \in [0,1]^{K+1}$ to denote the multi-channel values of $\mathbf{D}$ at (u,v). The same notation convention applies to other tensors of three axes.

Taking $\mathbf{D}_k$ as a weighting map, we use the weighted mean coordinate as the location of the k-th landmark, i.e.,

$$(x_k, y_k) = \frac{1}{\zeta_k}\sum_{v=1}^{H}\sum_{u=1}^{W}(u,v)\cdot\mathbf{D}_k(u,v), \quad (3)$$

where $\zeta_k = \sum_{v=1}^{H}\sum_{u=1}^{W}\mathbf{D}_k(u,v)$ is the spatial normalization factor. This formulation enables back-propagating the gradient from the downstream neural network through the landmark coordinates unless $\mathbf{D}_k$'s mass is totally concentrated in a single pixel or totally uniformly distributed, which rarely happens in practice. As a shorthand notation, we write the landmarks and landmark detector as

$$\boldsymbol{\ell} = [x_1, y_1, \ldots, x_K, y_K]^\top = \text{landmark}(\mathbf{I}; \theta_\ell). \quad (4)$$

The left half of the blue pathway in Figure 1 illustrates the landmark detector.

3.2. Visual concept of landmarks

The elements in $\boldsymbol{\ell}$ are supposed to be the discovered landmark coordinates, but so far, there is no guarantee to prevent them from being arbitrary latent representations. Therefore, we propose the following soft constraints as regularizers to enforce the desirable properties for landmarks.

Concentration constraint As a detection confidence map for a single location, the mass of $\mathbf{D}_k$ need to be concentrated in a local region. Taking $\mathbf{D}_k/\zeta_k$ (spatially normalized as in (3)) as the density of a bivariate distribution on the image coordinate, we compute its variance $\sigma^2_{\text{det},u}$ and $\sigma^2_{\text{det},v}$ along the two axes. We define the concentration constraint loss as follows to encourage both variances to be small:

$$L_{\text{conc}} = 2\pi e\left(\sigma^2_{\text{det},u} + \sigma^2_{\text{det},v}\right)^2. \quad (5)$$

This equation makes L_{conc} the exponential of the entropy of the isotropic Gaussian distribution $\mathcal{N}((x_k, y_k), \sigma^2_{\text{det}}\mathbb{I})$, where $\sigma^2_{\text{det}} = (\sigma^2_{\text{det},u} + \sigma^2_{\text{det},v})/2$, and $\mathbb{I}$ is the identity matrix. This Gaussian distribution is an approximation of $\mathbf{D}_k/\zeta_k$, and lower entropy means a more peaked distribution. Note that, formally, this approximation is

$$\overline{\mathbf{D}}_k(u,v) = (1/WH)\mathcal{N}\left((u,v); (x_k, y_k), \sigma^2_{\text{det}}\mathbb{I}\right). \quad (6)$$

Separation constraint Ideally, the autoencoder training objective can automatically encourage the K landmarks to be distributed at *different* local regions so that the whole image can be reconstructed. However, the initial randomness can make the landmarks, defined as the mean coordinates weighted by $\mathbf{D}$ as in (3), all around the image center in the beginning of the training. This can lead to local optima from which the gradient descent may not escape (see Appendix F.2). To circumvent this difficulty, we introduce

an explicit loss to spatially separate the landmarks:

$$L_{\text{sep}} = \sum_{k \neq k'}^{1,\dots,K} \exp\left(-\frac{\|(x_{k'}, y_{k'}) - (x_k, y_k)\|_2^2}{2\sigma_{\text{sep}}^2}\right). \quad (7)$$

Equivariance constraint A landmark should locate a stable local pattern (with definite semantics). This requires landmarks to show equivariance to image transformations. More specifically, a landmark should move according to the transformation (e.g., camera and object motion) applied to the image if the corresponding visual semantics still exist in the transformed image. Let $g(\cdot, \cdot)$ be a coordinate transformation that map image $\mathbf{I}$ to $\mathbf{I}'(u, v) = \mathbf{I}(g(u, v))$, and $\boldsymbol{\ell}' = [x'_1, y'_1, \dots, x'_K, y'_K]^\top = \text{landmark}(\mathbf{I}')$. We ideally have $g(x'_k, y'_k) = (x_k, y_k)$, inducing the soft constraint

$$L_{\text{eqv}} = \sum_{k=1}^{K} \|g(x'_k, y'_k) - (x_k, y_k)\|_2^2, \quad (8)$$

This loss function is well-defined when g is known. Inspired by Thewlis et al. [59], we simulate g by a thin plate spline (TPS) [4] with random parameters. We use random translation, rotation, and scaling to determine the global affine component of the TPS; and, we spatially perturb a set of control points to determine the local TPS component. Besides the conventional way of selecting TPS control points at a predefined uniform grid (as used in [59]), we also take the landmarks detected by the current model as the control points to improve simulated transformation's focus on key image patterns. The two sets of control points are alternatively used in each optimization iteration (see Appendix F.3 for details). Moreover, when training sample appear in the form of video, we can also take the dense motion flow as g and the actual next frame as $\mathbf{I}'$.

