AutoLink: Self-supervised Learning of Human Skeletons and Object Outlines by Linking Keypoints

Anonymous Author(s)
Affiliation
Address
email

Abstract

Structured representations such as keypoints are widely used in pose transfer, conditional image generation, animation, and 3D reconstruction. However, their supervised learning requires expensive annotation for each target domain. We propose a self-supervised method that learns to disentangle object structure from the appearance with a graph of 2D keypoints linked by straight edges. Both the keypoint location and their pairwise edge weights are learned, given only a collection of images depicting the same object class. The graph is interpretable, for example, AutoLink recovers the human skeleton topology when applied to images showing people. Our key ingredients are i) an encoder that predicts keypoint locations in an input image, ii) a shared graph as a latent variable that links the same pairs of keypoints in every image, iii) an intermediate edge map that combines the latent graph edge weights and keypoint locations in a soft, differentiable manner, and iv) an inpainting objective on randomly masked images. Although simpler, AutoLink outperforms existing self-supervised methods on the established keypoint and pose estimation benchmarks and paves the way for structure-conditioned generative models on more diverse datasets.

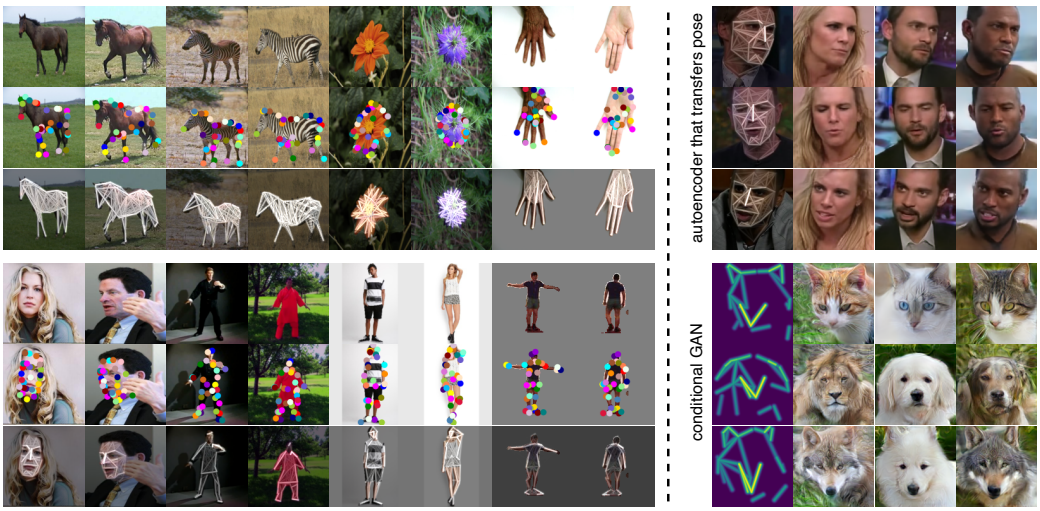
1 Introduction

Object structure representations are widely used in modern computer graphics and computer vision techniques, including keypoints for image generation [60, 61, 89] and skeletons for 3D reconstruction [22, 37, 111, 77, 91]. However, the structure is usually supervised on large annotated datasets [54, 3, 2, 64] or via hand-crafted parametric models [58, 76, 102, 73, 4, 52]. Neither approach generalizes well to new domains and both require additional manual annotation whenever more detail is needed [81].

Our goal is to reconstruct the keypoint locations of an object by learning from an unlabelled image collection, thereby sidestepping the generalization problem. Our key idea is to leverage that the same object shares the same topology by introducing an explicit graph that links the same pairs of keypoints in all instances. By contrast, existing self-supervised keypoint learning methods model objects as a set of independent parts. Consistency over different instances of the same object is encouraged by either enforcing parts to follow hand-crafted image transformations [93, 115, 38, 59, 33, 55] or by adding implicit bias in the network architecture that encodes such spatial equivariances [31, 30]. Only [39, 85] use an explicit skeleton representation, but both require predefined topology, rely on video input, and [39] is trained in a CycleGAN setting that still requires manually labeled examples.

We propose a simple yet effective method to learn both the keypoints and their links without supervision in terms of a sparse graph serving two purposes. First, the graph acts as a bottleneck that

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.



(a) detected keypoints and visualized graph representation (b) applications

Figure 1: **Teaser.** (a) AutoLink applies to diverse collections of images and automatically yields keypoints linked to a graph without ground truth. It recovers animal and human pose and object shape in settings where previous methods struggle, including cluttered background, structured stripe textures, articulated fingers, and detailed faces. (b) Example applications are conditional image generation with autoencoders (top) and GANs (bottom) that are driven by the learned keypoints.

can only store structural information disentangled from appearance. Second, it forms a constraint that associates observations across training images. We enforce the same topology across instances of the same class by learning a shared graph with a single set of edge weights. In the absence of labels, we train AutoLink with only an autoencoder reconstruction objective. Since the graph bottleneck should not model appearance, we additionally feed the decoder with the input image after masking the majority of its pixels. In turn, it is important for the disentangling that the heavily masked image models appearance without leaking structural information. This is the case as inpainting methods are unable to infer the original image precisely from only sparse pixel colors [117, 112] unless conditioning on structure representations, such as edge maps [100, 68, 51, 49, 27]. Therefore, by forcing the autoencoder to reconstruct the original image, the detector converges to generate representative structures of images.

We demonstrate on 4 benchmarks that the trained detector has a significantly improved keypoint localization accuracy and on 6 additional datasets that it applies to a broader set of images spanning portraits, persons, animals, hands and flowers, which we attribute to the explicit modeling of links in the graph. Figure 1 shows the diverse set of image domains it applies to, including challenging textures and uncontrolled background, how both skeleton representations as well as object outlines are learned by varying the number of keypoints, and exemplifies applications to controlled image generation. The code and the pre-trained models will be released upon publication.

Ethics - Risks. The estimated keypoints and edges could be abused for deep fakes as the driving signal for generative models or for unwanted surveillance applications. However, our method works towards improved generality, including objects and animals, and does not improve upon supervised models that already exist in high detail for humans. **Benefits.** Since our method is entirely self-supervised, it can be applied to a diverse set of persons, objects, animals, or situations that have not yet been labeled.

