Unsupervised Keypoints from Pretrained Diffusion Models

Eric Hedlin[1] Gopal Sharma[1] Shweta Mahajan[1, 2] Xingzhe He[1] Hossam Isack[3]
Abhishek Kar[3] Helge Rhodin[1] Andrea Tagliasacchi[4, 5, 6] Kwang Moo Yi[1]
[1] University of British Columbia [2] Vector Institute for AI [3] Google Research
[4] Google DeepMind [5] Simon Fraser University [6] University of Toronto

https://stablekeypoints.github.io/

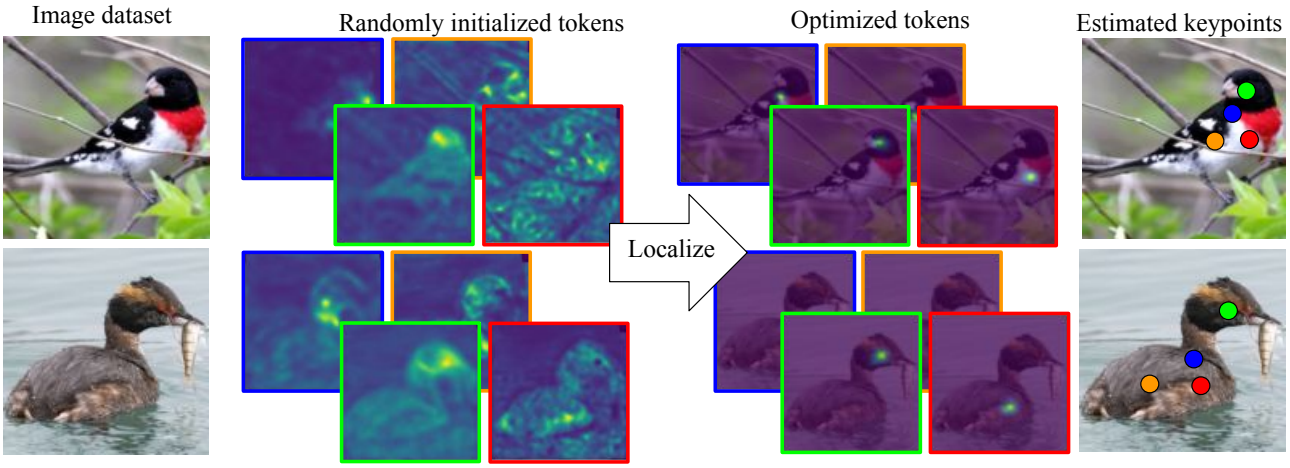


Figure 1. **Teaser** – we propose an unsupervised method to learn keypoints based on optimizing text embeddings of latent diffusion models [44]. Our method is motivated by the fact that random text tokens already respond roughly consistently to semantically similar regions. By promoting localization we obtain unsupervised keypoints that outperform the state-of-the-art.

Abstract

Unsupervised learning of keypoints and landmarks has seen significant progress with the help of modern neural network architectures, but performance is yet to match the supervised counterpart, making their practicability questionable. We leverage the emergent knowledge within text-to-image diffusion models, towards more robust unsupervised keypoints. Our core idea is to find text embeddings that would cause the generative model to consistently attend to compact regions in images (i.e. keypoints). To do so, we simply optimize the text embedding such that the cross-attention maps within the denoising network are localized as Gaussians with small standard deviations. We validate our performance on multiple datasets: the CelebA, CUB-200-2011, Tai-Chi-HD, DeepFashion, and Human3.6m datasets. We achieve significantly improved accuracy, sometimes even outperforming supervised ones, particularly for data that is non-aligned and less curated. Our code is publicly available at the project page.

1. Introduction

Keypoints or landmarks have played a critical role in computer vision for various task including image matching [31], 3D reconstruction [18], and motion tracking [32, 60]. Similarly to many other areas of computer vision, research has quickly adopted supervised learning to tackle this problem [3, 27]. However, labeling is tedious and sometimes even ambiguous—for example, it is difficult to consistently decide which keypoints on a human face are the "most important". Researchers have therefore been investigating unsupervised approaches [12, 13, 19, 30, 55, 67]. These are typically implemented as autoencoders paired with handcrafted intermediate layers, or losses that enforce spatial locality and equivariance of keypoint locations under deformation. However, as we will show later, these methods struggle with non-preprocessed data, and their performance is heavily reliant on knowing the ground truth location of objects, clearly limiting their practical applicability.