Cross-object correspondence Our model does not explicitly ensure the semantic correspondence among the landmarks discovered on different object instances. The cross-object semantic stability of the landmarks mainly relies on the fact that visual patterns activating the same convolutional filter are likely to share semantic similarities.

3.3. Local latent descriptors

For simple images, like in MNIST [29] (see results for MNIST in Appendix B), multiple landmarks can be enough to describe the object shapes. For most natural images, however, landmarks are insufficient to represent all visual content, so extra latent representations are needed to encode complementary information. Though necessary, the latent representations should not encode too much holistic information that can overwhelm the image structures reflected by the landmarks. Otherwise, the autoencoder would not provide enough driving force to localize landmarks at meaningful locations. To achieve this trade-off, we attach a low-dimensional local descriptor to each landmark.

An hourglass-style neural network (see Appendix G.2) is introduced to obtain a feature map $\mathbf{F}$, which has the same size as the detection confidence map $\mathbf{D}$:

$$\mathbf{F} = \text{hourglass}_f(\mathbf{I}; \theta_f) \in \mathbb{R}^{W \times H \times S}. \quad (9)$$

Note that $\mathbf{F}$ is in a feature space shared among all landmarks and has S channels.

For each landmark, we use an average pooling weighted by a soft mask centered at the landmark to extract the local feature in the shared space. In particular, we take $\overline{\mathbf{D}}_k$, which is the Gaussian approximation of the detection confidence map defined in (6), as the soft mask. Then, a learnable linear operator is introduced for each landmarks to map the feature representation into a lower-dimensional individual space. Thus, the latent descriptor for the k-th landmark is

$$\mathbf{f}_k = \mathbf{W}_k \sum_{v=1}^{H} \sum_{u=1}^{W} \left(\overline{\mathbf{D}}_k(u, v) \cdot \mathbf{F}(u, v)\right) \in \mathbb{R}^C, \quad (10)$$

where $C < S$. The landmark-specific linear operator enables each landmark descriptor to encode a particular pattern in limited bits. We can also use (10) to extract a low-dimensional background descriptor. Since it is unreasonable to approximate the background confidence map with a Gaussian distribution, we exactly set $\overline{\mathbf{D}}_{K+1} = \mathbf{D}_{K+1}/\zeta_{K+1}$. Note that $\mathbf{f}_k$ is differentiable regarding both the feature map and the detection confidence map.

Putting all latent descriptors together, we have $\mathbf{f} = \text{vec}\left([\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_{K+1}] \in \mathbb{R}^{C \times (K+1)}\right)$. The left half of the red pathway in Figure 1 illustrates the neural network architecture to extract the landmark descriptors.

3.4. Landmark-based decoder

We approximately invert the landmark coordinates to the detection confidence map $\tilde{\mathbf{D}} \in \mathbb{R}^{W \times H \times (K+1)}$. Concretely, we use the probability density of an isotropic Gaussian distribution centered at each landmark to get raw score maps

$$\tilde{\mathbf{R}}_k(u, v) = \mathcal{N}\left((u, v); (x_k, y_k), \sigma_{\text{dec}}^2 \mathbb{I}\right), \tilde{\mathbf{R}}_{K+1} = \mathbf{1}. \quad (11)$$

and the background channel is set to 1. $\tilde{\mathbf{R}}$ is then normalized across channels to obtain the reconstructed detection confidence map

$$\tilde{\mathbf{D}}(u, v) = \tilde{\mathbf{R}}_k(u, v) / \sum_{k=1}^{K+1} \tilde{\mathbf{R}}_k(u, v). \quad (12)$$

Figure 1 (right half of the blue pathway) illustrates this.

For each landmark (including the background) descriptor $\mathbf{f}_k$, we transform it into a shared feature space by the landmark-specific operator $\tilde{\mathbf{W}}_k$ and an activation function (e.g., LeakyReLU [34]). Using $\tilde{\mathbf{D}}$ as the soft switches for global unpooling, we recover the feature map

$$\tilde{\mathbf{F}}(u, v) = \sum_{k=1}^{K+1} \tilde{\mathbf{D}}_k(u, v) \cdot \tau(\tilde{\mathbf{W}}_k \mathbf{f}_k) \in \mathbb{R}^{W \times H \times S}, \quad (13)$$

where $\tau(\cdot)$ is the non-linear activation function. This is illustrated by the right half of the red pathway in Figure 1.