2 Related Work

Most representation learning methods focus on generic feature vectors for entire images to initialize deep networks for improved object classification [98, 10, 72, 29, 28]. By contrast, our method introduces explicit object structure and we review the most related approaches in the following.

Self-supervised Keypoints Detection. The most common idea to discover keypoints in an unsupervised manner is to rely on the notion that keypoints move as the image changes. Various losses have been designed to enforce that the keypoints follow a known transformation. For instance, the transformation can be the view change in a multi-view recording [92, 80, 79] or the natural motion in videos [19, 86, 48, 67, 16, 46, 39]. When only single images are available, artificial image deformation is applied, either from randomized [93, 115, 38, 59] or learned [97, 106] transformations within a predefined deformation space that is usually tuned for each dataset. However, the learned keypoints may model the background [115, 86] and struggle with large pose variation [33] as image deformations do not separate foreground from background and are bound. Most models leverage multi-branch network architectures to encode the structure and appearance separately and utilize multiple losses that need to be balanced. By contrast, we do not apply artificial transformations, use a single branch, and a single loss which eases and stabilizes training. To overcome the need for artificial image deformation, He et al. [30, 31] exploit GANs to generate images along with corresponding keypoints and later use them to train a detector. However, this leads to even more complex network architectures and comes with instabilities in GAN training, limiting their applicability to complex objects like human bodies.

Skeleton Representations. Bone maps representing the keypoints connectivity are, for example, used in the supervised bottom-up approaches for human pose estimation. They learn an *affinity field* from manual labels that connects joints with bone maps [8] or via explicit offsets [75]. We use a similar edge map representation but learn both the location and linking from scratch without annotations. In the weakly-supervised setting, Jakab et al. [39] exploit CycleGAN [118] to translate between image and edge maps. The graph connectivity is predefined to the human skeleton and edges are supervised by a large dataset of unpaired ground truth object edges, which can come from a different dataset but are manually annotated. Schmidtke et al. [85] overcome the manual labeling by deforming a template skeleton. However, they both require the known connectivity of the keypoints and require videos for training while ours learns both the keypoints and connectivity from a collection of single images. Noguchi et al. [71] heuristically generate the skeleton by linking the centers of the learned Signed Distance Field [63] for parts. However, they require videos without the background of the same object, and the learned skeleton does not generalize to other objects of the same class.

Object Sketch Learning. Sketches are made of strokes drawn by a pen. It is a concise and abstract representation, which can be used in object recognition [110, 105] and image retrieval [94, 109, 104]. There are two common sketch representations used in neural models [103]. The first one is a black-white raster image [95, 84, 108], often used for image-to-image translation [36, 118, 74, 43]. Alternatively, a sketch can also be represented as a sequence of points (pen coordinates) [26, 20, 83], which is usually used by recurrent generation models [26, 11, 7, 42, 15, 24]. This graph representation is similar to ours. However, instead of learning to mimic human drawings, ours directly predicts both the keypoints and their connectivity on real natural images.

Structure-enhanced Image Inpainting. When key parts are missing in an image, e.g., eyes on faces or arms of humans, it is hard for inpainting networks to imagine the content accurately from scratch. Recently researchers try to use structural cues to guide the networks. These methods usually have two stages, where structure cues are detected in the first stage, and images are generated based on the structural cues in the second stage. The structural cues can be edges [68, 51, 40, 101, 6], segmentation masks [116, 53, 27, 90], landmarks [49, 113, 107], low-frequency subband [96, 78], and foreground contours [100]. These structural cues are detected from the masked images and supervised by the ground truth calculated from the original images. Our reconstruction objective can be seen as such two-stage inpainting, but self-supervised and with the image edges replaced with the learned graph edge representation.

3 Method

We leverage that the objects in the dataset share the same topology and can be represented as a graph that connects keypoints by a set of edges. To learn the keypoints and edges, we design an autoencoder that aims to accurately reconstruct the input image, with the graph as the intermediate

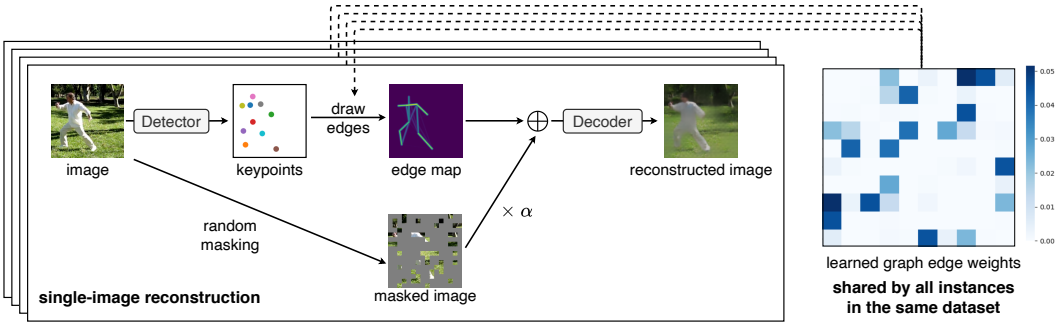


Figure 2: **Overview.** Given an image, we detect keypoints and draw differentiable edges between keypoints according to the learned graph edge weights that is visualized as a color matrix. The method is self-supervised in that the latent edge map and keypoints are learned by reconstructing the masked input images. Note that keypoints are image specific and edge maps are shared.

representation. To encode the input image into a graph, we detect the keypoints and create an edge heatmap based on learnable edge weights. To mostly obtain appearance information, we mask out the majority of the image, which reduces the structure information to a very low level. The edge heatmap is combined with the masked image to reconstruct the original image. Since the missing structure is important to reconstruct the original image, the network is forced to learn the structure of the object in a self-supervised manner. Figure 2 shows an overview of our method.