To enhance the learning of unsupervised keypoints, we draw inspiration from the demonstrated success of scaling up datasets [52]. For example, in natural language pro-

Figure 2. **Example attention maps –** we show example attention maps for a selected learned keypoint for the CUB-200-2011 dataset, on the CUB-aligned subset. As shown, our keypoint attention map responds consistently across varying images.

cessing performance has recently improved to a great extent thanks to large models and data [7, 39, 57]. Similarly, in computer vision, the performance of text-to-image models [43, 44, 46] has drastically improved thanks to the availability of *extra* large datasets [48]. However, unsupervised keypoint learning typically assumes class-specific datasets, *e.g.*, animals that have a shared skeleton that connects keypoints, and these datasets are small in scale.

Rather than collecting larger domain-specific datasets, we instead propose to leverage the knowledge stored within large generative models, such as Stable Diffusion [44]. This has been shown to be very effective across a number tasks [1, 2, 4, 8, 15, 24, 26, 33, 35, 41, 54, 56, 59, 61, 63, 64, 66], but, to the best of our knowledge, it has not yet found application for the task of keypoint learning. Our main idea is to localize "important" keypoints by finding text embeddings that *consistently* correspond to a distinct location in images of a certain object class. This idea is rooted in the observation that, even with random text embeddings, the attention maps for various images roughly correspond to regions that are semantically similar; see Fig. 1. Therefore, text embeddings carry semantic meaning, which could be used to relate collections of images to each other; see Fig. 2.

We find embeddings that are specific to certain locations by enforcing localized attention maps. In more detail, we propose to find (*i.e.*, optimize) a set of tokens in a text embedding that locally responds in the Stable Diffusion cross-attention layers. We enforce locality by maximizing the similarity of the attention responses of each token to a single-mode Gaussian distribution. Thanks to the way the cross-attention layers are constructed within Stable Diffusion, this simple objective also prevents the different tokens from attending to the same locations in an image, a common degenerate solution that typically requires explicit workarounds [23].

We evaluate our method on established benchmarks: CelebA [28], CUB-200-2011 [58], Tai-Chi-HD [50], DeepFashion [29], and Human3.6m [17]. Our approach yields results on par with state-of-the-art methods for well-curated and aligned datasets, while notably enhancing performance for in-the-wild setups, particularly with unaligned data, sometimes even surpassing fully supervised baselines.

2. Related Work

Below we review the literature of finding keypoints in unsupervised and supervised fashion, along with work that exploits large pre-trained models like stable-diffusion for lower-level computer vision tasks.

Learning keypoints with supervision. Pose estimation and landmark estimation are fundamental problems in computer vision. They naturally arise in various tasks, including human [69] and animal pose estimation [22], hand [5] and face landmark estimation [62], and object pose tracking [34]. Many fully supervised methods find different ways to induce some prior within the model to better capture the task at hand, such as using part affinity fields [3], temporal consistency for video data [45], spacial relationships [65], and geometry constraints [21] among others. While fully supervised methods have excelled in categories with abundant labeled data, such as human pose estimation, their major drawback is the insatiable need for large and high-quality datasets [22, 37, 42]. The scalability of gathering such extensive and meticulously annotated data for every conceivable object category remains a significant drawback [22, 37, 42].

Learning keypoints via self-supervision. The amount of unlabeled data far exceeds that of labeled data, so unsupervised keypoint estimation methods attempt to take advantage of this. Self-supervised keypoint detection often relies on tracking how keypoints move with image changes and uses various constraints for known transformations [16, 19, 30, 49, 55, 67], but these methods can struggle with background modeling [49, 67] and pose variations [16]. One can also rely on image reconstruction to learn keypoints. Some methods use GANs to generate images from keypoints [12, 14], but this often results in training instability. Alternatively, auto-encoders can be also used [13, 67], but these require training from scratch on each dataset. Our method neither suffers from GAN

training instability, nor requires dataset fine-tuning. Finally, there exist self-supervised methods that exploit skeletal representations [13, 20, 40]. However, many of these approaches generally require known keypoint connectivity and video data [20, 40], and often face limitations in background handling and generalizability to objects within the same class [13, 20, 40]. Our method has no object-specific priors, and generalizes well due to the large dataset used by the pre-trained diffusion models.