Though alternative neural network architectures are available (e.g., in [46, 47]) for landmark-conditioned im-

age decoding, our proposed architecture enables back-propagation through the landmark coordinates. The Gaussian variance σ^2_{dec} determines how much the neighboring pixels can contribute to the gradients for the landmark coordinates and how sharp the descriptor is localized in the recovered feature map. While it is important to include more pixels for back-propagation in the early stage of training, sharpness becomes more important as training goes on. To balance the two needs, we obtain multiple versions of $\tilde{\mathbf{D}}, \tilde{\mathbf{F}}$ under different values of σ_{dec}, say, $(\tilde{\mathbf{D}}^1, \tilde{\mathbf{F}}^1), (\tilde{\mathbf{D}}^2, \tilde{\mathbf{F}}^2), \ldots, (\tilde{\mathbf{D}}^M, \tilde{\mathbf{F}}^M)$.

Let $[\![\cdots]\!]$ be the channel-wise concatenation. We use another hourglass-style network to reconstruct the image

$$\tilde{\mathbf{I}} = \text{hourglass}_d([\![\tilde{\mathbf{D}}^1, \tilde{\mathbf{F}}^1, \ldots, \tilde{\mathbf{D}}^M, \tilde{\mathbf{F}}^M]\!]; \theta_d) \quad (14)$$

The gray pathway in Figure 1 illustrates the image decoder.

3.5. Overall training objective

The image reconstruction loss L_{recon} drives the training of the entire autoencoder. We define L_{recon} as $\|\mathbf{I} - \tilde{\mathbf{I}}\|^2_F$, and $\mathbf{I}$ is normalized to $[0, 1]$. The full loss is $L_{\text{AE}} =$

$$\lambda_{\text{recon}} L_{\text{recon}} + \lambda_{\text{conc}} L_{\text{conc}} + \lambda_{\text{sep}} L_{\text{sep}} + \lambda_{\text{eqv}} L_{\text{eqv}}. \quad (15)$$

4. Experiments

We evaluate our method on a variety of datasets, including CelebA [31] and AFLW [35] for human faces, the cat head dataset [79], a car dataset built from PASCAL 3D [69], shoe images from UT Zappos50k [74], human pose images from Human3.6M [22, 8], MNIST (Appendix B), and animal images from AwA [28] (Appendix D).

Section 4.1 describes the datasets and shows the qualitative results of landmark discovery. In Section 4.2, we use the discovered landmarks to predict human-annotated landmarks, and we take the landmark detection accuracy as an indicator of the quality of discovered landmark. Section 4.3 demonstrates that our discovered landmarks can serve as effective image representations to predict shape-related facial attributes on CelebA. In Section 4.3, we show that our decoding module and the automatically discovered landmarks can be used to manipulate the object shapes.

4.1. Landmark discovery on multiple datasets

We train and evaluate landmark discovery models on a variety of objects. The detailed architectures of the neural network modules (i.e., $\text{hourglass}_{\ell|f|d}$) depend on the image sizes on different datasets. Appendix G describes implementation details, including data preprocessing, network architectures, model parameters, and optimization methods.

CelebA Following [59], we use all facial images in the CelebA training set excluding those also appearing in the MAFL the test set[1] (then 16,1962 images in total) to train models for landmark discovery. We use the MAFL testing set (1000 images) for all testing cases and reserve the MAFL training set (19,000 images) to train prediction models for manually-annotated landmarks. By default, we use the cropped and aligned images provided in the dataset.

[1]The MAFL dataset [81] is a subset of CelebA.

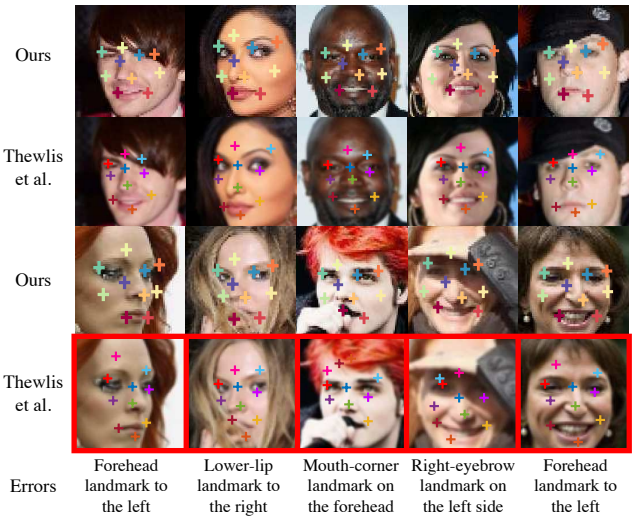

Figure 2: Discovering 10 landmarks on CelebA images. All figures for Thewlis et al. [59]'s come from their paper. The last row shows unsuccessful cases from [59] with error descriptions below.