Formally, given an image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ with height H and width W we aim to learn a set of keypoints $\{\mathbf{k}_i\}_{i=1}^{K}$, where $\mathbf{k}_i \in [-1, 1] \times [-1, 1] \subset \mathbb{R}^2$ is the normalized keypoint coordinate, and K is the number of keypoints. We use a ResNet with upsampling [99] to detect keypoints. Afterward, we draw a differentiable edge [66] between each pair of keypoints. The edge map $\mathbf{S} \in \mathbb{R}^{H \times W}$ is concatenated along the channel dimension with the randomly masked image $\mathbf{I}_m \in \mathbb{R}^{H \times W \times 3}$ and fed into a UNet [82] to obtain the reconstructed image $\mathbf{I}'$. The detailed network architectures can be found in Appendix D.

3.1 Image Structure Representation

In this section, we introduce the generation of keypoints and the edge map from the image. Let $\mathbf{H} \in \mathbb{R}^{H \times W \times K}$ be the K heatmaps generated by a ResNet with upsampling [99] from the image $\mathbf{I}$. The keypoint $\mathbf{k}_i$ is calculated by the differentiable soft-argmax function,

$$\mathbf{k}_i = \sum_{\mathbf{p}} \frac{\exp(\mathbf{H}(\mathbf{p}))}{\sum_{\mathbf{p}} \exp \mathbf{H}(\mathbf{p}))} \mathbf{p}, \tag{1}$$

where $\mathbf{p} \in [-1, 1] \times [-1, 1]$ is the normalized pixel coordinates.

Given two keypoints $\mathbf{k}_i, \mathbf{k}_j$, a differentiable edge map [66] $\mathbf{S}_{ij}$ is defined as

$$\mathbf{S}_{ij}(\mathbf{p}) = \exp\left(d_{ij}^2(\mathbf{p})/\sigma^2\right), \tag{2}$$

where σ is a hyperparameter controlling the thickness of the edge, and $d_{ij}(\mathbf{p})$ is the L_2 distance between the pixel $\mathbf{p}$ and the edge drawn by keypoints $\mathbf{k}_i$ and $\mathbf{k}_j$,

$$\mathbf{d}_{ij}(\mathbf{p}) \begin{cases} \|\mathbf{p} - \mathbf{k}_i\|_2 & \text{if } t \leq 0, \\ \|\mathbf{p} - (\mathbf{k}_i + t\mathbf{k}_j)\|_2 & \text{if } 0 < t < 1, \\ \|\mathbf{p} - \mathbf{k}_j\|_2 & \text{if } t \geq 1, \end{cases} \quad \text{where} \quad t = \frac{(\mathbf{p} - \mathbf{k}_i) \cdot (\mathbf{k}_j - \mathbf{k}_i)}{\|\mathbf{k}_i - \mathbf{k}_j\|_2^2}. \tag{3}$$

We assign a weight $w_{ij} > 0$ to each edge, which is enforced to be positive by SoftPlus [18]. This weight is learned during training and shared across all object instances in a dataset. Finally, we take the maximum at each pixel of the heatmaps to obtain the final edge map $\mathbf{S} \in \mathbb{R}^{H \times W}$,

$$\mathbf{S}(\mathbf{p}) = \max_{ij} w_{ij} \mathbf{S}_{ij}(\mathbf{p}). \tag{4}$$

Taking the maximum at each pixel avoids the entanglement of the edge weights and the convolution kernel weights, which is further explained in Section 4.4.

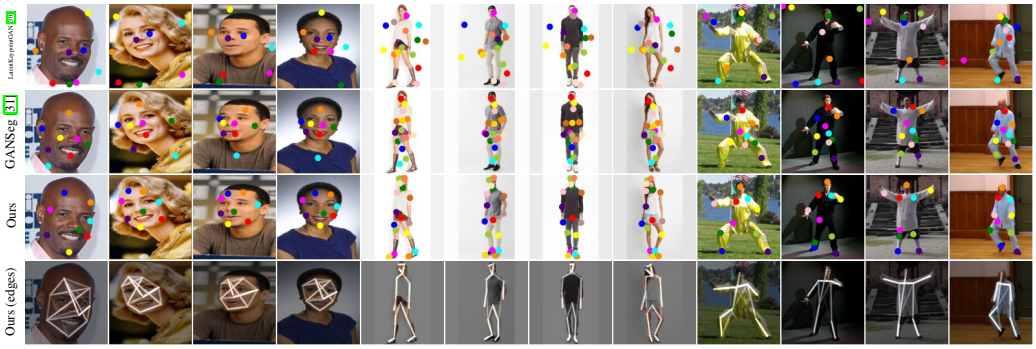


Figure 3: **Qualitative comparison on detected keypoints**. Our model is more robust on wild face poses, and depicts more details on human bodies compared to [30] and [31]. For example, the feet poses in the middle four images are clearly detected.

3.2 Image Reconstruction

The masked image $\mathbf{I}_m$ is generated by first uniformly dividing the image $\mathbf{I}$ into a 16×16 grid, and randomly masking out 80% of the grid cells, similar to [28]. We concatenate the masked image with the edge map and feed it into a UNet decoder [82] to reconstruct the original image,

$$\mathbf{I}' = \text{Decoder}(\alpha \mathbf{I}_m \oplus \mathbf{S}) \tag{5}$$

where $\oplus$ means concatenation along the channel dimension, and α is a learnable parameter that compensates for the change of the edge weight magnitude during training. α is initialized to 1. We found this parameter to be helpful in training stability. Different to [39], we have no ground truth for the edge map. Our edge map is an unobserved latent variable. Thus we only minimize the difference of the original image and the reconstructed image by the perceptual loss [41],

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \|\Gamma(I_i) - \Gamma(I_i')\|_2^2 \tag{6}$$

where N is the number of examples and Γ is the feature extractor. The perceptual loss is believed to measure the structure similarity [41, 23, 17], and leads to more robust training [38, 39].

3.3 Implementation Details

We use the Adam optimizer [47] with a learning rate of 10^{-4} with $\beta_1 = 0.9$, $\beta_2 = 0.99$. The batch size is 64. We train for 20k iterations. It takes 3 hours to train on a single V100 GPU. All images are resized to 128×128. The learning rate for the edge weights is multiplied by 512 due to the small gradient of SoftPlus [18] when the value is close to 0. To show the robustness of our model, we report all experiments on the sampling strategy of masking 80% of the 16×16 patches. We perform experiments with the same edge thickness of $\sigma^2 = 5e-5$ for all benchmark datasets. We train for 10 times and report the mean and the standard deviation of the evaluation metrics. Although it already outperforms other work in most experiments, we also tune thicknesses to each individual dataset, as others did for their hyperparameters, which further improve the results. The tuned thicknesses can be found in Appendix B. The only other hyperparameter is the number of keypoints, which we range from 4 to 32 and set to that of the established benchmarks for quantitative comparisons.