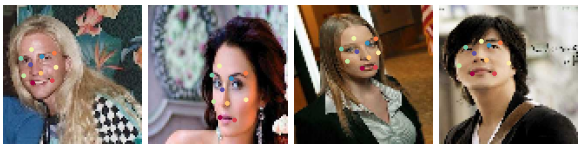

Figure 3: Discovering 10 landmarks on unaligned head-shoulder images using our model trained on aligned facial images.

Figure 4: Discovering 30 landmarks on unaligned CelebA images using our method.

As shown in Figure 2, our method can automatically discover facial landmarks at semantically meaningful and stable locations, such as the forehead center, eyes, eyebrows, nose, and mouth corners. Compared to Thewlis et al. [59]'s method, which results in a few significant errors, our method can locate landmarks more robustly against pose variations and occlusions. Interestingly, our method can work out-of-the-box on head-shoulder portraits without training on exactly the same type of images (Figure 3). Figure 4 shows that our method can also learn and detect a larger number (e.g., 30) of high-quality landmarks on unaligned facial images. Appendix E.1 shows more results.

AFLW Face images in AFLW are cropped differently from CelebA. The landmark discovery models (both ours and Thewlis et al. [59]'s) are pretrained on CelebA and fine-tuned on the AFLW training set (10,122 images) for adap-

Figure 5: Discovering landmarks on cat head images using our method. Top row: 10 landmarks; Bottom row: 20 landmarks.

Figure 6: Discovering 8 landmarks on shoes.

Figure 7: Discovering 10 landmarks on the profile images of cars.

tation. Sampled results on the AFLW testing set (2,991 images) are in Appendix E.2.

Cat heads Our model is trained on 7,747 cat head images and tested on 1,257 images. Compared to human faces, cat heads show more holistic appearance variations. As shown in Figure 5, our model can discover consistent landmarks (e.g., ears, nose, mouth) across different cat species and interestingly predict landmark locations under significant occlusion (the first image). Appendix E.3 shows more results.

Cars We build the profile-view car dataset by cropping the car images from the PASCAL 3D dataset. This dataset has a limited number of samples (567 images for training and 63 images for testing). As shown in Figure 7, our method can still learn meaningful landmarks (e.g., the windshield, driver-side door, wheels, rear) using a relatively small training set. Note that we transform the 3D annotations of the cars to 2D landmarks, so this dataset is ready for quantitative evaluation. Appendix E.4 shows more results.

Shoes We use the same setting as in [59] (49,525, training images and 500 testing images). As shown in Figure 6, landmarks are detected at semantically stable locations for different types of shoes. Appendix E.5 shows more results.

Human3.6M Human3.6M contains human activity videos in stable backgrounds. We use all 7 subjects in Human3.6M training set for our evaluation (6 for training and 1 for validation)[2]. We consider six activities (direction, discussion, posing, waiting, greeting, walking), in which human bodies are in the upright direction most of the time, resulting in 796,648 image frames for training and 87,975 image

[2]Training subject IDs: S1,S5,S6,S7,S8,S9; Validation subject IDs: S11.

Figure 8: Discovering 16 landmarks on Human3.6M dataset.

frames for testing. We removed the background using the off-the-shelf unsupervised background subtraction method provided in the dataset. The human bodies are cropped and roughly aligned regarding the foot location so that the excessive background regions are removed.

Compared to previously mentioned object types, human bodies have much more shape variations. As shown in Figure 8, our method can discover roughly consistent landmarks across a range of poses. In particular, the landmarks at the head, back, waist, and legs are stable across images. The landmarks at the arms are relatively less consistent across different poses, but they are still at semantically meaningful locations. Since the human body appearances in the frontal and back views are similar, we do not expect our discovered landmarks to distinguish the left and right sides of the human body, which means that a landmark at the left leg in the frontal view can locate the right leg in the back view. Since the training data is in the video format, optical flows are used as a short-term self-supervision for the eqvuivariance constraint in (8). Appendix C describes more details and results for Human3.6M experiments.

4.2. Prediction of ground truth landmarks

Unsupervised landmark learning is useful because of its potential to discover object structures that are coherent with the human's perception. We evaluate discovered landmarks' quality by predicting manually-annotated landmarks. Specifically, we use a linear model without a bias term to regress from the discovered landmarks to the human-annotated landmarks. Ground truth landmark annotations are needed to train this linear regressor. Thewlis et al. [59] extensively used random TPS to augment both discovered and labeled landmarks for training (on CelebA and ALFW). However, we do not use data augmentation for our method to minimize the complexity of training. Even in this case, our method shows stronger performance.