4 Experiments

In this section, we compare our results to the related methods, showing that our model is simple yet effective. Besides, we perform a number of ablation studies on hyperparameters and algorithm variants, exhibiting the robustness of our model and justifying the necessity of every model component.

Table 1: **Landmark detection on CelebA**. The metric is the landmark regression (without bias) error in terms of L_2 distance normalized by inter-ocular distance (lower is better). While all methods perform well on aligned CelebA, ours is more robust on Wild CelebA. The sign $\star$ means being reported by [33] and † means being reported by [55].

Method	Type	Aligned (K=10)	Wild (K=4)	Wild (K=8)
DEF [14] by [33]	Part Segmentation	-	-	31.30% $\star$
SCOPS [33] (w/o saliency)	Part Segmentation	-	46.62%	22.11%
SCOPS [33] (w/ saliency)	Part Segmentation	-	21.76%	15.01%
Liu et al. [55]	Part Segmentation	-	15.39%	12.26%
Huang et al. [32] (w/ detailed label)	Part Segmentation	-	-	8.40%
GANSeg [31]	Part Segmentation	3.98%	12.26%	6.18%
Thewlis et al. [93]	Landmark	7.95%	-	31.30% $\star$
Zhang et al. [115]	Landmark	3.46%	-	40.82% $\star$
LatentKeypointGAN [30]	Landmark	5.85%	25.81%	21.90%
Lorenz et al. [59]	Landmark	3.24%	15.49% †	11.41% †
IMM [38]	Landmark	**3.19**%	19.42% †	8.74% †
LatentKeypointGAN-tuned [30]	Landmark	3.31%	12.10%	5.63%
Ours (general)	Landmark	3.92±0.69%	**7.72**±0.47%	5.66±0.29%
Ours (thickness-tuned)	Landmark	3.54%	**6.11**%	**5.24**%

Table 2: **Landmark detection on Human Body**. Our model outperforms all the other baselines. The metric for each dataset follows the corresponding description in the text. The sign † means being reported by [88] and the sign $\star$ means being reported by [5]. The number of keypoints are $K = 16$ for Human3.6m and DeepFashion and $K = 10$ for Taichi.

Method	Supervised by	Human3.6m ↓	DeepFashion ↑	Taichi ↓
DEF [14]	testing dataset	-	-	494.48 †
SCOPS [33]	saliency maps	-	-	411.38 †
Siarohin et al. [88]	videos	-	-	389.78
Zhang et al. [114]	videos	-	-	343.67
Zhang et al. [115]	videos	4.14	-	-
Schmidtke et al. [85]	video & T-pose template	3.31	-	-
Jakab et al. [39]	video & unpaired ground truth	2.73	-	-
Newell et al. [39]	paired ground truth	**2.16**	-	-
Thewlis et al. [93]	n/a	7.51	-	-
Zhang et al. [115]	n/a	4.91	-	-
LatentKeypointGAN [30]	n/a	-	49%	437.69
Lorenz et al. [59]	n/a	2.79	57% $\star$	-
GANSeg [31]	n/a	-	59%	417.17
Ours (general)	n/a	2.81±0.07	**65**±1.2%	**337.50**±25.08
Ours (thickness-tuned)	n/a	**2.76**	**66**%	**316.10**

4.1 Datasets and Evaluation Metrics

CelebA-aligned [57] contains 200k celebrity faces aligned in center. We follow [93] splitting it into three subsets: CelebA training set without MAFL (160k images), MAFL training set (19k), MAFL test set (1k). We train our network on the CelebA training set without MAFL. To quantitatively evaluate the consistency of our predicted keypoints, we follow [93] training a linear regression without bias from our detected keypoints to the ground truth keypoints on the MAFL training set and reporting the mean L_2 error normalized by inter-ocular distance on the MAFL test set.

CelebA-in-the-wild [57] contains celebrity faces in unconstrained conditions. We follow [33] and first split it into three subsets as for CelebA-aligned, and then remove the images where a face covers less than 30% of the area, which results in 45,609 images for model training, 5,379 with keypoint labels for regression, and 283 for testing. The evaluation metric is the same as CelebA-aligned.

Human3.6m [35] contains human activity videos in static backgrounds. We follow [115] considering six activities (direction, discussion, posing, waiting, greeting, walking), and using subjects 1, 5, 6, 7, 8, 9 for training and 11 for testing. This results in 796,648 images for training and 87,975 images for testing. The evaluation metric is the regressed (without bias) mean L_2 error normalized by the image size. We remove the background as in [115, 59] to make a fair comparison to others. To underline the robustness against structured backgrounds we also report the numbers including background.

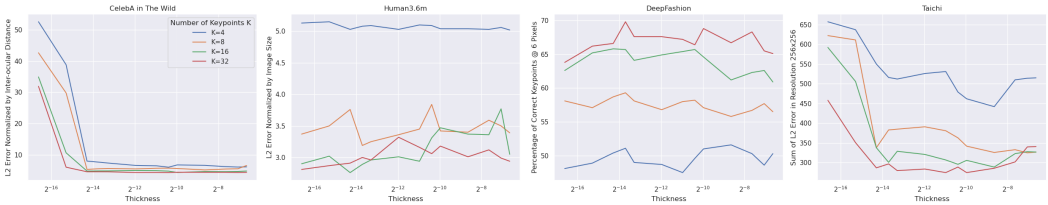


Figure 4: **Ablation tests on number of keypoints and edge thickness.** While the model shows better performance with more keypoints, it is robust to the edge thickness.

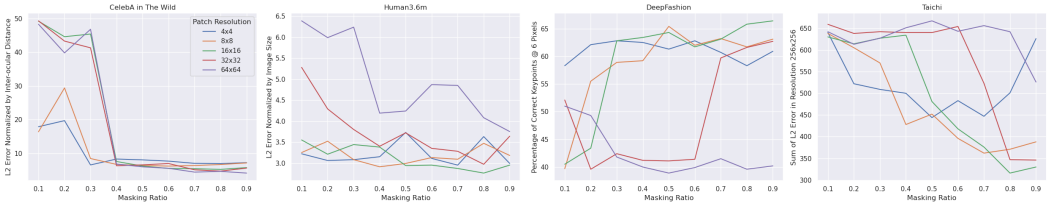


Figure 5: **Ablation tests on masking strategy.** Overall, the performance increases as the mask ratio increases. Too small or too large patch size can decrease the performance. Empirically, masking 80% of the 16×16 patches is a golden rule.

DeepFashion [56] contains 53k in-shop clothes images. We follow [59] only keeping the full-body images. We use 10604 images for training and 1179 images for testing as in [5]. We use the keypoints generated by AlphaPose [21] as the ground truth. The evaluation metric is Percentage of Correct Keypoints of d=6 pixels in resolution 256×256.

Taichi [87] contains 3049 training videos and 285 test videos of people performing Tai-Chi, with various appearance of foreground and background. We follow [88] using 5000 and 300 images (not contained in training data) for training a linear regression and for testing, respectively. The evaluation metric, mean average error (MAE), is calculated as the sum of the L_2 error in resolution 256×256.

Flower [70], **11k Hands** [1], **Horses** [118], and **Zebras** [118] are used for qualitative experiments. **VoxCeleb2** [13] and **AFHQ** [12] are used for pose transfer and conditional image generation, respectively. Horses and Zebras are extracted from the CycleGAN dataset [118] by removing the images with multiple horses and aligning them to face left. The train/test split of Flower follows [9]. All the other datasets follow their official train/test split.

4.2 Qualitative Analysis

We qualitatively compare our detected keypoints with other methods and show the examples of the learned edges in Figure 3. For visualization purposes, we scale the edge weights to obtain visible edges. We use the same number of keypoints as the previous method [31] for a fair comparison, which are 8 for CelebA-in-the-Wild, 16 for DeepFashion, and 10 for Taichi. We will discuss more on the choice of the number of keypoints in Ablation Study 4.4. As shown in Figure 3, our model not only detects consistent keypoints but also learns reasonable edges, such as human skeletons in DeepFashion and Taichi. For example, the feet are clearly connected with the corresponding knees, and there is no edge between the left and right hands. We show 105 images with detected keypoints and visualized graph structure for each dataset in Figure 9-21 in the Appendix, demonstrating that our model works on various classes of objects of diverse appearance and complex backgrounds.

4.3 Quantitative Analysis

We compare the keypoint detection results with other methods in Table 1 (CelebA), and Table 2 (Human3.6, DeepFashion, Taichi). Our simple model outperforms other methods in all benchmarks,

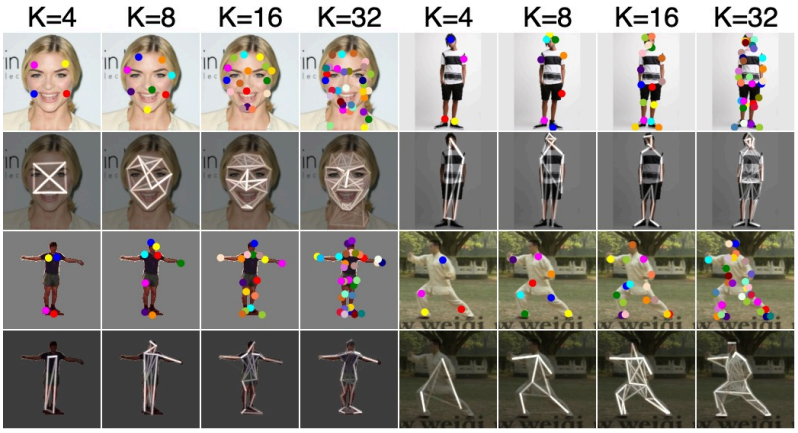


Figure 6: **Examples of different numbers of detected keypoints.** With very few keypoints, the model only models a very basic shape, such as the box on the face, and sometimes cannot fully capture the structure. For example, with K=4, only legs are modeled on humans. With abundant keypoints, it is able to model details.

Table 3: **Ablation tests on variants of edge heatmap generation**. The original design is proved to be the most robust one. Although in some cases it is not the optimal, the difference is almost trivial.

Model	CelebA in The Wild ↓	Human3.6m ↓	DeepFashion ↑	Taichi ↓
original model	**5.24**%	**2.76**%	65.8%	316
fixed α	6.39%	2.87%	**66.0**%	374
shared learnable thickness	6.12%	3.25%	49.1%	425
independent learnable thickness	5.94%	3.73%	50.2%	**311**
edge-specific heatmap	5.65%	3.83%	65.1%	407
only using keypoints without edges	6.55%	3.58%	52.9%	722

except for the most constrained setting CelebA-aligned where all methods perform well and the results are comparable. The results on CelebA in Table 1 confirms that our model is more robust to poses in the wild. Since self-supervised part segmentation methods are usually more robust on wild faces [55, 33, 31], we also include them for comparison, demonstrating the robustness of our model over existing baselines. The results on Human3.6m and DeepFashion in Table 2 show the capability of our model to detect keypoints on human bodies of either similar or diverse appearances. The results on Taichi in Table 2 demonstrates the general applicability of our model to human bodies of diverse pose in complex backgrounds.

4.4 Ablation Tests

In this section, we analyze the hyperparameters and demonstrate the robustness of our model. We also discuss the possible variants of our model and show the superiority of our design.

Number of Keypoints & Thickness. We show ablation test results on the different numbers of keypoints and edge thicknesses in Figure 4. The exact numbers can be found in Table 4 in Appendix B. Our model shows very strong robustness to edge thickness. The accuracy remains state-of-the-art while the thickness of the edges varies by multiple orders of magnitude. On the other hand, with the increasing number of keypoints the accuracy increases. This is expected since more keypoints are able to capture structure in more detail, as shown in Figure 4.4.