Stronger relevance to human-designed landmarks. In Table 1a, we regress the landmarks discovered using the models trained on the CelebA training set to the 5 anno-

# discovered landmarks	Regressor training set	Thewlis et al. [59]	Ours
10	CelebA	6.32	3.46
30	CelebA	5.76	**3.15**
50	CelebA	5.33	-
10	MAFL	7.95	3.46
30	MAFL	7.15	**3.16**
50	MAFL	6.67	-

(a) Comparison with unsupervised landmark learning methods on the MAFL testing set.

Method		MAFL	ALFW
Fully supervised	RCPR [5]	-	11.60
	CFAN [77]	15.84	10.94
	TCDCN [82]	7.95	7.65
	Cascaded CNN [55]	9.73	8.97
	RAR [70]	-	7.23
	MTCNN [81]	**5.39**	**6.90**
Discovery	Thewlis et al. [59] (50 landmarks)	6.67	10.53
	Ours (10 landmarks)	3.46	7.01
	Ours (30 landmarks)	**3.15**	**6.58**

(b) Comparison with supervised methods on the MAFL and ALFW testing sets.

Full L	w/o L_{recon}	w/o L_{conc}	w/o L_{sep}	w/o L_{eqv}
3.15	3.45	3.91	16.56	8.42

(c) Using ablative training losses of our method. Refer to (15) for each loss terms. Results are obtained on the MAFL testing set using 10 discovered landmarks.

Table 1: Mean errors of the annotated landmark prediction on human face datasets. Errors are in % regarding the biocular distance.

tated landmarks. The landmark labels in either the CelebA training set or the much smaller MAFL training set are used to train the regressor. Our method is not sensitive to the decreased size of the labeled training set. It outperforms Thewlis et al. [59]'s by 55% decrease of the landmark detection error. Notably, we achieve this with 30 discovered landmarks and a smaller labeled set while theirs uses 50 landmarks on a larger labeled set. Additionally, Table 2 demonstrates the consistent superiority of our method on the cat head dataset (7 target landmarks[3]), the car dataset (6 target landmarks), and Human3.6M[4] (32 target landmarks). Figure 9 illustrates the landmark regression results.

Competitive performance compared to fully supervised methods. Putting the landmark discovery model together with the linear regressor, we obtain a detector of human-designed landmarks. Unlike fully supervised methods, our model is trainable with a huge amount of unlabeled data, and the linear regressor can be trained using a relatively small amount of labeled data within a few minutes. Table 1b demonstrates that our model outperforms previous unsupervised methods and off-the-shelf pretrained fully-supervised models on the MAFL and AFLW testing sets. On AFLW,

[3]9 annotated landmarks in total. We do not use the 2 at the ears.

[4]See Appendix C for details

Figure 9: Prediction of annotated landmarks. Colorful cross: discovered landmark; Red dot: annotated landmark; Circle: regressed landmark, whose color represent its distance to the annotated landmarks. See the color bar for the distance (i.e., prediction error).

Dataset	Car		Cat head		Human3.6M
# discovered landmarks	10	24	10	20	16
Thewlis et al. [59]	11.42	11.11	26.76	26.94	7.51
Ours	**5.87**	**5.80**	**15.35**	**14.84**	**4.14**

Table 2: Mean errors of the annotated landmark prediction on the cat heads, cars, and human bodies. Errors are in % regarding the biocular distance, bi-wheel distance, and image size, respectively.

we take the 5 always-visible landmarks as the regression target. All models reported are either trained on the MAFL training set or publicly available.

Landmark detection with few labeled samples. Taking our model as a detector of manually annotated landmarks, we find that less than 200 samples are enough for our model to achieve less than 4% mean error on the MAFL testing set, which is better than the performance of TCDCN and MTCNN. Learning curves are provided in Appendix F.1.

Effectiveness of different loss terms. Our method combines several loss terms in the training objective (15). Table 1c shows that the removal of any term can cause performance drop of our model. In particular, the removal of the separation loss can devastate the model, and more detailed discussion about this loss term is in Appendix F.2. Our new differentiable formulation of the landmark validity constraints can already lead to a lower landmark detection error than Thewlis et al. [59]'s. Adding the reconstruction loss can further improve the accuracy.

4.3. Visual attribute recognition

Landmarks reflect object shapes. We use our discovered landmarks as a feature representation to recognize the shape-related binary facial attributes (13 labeled attributes are found) on CelebA. We still take the MAFL testing set for the quantitative evaluation. A linear SVM is trained for each attribute on the CelebA training set. We also compare our landmark coordinates with pretrained FaceNet [51] (InceptionV1) *top-layer* (128-dim) and top *conv-layer* (1792-dim) features for the attribute recognition task. As shown in Table 3, our discovered landmarks (60-dim) outperforms the FaceNet top-layer features for most attributes. The conv-layer features outperform our landmarks slightly but have a

Methods	Feature Dimension	Arched Eyebrows	Bags Under Eyes	Big Lips	Big Nose	Double Chin	High Cheek-bones	Male	Mouth Slightly Open	Narrow Eyes	Oval Face	Pointy Nose	Receding Hairline	Smiling	Average
Ours (discovered landmarks)	60	79.4	80.9	76.9	82.3	94.5	82.5	88.4	81.3	88.0	73.2	73.7	92.1	88.8	83.2
FaceNet [51] (top-layer)	128	76.4	80.3	76.8	80.4	**94.5**	72.6	82.7	74.4	87.9	72.7	73.1	92.2	76.2	80.0
FaceNet (top-layer) + Ours	188	**81.3**	**81.3**	**77.5**	**82.6**	**94.5**	**83.5**	**91.2**	**83.8**	**88.4**	**73.7**	**75.0**	**92.7**	**89.9**	**84.3**
FaceNet [51] (conv-layer)	1792	78.8	81.5	**77.4**	80.5	94.6	77.3	90.0	80.9	88.4	74.2	73.6	92.4	81.5	82.4
FaceNet (conv-layer) + Ours	1852	**80.1**	**81.8**	77.2	**82.3**	**94.7**	**82.1**	**90.8**	**85.0**	**88.6**	**74.5**	**73.6**	**92.4**	**90.5**	**84.1**

Table 3: Visual attribute recognition using pretrained FaceNet features and our discovered 30 landmarks on the MAFL test set.

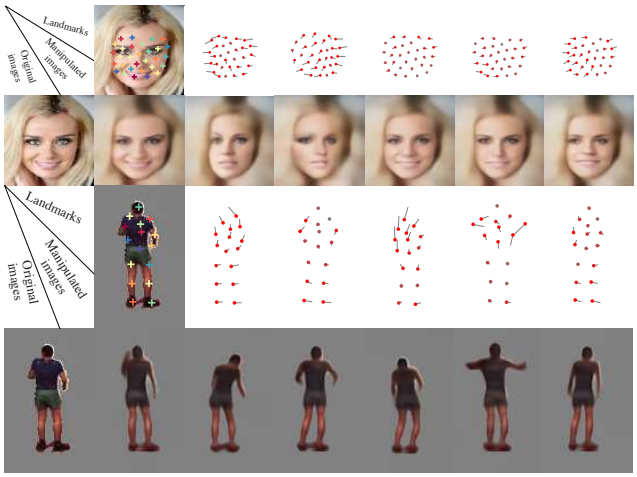

Figure 10: Image manipulation with our discovered landmarks and landmark-based decoder on the MAFL and Human3.6M testing set. 1st column: input images; 2nd column: discovered landmarks and reconstructed images; other columns: the red dots for new landmark locations, the gray lines for the synthetic adjustment of landmarks, and the images for the decoder outputs.

much higher dimension. Combining the landmark coordinates and the FaceNet features, higher accuracy is achieved. This suggests that the discovered landmarks are complementary to image features pretrained on classification tasks.

4.4. Image manipulation and generation

Our jointly trained image decoding module conditioned its outputs on the input landmarks and the their latent descriptors. If the two conditions are disentangled, we should be able to manipulate the object shape without changing other appearance factors by adjusting only the landmarks; or, *vice versa*. Note that landmark-based image morphing is not a new topic, and landmark-based hierarchical image decoding has also been explored recently [46, 64, 47]. However, these landmarks are all designed and annotated by humans. So far, little evidence has suggested that the automatically discovered landmarks are accurate and representative enough as a reliable condition for image generation.

In Figure 10, we synthesize flows to adjust the discovered landmarks of an input image. Fixing the landmark latent descriptors, we obtain realistic facial and human-body images whose shapes agree with the new landmarks. Other than the facial and body shape, then appearance factors of the input image are not visually changed. This result suggests that our image decoding module can synthesize realistic image using the landmarks learned without supervision,

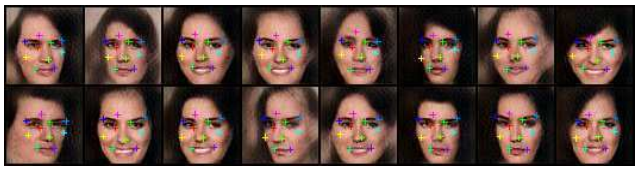

Figure 11: Face generation conditioned on discovered landmarks.

and it also suggests that our discovered landmarks have become an explicit representation disentangled from other factors of variations for image modeling. Implementation details and more results about unsupervised landmark-based face manipulation are available in Appendix A.