Masking Strategy. In our standard setting, the image is divided into 16×16 patches and 80% of the patches are randomly masked. We investigate how the patch size and masking ratio affect the model performance. Figure 5 shows that a too low masking ratio enables the network to directly infer the structure from the masked image which is undesired in our case. In these cases, the network would not choose to infer a set of compact keypoints from the original image. Figure 5 illustrates that the patch size cannot be too small (4×4) or too large (64×64). Although in some cases, such as CelebA-in-the-Wild, a different masking strategy gives better results (4.14 vs 5.24), we choose to report a unified strategy that we mask 80% of 16×16 patches for simplicity and conciseness.

Variants of Edge Heatmap Generation. Besides the heatmap generation described in Method 3, we test four more ways to generate the edge maps: 1) we define the thickness as a globally learnable

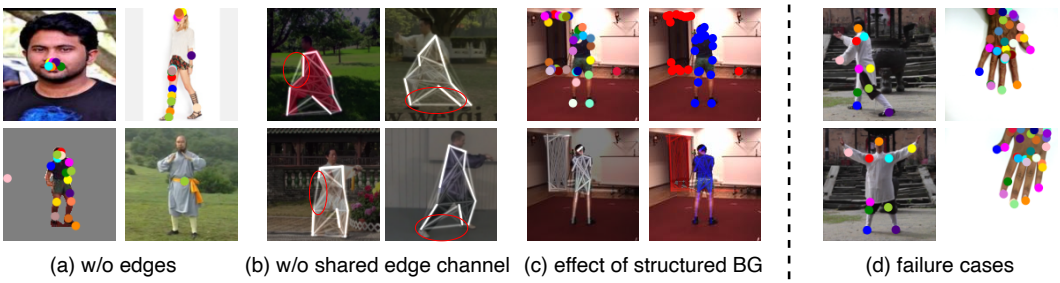


Figure 7: (a) If we do not model the edges, the model may degenerate. (b) If we give different edges a different channel in the feature map, the model would give dummy edges. (c) If the model is trained on the dataset with structured background, the background would be modeled. However, the keypoints can be separated into two sets by spectral clustering. (d) failure case: left) the model cannot model the occlusion well; right) the model has left and right ambiguity.

parameter; 2) we learn each edge thickness independently as a learnable parameter; 3) we treat each edge heatmap as an independent channel of the feature map, instead of making them a single channel heatmap with Equation 4; 4) instead of using the edge heatmap, we generate Gaussian heatmaps for the keypoints and use Equation 4 to combine them in a single channel heatmap. The results are listed in Table 3. Overall, these variants have worse performance. Although in some cases, the model has slightly better results on specific datasets, the performance boost does not hold in general. We observe that with only keypoints without edges, the model may degenerate, as shown in Figure 7a. Interestingly, assigning each edge a different channel performs worse than simply combining all edges into a single channel. We believe it is caused by entangling the edge weights with the convolution kernel weights. As visualized in Figure 7b, there exists dummy edges that do not model the object structure. In addition, we tried to remove the learnable α in Equation 5, fixing $\alpha = 1$, but the overall performance decreases as shown in Table 3.

Does Texture Matter? As shown in Figure 1a, the horse and the zebra share similar shape structures but only one is textured. The striped texture not having a significant impact on the learned structure shows that our model primarily learns the structure instead of texture features.

What if the model is trained on the structured background? We tested on Human3.6m with background, where all images are taken in a single room. The error is 5.02. As shown in Figure 7c, our model captures the entrance on the background. It is expected since we assume the foreground object is structured. If we apply spectral clustering [69] on the learned graph, the keypoints are clearly divided into two clusters.

5 Limitations and Future Work

As shown before, if the background is highly structured, the keypoints will appear on the background. We can use a post graph clustering to separate the object from the background. Similar to previous 2D self-supervised methods [59, 88, 31], our model cannot model occlusion well. We show in Figure 7d left that the occluded right arm becomes the back when the person turns to the left. In addition, as for all other methods, the model cannot distinguish left and right side of the human body as shown in Figure 7d right. We believe it is necessary to model the structure in 3D to solve these problems.

6 Conclusion

We presented a simple approach for learning a spatial graph representation from unlabelled image collections by reconstructing a masked images. The crucial part is our learnable graph design modeling the relationship between different keypoints. It is simpler than existing alternatives and opens up a path for image understanding, image editing and learning 3D models from 2D images.

References

[1] Mahmoud Afifi. 11k hands: gender recognition and biometric identification using a large dataset of hand images. *Multimedia Tools and Applications*, 2019.

[2] M. Andriluka, U. Iqbal, E. Insafutdinov, L. Pishchulin, A. Milan, J. Gall, and Schiele B. PoseTrack: A benchmark for human pose estimation and tracking. In *CVPR*, 2018.

[3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[4] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999.

[5] Sandro Braun. unsupervised-disentangling. `https://github.com/theRealSuperMario/unsupervised-disentangling/tree/reproducing_baselines`, 2020.

[6] Chenjie Cao and Yanwei Fu. Learning a sketch tensor space for image inpainting of man-made scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14509–14518, 2021.

[7] Nan Cao, Xin Yan, Yang Shi, and Chaoran Chen. Ai-sketcher : A deep generative model for producing high-quality sketches. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):2564–2571, Jul. 2019.

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

[9] Mickaël Chen, Thierry Artières, and Ludovic Denoyer. Unsupervised object segmentation by redrawing. In *Advances in Neural Information Processing Systems 32 (NIPS 2019)*, pages 12705–12716, 2019.

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[11] Yajing Chen, Shikui Tu, Yuqi Yi, and Lei Xu. Sketch-pix2seq: a model to generate sketches of multiple categories. *arXiv preprint arXiv:1709.04121*, 2017.

[12] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020.

[13] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018.

[14] Edo Collins, Radhakrishna Achanta, and Sabine Susstrunk. Deep feature factorization for concept discovery. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 336–352, 2018.

[15] Ayan Das, Yongxin Yang, Timothy Hospedales, Tao Xiang, and Yi-Zhe Song. Bézier sketch: A generative model for scalable vector sketches. In *European Conference on Computer Vision*, pages 632–647. Springer, 2020.

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

[17] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. *Advances in neural information processing systems*, 29, 2016.