In Figure 11, instead of adjusting the landmark coordinates, we use the discovered landmarks of a reference image as the control signal to generate new facial images. Following the GAN framework [18], the latent representation of the generated image is randomly drawn from a prior distribution. As in Reed et al. [46], the landmark coordinates and latent representation are combined for image generation. We adopt BEGAN [3] for the discriminator and training objective. In addition, we apply a cyclic loss for the landmark coordinates, which encourages the same landmarks to be detected on the generated images as on the reference image. Our results provide additional evidence on the usefulness of the discovered landmarks for image modeling. Implementation details are in Appendix G.5.

5. Conclusion

We address the problem of unsupervised object landmark discovery and take it as an intermediate step of image representation learning. In particular, a fully differentiable neural network architecture is proposed for determining the landmark coordinates, together with soft contraints to enforce the validity of the detected landmarks. The discovered landmarks are visually meaningful and quantitatively more relevant to human-designed landmarks. In our framework, the discovered landmarks are an explicit part of the learned image representations. They are disentangled from the latent representations of the other appearance factors. The landmark-based explicit representations not only provide an interface for manipulating the image generation process but also appear to be complementary to pretrained deep-neural-network features for solving discriminative tasks.

Acknowledgements This work was supported in part by ONR N00014-13-1-0762, NSF CAREER IIS-1453651, and Sloan Research Fellowship.

References

[1] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *CVPR*, 2017. 1

[2] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. In *NIPS*, 2007. 2

[3] D. Berthelot, T. Schumm, and L. Metz. BEGAN: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017. 8, 35

[4] F. L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6):567–585, 1989. 2, 4

[5] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *ICCV*, 2013. 7

[6] M. C. Burl, M. Weber, and P. Perona. A probabilistic approach to object recognition using local photometry and global geometry. In *ECCV*, 1998. 2

[7] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190, 2014. 2

[8] C. S. Catalin Ionescu, Fuxin Li. Latent structured models for human pose estimation. In *ICCV*, 2011. 5

[9] M. Cho, S. Kwak, C. Schmid, and J. Ponce. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 2

[10] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001. 2

[11] D. Cristinacce and T. Cootes. Automatic feature localisation with constrained local models. *Pattern Recognition*, 41(10): 3054–3067, 2008. 2

[12] M. Dantone, J. Gall, G. Fanelli, and L. Van Gool. Real-time facial feature detection using conditional regression forests. In *CVPR*, 2012. 2

[13] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NIPS*, 2014. 1

[14] H. Fan, H. Su, and L. Guibas. A point set generation network for 3D object reconstruction from a single image. In *CVPR*, 2017. 1

[15] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010. 2

[16] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, 2003. 2

[17] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Region-based convolutional networks for accurate object detection and segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):142–158, Jan 2016. 1

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014. 1, 2, 8

[19] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(5): 855–868, 2009. 2

[20] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[21] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *ICCV*, 2017. 1, 2

[22] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014. 5

[23] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *NIPS*, 2015. 2

[24] A. Kanazawa, D. W. Jacobs, and M. Chandraker. WarpNet: Weakly supervised matching for single-view reconstruction. In *CVPR*, 2016. 2

[25] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *ICLR*, 2014. 1, 2

[26] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012. 1

[27] M. Lam, B. Mahasseni, and S. Todorovic. Fine-grained recognition as hsnet search for informative image parts. In *CVPR*, 2017. 2

[28] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 5, 17

[29] Y. LeCun. The MNIST database of handwritten digits. `http://yann.lecun.com/exdb/mnist/`, 1998. 4

[30] K. Lenc and A. Vedaldi. Learning covariant feature detectors. In *ECCV Workshop on Geometry Meets Deep Learning*, 2016. 2

[31] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 5

[32] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1

[33] J. Lv, X. Shao, J. Xing, C. Cheng, and X. Zhou. A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In *CVPR*, 2017. 2

[34] A. Maas, A. Hannun, and A. Ng. Rectifier nonlinearities improve neural network acoustic models. In *ICML*, 2013. 4, 34

[35] P. R. Martin Koestinger, Paul Wohlhart and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011. 5

[36] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In *International Conference on Artificial Neural Networks*, 2011. 2

[37] M. F. Mathieu, J. J. Zhao, J. Zhao, A. Ramesh, P. Sprechmann, and Y. LeCun. Disentangling factors of variation in deep representation using adversarial training. In *NIPS*, 2016. 2

[38] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004. 2

[39] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 1, 2, 3, 13, 14

[40] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *ICCV*, 2015. 1

[41] D. Novotny, D. Larlus, and A. Vedaldi. AnchorNet: A weakly supervised network to learn geometry-sensitive features for semantic matching. 2017. 2

[42] M. Pedersoli, T. Tuytelaars, and L. Van Gool. Using a deformation field model for localizing faces and facial points under weak supervision. In *CVPR*, 2014. 2

[43] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016. 1, 2

[44] S. Reed, K. Sohn, Y. Zhang, and H. Lee. Learning to disentangle factors of variation with manifold interaction. In *ICML*, 2014. 2

[45] S. Reed, Y. Zhang, Y. Zhang, and H. Lee. Deep visual analogy-making. In *NIPS*, December 2015. 2

[46] S. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee. Learning what and where to draw. In *NIPS*, 2016. 2, 4, 8, 35

[47] S. Reed, A. v. d. Oord, N. Kalchbrenner, S. G. Colmenarejo, Z. Wang, D. Belov, and N. de Freitas. Parallel multiscale autoregressive density estimation. In *ICML*, 2017. 2, 4, 8

[48] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 FPS via regressing local binary features. In *CVPR*, 2014. 2

[49] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 1

[50] I. Rocco, R. Arandjelović, and J. Sivic. Convolutional neural network architecture for geometric matching. In *CVPR*, 2017. 2

[51] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 7, 8

[52] R. Sicre, Y. Avrithis, E. Kijak, and F. Jurie. Unsupervised part learning for visual recognition. In *CVPR*, 2017. 2

[53] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 1

[54] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012. 2

[55] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *CVPR*, 2013. 2, 7

[56] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, et al. Going deeper with convolutions. In *CVPR*, 2015. 1

[57] K. Tang, A. Joulin, L.-J. Li, and L. Fei-Fei. Co-localization in real-world images. In *CVPR*, 2014. 2

[58] J. Thewlis, H. Bilen, and A. Vedaldi. Unsupervised object learning from dense equivariant image labelling. In *NIPS*, 2017. 2

[59] J. Thewlis, H. Bilen, and A. Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *ICCV*, 2017. 1, 2, 4, 5, 6, 7, 13, 14, 15, 16, 19, 21, 27, 29

[60] A. Toshev and C. Szegedy. DeepPose: Human pose estimation via deep neural networks. In *CVPR*, 2014. 2

[61] M. Valstar, B. Martinez, X. Binefa, and M. Pantic. Facial point detection using boosted regression and graph models. In *CVPR*, 2010. 2

[62] A. van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, et al. Conditional image generation with pixelcnn decoders. In *NIPS*, 2016. 2

[63] A. Van Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks. In *ICML*, 2016. 2

[64] R. Villegas, J. Yang, Y. Zou, S. Sohn, X. Lin, and H. Lee. Learning to generate long-term future via hierarchical prediction. In *ICML*, 2017. 8

[65] X. Wang and A. Gupta. Generative image modeling using style and structure adversarial networks. In *ECCV*, 2016. 2

[66] M. Weber, M. Welling, and P. Perona. Towards automatic discovery of object categories. In *CVPR*, 2000. 2

[67] J. Wu, T. Xue, J. J. Lim, Y. Tian, J. B. Tenenbaum, A. Torralba, and W. T. Freeman. Single image 3D interpreter network. In *ECCV*, 2016. 1, 2

[68] Y. Wu, C. Gou, and Q. Ji. Simultaneous facial landmark detection, pose and deformation estimation under facial occlusion. In *CVPR*, 2017. 2

[69] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond PASCAL: A benchmark for 3D object detection in the wild. In *WACV*, 2014. 5

[70] S. Xiao, J. Feng, J. Xing, H. Lai, S. Yan, and A. Kassim. Robust facial landmark detection via recurrent attentive-refinement networks. In *ECCV*, 2016. 2, 7

[71] S. Xiao, J. Feng, L. Liu, X. Nie, W. Wang, S. Yan, and A. Kassim. Recurrent 3d-2d dual learning for large-pose facial landmark detection. In *ICCV*, 2017. 2

[72] W. Yang, W. Ouyang, H. Li, and X. Wang. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In *CVPR*, pages 3073–3082, 2016. 2

[73] Y. Yang. *Articulated Human Pose Estimation with Flexible Mixtures of Parts*. PhD thesis, University of California, Irvine, 2013. 2

[74] A. Yu and K. Grauman. Fine-grained visual comparisons with local learning. In *CVPR*, 2014. 5

[75] X. Yu, F. Zhou, and M. Chandraker. Deep deformation network for object landmark localization. In *ECCV*, 2016. 2

[76] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014. 1

[77] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment. In *ECCV*, 2014. 2, 7

[78] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 2016. 2

[79] W. Zhang, J. Sun, and X. Tang. Cat head detection - how to effectively exploit shape and texture features. *ECCV*, 2008. 5

[80] Y. Zhang, K. Sohn, R. Villegas, G. Pan, and H. Lee. Improving object detection with deep convolutional networks via Bayesian optimization and structured prediction. In *CVPR*, 2015. 1

[81] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*, pages 94–108. Springer, 2014. 2, 5, 7

[82] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Learning deep representation for face alignment with auxiliary attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(5):918–930, 2016. 2, 7

[83] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012. 2