[18] Charles Dugas, Yoshua Bengio, François Bélisle, Claude Nadeau, and René Garcia. Incorporating second-order functional knowledge for better option pricing. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2000.

[19] Aysegul Dundar, Kevin J Shih, Animesh Garg, Robert Pottorf, Andrew Tao, and Bryan Catanzaro. Unsupervised disentanglement of pose, appearance and background from images and videos. *arXiv preprint arXiv:2001.09518*, 2020.

[20] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM Trans. Graph. (Proc. SIGGRAPH)*, 31(4):44:1–44:10, 2012.

[21] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017.

[22] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 534–551, 2018.

[23] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

[24] Songwei Ge, Vedanuj Goswami, Larry Zitnick, and Devi Parikh. Creative sketch generation. In *International Conference on Learning Representations*, 2021.

[25] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[26] David Ha and Douglas Eck. A neural representation of sketch drawings. In *International Conference on Learning Representations*, 2018.

[27] Xintong Han, Zuxuan Wu, Weilin Huang, Matthew R. Scott, and Larry S. Davis. Finet: Compatible and diverse fashion image inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[28] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.

[29] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[30] Xingzhe He, Bastian Wandt, and Helge Rhodin. Latentkeypointgan: Controlling gans via latent keypoints. *arXiv preprint arXiv:2103.15812*, 2021.

[31] Xingzhe He, Bastian Wandt, and Helge Rhodin. Ganseg: Learning to segment by unsupervised hierarchical image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[32] Zixuan Huang and Yin Li. Interpretable and accurate fine-grained recognition via region grouping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8662–8672, 2020.

[33] Wei-Chih Hung, Varun Jampani, Sifei Liu, Pavlo Molchanov, Ming-Hsuan Yang, and Jan Kautz. Scops: Self-supervised co-part segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 869–878, 2019.

[34] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[35] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014.

[36] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[37] Aaron S Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *Proceedings of the IEEE international conference on computer vision*, pages 1031–1039, 2017.

[38] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. In *Advances in neural information processing systems*, pages 4016–4027, 2018.

[39] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Self-supervised learning of interpretable keypoints from unlabelled videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8787–8797, 2020.

[40] Yong Shi Jie Yang, Zhiquan Qi. Learning to incorporate structure knowledge for image inpainting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12605–12612, 2020.

[41] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[42] Kurmanbek Kaiyrbekov and Metin Sezgin. Deep stroke-based sketched symbol reconstruction and segmentation. *IEEE computer graphics and applications*, 40(1):112–126, 2019.

[43] Moritz Kampelmuhler and Axel Pinz. Synthesizing human-like sketches from natural images using a conditional convolutional decoder. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3203–3211, 2020.

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

[45] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

[46] Yunji Kim, Seonghyeon Nam, In Cho, and Seon Joo Kim. Unsupervised keypoint learning for guiding class-conditional video prediction. In *Advances in Neural Information Processing Systems*, pages 3814–3824, 2019.

[47] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[48] 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. In *Advances in neural information processing systems*, pages 10724–10734, 2019.

[49] Avisek Lahiri, Arnav Kumar Jain, Sanskar Agrawal, Pabitra Mitra, and Prabir Kumar Biswas. Prior guided gan based semantic inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[50] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.

[51] Jingyuan Li, Fengxiang He, Lefei Zhang, Bo Du, and Dacheng Tao. Progressive reconstruction of visual structure for image inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[52] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017.

[53] Liang Liao, Jing Xiao, Zheng Wang, Chia-Wen Lin, and Shin'ichi Satoh. Guidance and evaluation: Semantic-aware image inpainting for mixed scenes. In *European Conference on Computer Vision*, pages 683–700. Springer, 2020.

[54] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[55] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. Unsupervised part segmentation through disentangling appearance and shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8355–8364, 2021.

[56] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[57] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[58] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

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

[60] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. *Advances in neural information processing systems*, 30, 2017.

[61] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 99–108, 2018.

[62] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *in ICML Workshop on Deep Learning for Audio, Speech and Language Processing*. Citeseer, 2013.

[63] R. Malladi, J.A. Sethian, and B.C. Vemuri. Shape modeling with front propagation: a level set approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(2):158–175, 1995.

[64] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017.

[65] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International conference on machine learning*, pages 3481–3490. PMLR, 2018.

[66] Daniela Mihai and Jonathon Hare. Differentiable drawing and sketching. *arXiv preprint arXiv:2103.16194*, 2021.

[67] Matthias Minderer, Chen Sun, Ruben Villegas, Forrester Cole, Kevin P Murphy, and Honglak Lee. Unsupervised learning of object structure and dynamics from videos. In *Advances in Neural Information Processing Systems*, pages 92–102, 2019.

[68] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Structure guided image inpainting using edge prediction. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.

[69] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, page 849–856, Cambridge, MA, USA, 2001. MIT Press.

[70] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

[71] Atsuhiro Noguchi, Umar Iqbal, Jonathan Tremblay, Tatsuya Harada, and Orazio Gallo. Watch it move: Unsupervised discovery of 3D joints for re-posing of articulated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[72] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[73] Ahmed AA Osman, Timo Bolkart, and Michael J Black. Star: Sparse trained articulated human body regressor. In *European Conference on Computer Vision*, pages 598–613. Springer, 2020.

[74] Kaiyue Pang, Da Li, Jifei Song, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Deep factorised inverse-sketching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 36–52, 2018.

[75] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *Proceedings of the European conference on computer vision (ECCV)*, pages 269–286, 2018.

[76] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.

[77] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019.

[78] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H Li, Shan Liu, and Ge Li. Structureflow: Image inpainting via structure-aware appearance flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 181–190, 2019.

[79] Helge Rhodin, Victor Constantin, Isinsu Katircioglu, Mathieu Salzmann, and Pascal Fua. Neural scene decomposition for multi-person motion capture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[80] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation for 3d human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 750–767, 2018.

[81] Iasonas Kokkinos Riza Alp Güler, Natalia Neverova. Densepose: Dense human pose estimation in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[82] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[83] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: Learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (proceedings of SIGGRAPH)*, 2016.

[84] Kazuma Sasaki, Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Learning to Restore Deteriorated Line Drawing. *The Visual Computer (Proc. of Computer Graphics International 2018)*, 34(6-8):1077–1085, 2018.

[85] Luca Schmidtke, Athanasios Vlontzos, Simon Ellershaw, Anna Lukens, Tomoki Arichi, and Bernhard Kainz. Unsupervised human pose estimation through transforming shape templates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2484–2494, 2021.

[86] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2377–2386, 2019.

[87] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Conference on Neural Information Processing Systems (NeurIPS)*, December 2019.

[88] Aliaksandr Siarohin, Subhankar Roy, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Motion-supervised co-part segmentation. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9650–9657. IEEE, 2021.

[89] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuiliere, and Nicu Sebe. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3408–3416, 2018.

[90] Yuhang Song, Chao Yang, Yeji Shen, Peng Wang, Qin Huang, and C-C Jay Kuo. Spg-net: Segmentation prediction and guidance network for image inpainting. *arXiv preprint arXiv:1805.03356*, 2018.

[91] Shih-Yang Su, Frank Yu, Michael Zollhoefer, and Helge Rhodin. A-neRF: Articulated neural radiance fields for learning human shape, appearance, and pose. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[92] Supasorn Suwajanakorn, Noah Snavely, Jonathan J Tompson, and Mohammad Norouzi. Discovery of latent 3d keypoints via end-to-end geometric reasoning. In *Advances in neural information processing systems*, pages 2059–2070, 2018.

[93] 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.

[94] Fang Wang, Le Kang, and Yi Li. Sketch-based 3d shape retrieval using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1875–1883, 2015.

[95] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019.

[96] Jin Wang, Chen Wang, Qingming Huang, Yunhui Shi, Jian-Feng Cai, Qing Zhu, and Baocai Yin. *Image Inpainting Based on Multi-Frequency Probabilistic Inference Model*, page 1–9. Association for Computing Machinery, New York, NY, USA, 2020.

[97] Wayne Wu, Kaidi Cao, Cheng Li, Chen Qian, and Chen Change Loy. Transgaga: Geometry-aware unsupervised image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8012–8021, 2019.

[98] Zhirong Wu, Yuanjun Xiong, X Yu Stella, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[99] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *European Conference on Computer Vision (ECCV)*, 2018.

[100] Wei Xiong, Jiahui Yu, Zhe Lin, Jimei Yang, Xin Lu, Connelly Barnes, and Jiebo Luo. Foreground-aware image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5840–5848, 2019.

[101] H. Xu, X. Su, M. Wang, X. Hao, and G. Gao. An edge information and mask shrinking based image inpainting approach. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, Los Alamitos, CA, USA, jul 2020. IEEE Computer Society.

[102] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, Bill Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (Oral)*, pages 6184–6193, 2020.

[103] Peng Xu. Deep learning for free-hand sketch: A survey. *IEEE transactions on pattern analysis and machine intelligence*, PP, 2022.

[104] Peng Xu, Yongye Huang, Tongtong Yuan, Kaiyue Pang, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, Zhanyu Ma, and Jun Guo. Sketchmate: Deep hashing for million-scale human sketch retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8090–8098, 2018.

[105] Peng Xu, Chaitanya K Joshi, and Xavier Bresson. Multigraph transformer for free-hand sketch recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[106] Yinghao Xu, Ceyuan Yang, Ziwei Liu, Bo Dai, and Bolei Zhou. Unsupervised landmark learning from unpaired data. *arXiv preprint arXiv:2007.01053*, 2020.

[107] Yang Yang and Xiaojie Guo. Generative landmark guided face inpainting. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 14–26. Springer, 2020.

[108] Ran Yi, Yong-Jin Liu, Yu-Kun Lai, and Paul L Rosin. APDrawingGAN: Generating artistic portrait drawings from face photos with hierarchical gans. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR '19)*, pages 10743–10752, 2019.

[109] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M. Hospedales, and Chen Change Loy. Sketch me that shoe. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[110] Qian Yu, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. Sketch-a-net that beats humans. In *BMVC*, 2015.

[111] Tao Yu, Kaiwen Guo, Feng Xu, Yuan Dong, Zhaoqi Su, Jianhui Zhao, Jianguo Li, Qionghai Dai, and Yebin Liu. Bodyfusion: Real-time capture of human motion and surface geometry using a single depth camera. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 910–919, 2017.

[112] Yingchen Yu, Fangneng Zhan, Rongliang WU, Jianxiong Pan, Kaiwen Cui, Shijian Lu, Feiying Ma, Xuansong Xie, and Chunyan Miao. *Diverse Image Inpainting with Bidirectional and Autoregressive Transformers*, page 69–78. Association for Computing Machinery, New York, NY, USA, 2021.

[113] Xian Zhang, Canghong Shi, Xin Wang, Xi Wu, Xiaojie Li, Jiancheng Lv, and Imran Mumtaz. Face inpainting based on gan by facial prediction and fusion as guidance information. *Applied Soft Computing*, 111:107626, 2021.

[114] Yanping Zhang, Qiaokang Liang, Kunlin Zou, Zhengwei Li, Wei Sun, and Yaonan Wang. Self-supervised part segmentation via motion imitation. *Image and Vision Computing*, 120:104393, 2022.

[115] 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.

[116] Zibo Zhao, Wen Liu, Yanyu Xu, Xianing Chen, Weixin Luo, Lei Jin, Bohui Zhu, Tong Liu, Binqiang Zhao, and Shenghua Gao. Prior based human completion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7947–7957, 2021.

[117] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1438–1447, 2019.

[118] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 5
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 1
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [No] This paper does not work towards theoretical analysis.
   (b) Did you include complete proofs of all theoretical results? [No] This paper does not work towards theoretical analysis.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] The code, instructions, and the pre-trained models will be released upon acceptance. The data we used is already publicly available.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 3.3 and Appendix B,D.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Section 3.3 and Table 1& 2.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 3.3
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No] All the datasets are commonly used in research papers and their licenses allow the usage for research purposes.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No] All materials, including codes and pre-trained models, will be released upon acceptance.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] All the datasets are commonly used in research papers and the licenses allow the usage.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] All the datasets are commonly used in research papers and all datasets that contain humans are established datasets.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]