Diffusion models for image understanding. Recently, large image diffusion models have reached impressive image generation quality [43, 44, 46]. These models learn priors for real images within the latent space of the diffusion model, and provide a useful initialization for many downstream tasks such as image correspondence [15, 33, 54, 66], object detection [4], semantic segmentation [2, 24, 56, 59, 61, 63, 64], and image classification [1, 8]. Interestingly, without requiring any retraining, these models demonstrate an innate ability to understand 3D spatial configurations [26, 35, 41]. Recent work in each of these areas has shown the emergent power of these large models, most of them using the model without any modifications or extra supervision required. More relevant to our work, Mokady et al. [36] found that the pre-trained Stable Diffusion [44] model's cross-attention maps connect text tokens to semantically relevant areas in images.

Correspondences via diffusion models. Among works that re-purpose diffusion models, of high relevance, is the effectiveness of diffusion models in correspondence estimation tasks [15, 33, 54, 66]. Hedlin et al. [15] optimizes the attention map for a specific point in a source image and finds the corresponding activation in a target image. However, this method requires a *query* to be provided in the source image. While our method shares the same inspiration of utilizing attention maps, critically, rather than optimizing the embedding given a single image, we optimize an embedding given a dataset of images from a given object class. In other words, our method discovers on its own, where to focus, rather than relying on user input. Our task is therefore changed from image matching between *two* images, to semantic matching across *all* images within the dataset.

3. Method

To identify a set of representative keypoints across a dataset of images, we formulate our approach in an unsupervised framework leveraging conditional diffusion models; see Fig. 3. In particular, we utilize the cross-attention maps between the text embeddings and the image features, derived from the latent diffusion model [44], and force them to consistently concentrate their activation on highly localized regions within the images. While Hedlin et al. [15] employed a similar mechanism (*i.e.*, given a set of keypoint locations in one image, identify correspondences in another image), in this work we seek to identify semantic correspondences across all images within a class-specific dataset (*e.g.*, human faces), *without any given* knowledge on what and where to focus. We show that this is possible simply by enforcing locality and equivariance to transformations.

Let us start by quickly reviewing the fundamentals of diffusion models and formalizing the attention maps that we will utilize within these models (Sec. 3.1). We then detail the objectives used to learn the text embeddings that represent keypoints (Sec. 3.2) and discuss important implementation details (Sec. 3.3).

3.1. Attention maps in diffusion networks

Diffusion models are a class of generative models that approximate the data distribution by denoising a base (typically Gaussian) distribution [38]. A *latent* diffusion model operates on a latent representation $\mathbf{z}$ rather than the image itself, with an encoder that maps an image $\mathbf{X}$ into a latent $\mathbf{z}$, and a decoder that maps $\mathbf{z}$ into $\mathbf{X}$. These models define a *forward* diffusion process, where the latent representation $\mathbf{z}$ is gradually transformed into Gaussian noise over a series of T time steps. The *inverse* process, over a sequence denoising steps $t = 1, \dots, T$ predicts the latent noise $\epsilon_{\boldsymbol{\theta}}(\mathbf{z}_t, t)$ which was gradually added in each iteration in order to recover the original (latent) signal.

In our work, we are interested in *conditional* diffusion models, and the explicit attentional relationship between the condition (i.e. text) and the generated outcome (i.e. image) that these models learn. Typically, diffusion models are made conditional on some text $\mathbf{y}$, by providing an embedding $\mathbf{e} = \tau_{\boldsymbol{\theta}}(\mathbf{y})$ from a text encoder $\tau_{\boldsymbol{\theta}}$ to the denoiser. They are then trained to optimize

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

where the denoiser $\epsilon_{\boldsymbol{\theta}}(\mathbf{z}_t, t, \mathbf{e})$ is typically implemented by a transformer architecture [38] involving a combination of self-attention and cross-attention layers. Of our interest here is the cross-attention layers that relate $\mathbf{e}$ to $\mathbf{z}_t$, which we now formalize.

Specifically, in the transformer part of the model, denote $\Phi_l^c(\cdot)$ and $\Psi_l^c(\cdot)$ as the c-th head and the l-th linear layers of the U-Net. We calculate the query as $\mathbf{Q}_l^c = \Phi_l^c(\mathbf{z}_{t=1}) \in \mathbb{R}^{(H\times W)\times D_l}$,[1] and the key from the language embedding $\mathbf{K}_l^c = \Psi_l^c(\mathbf{e}) \in \mathbb{R}^{N\times D_l}$, where N is the number of tokens, C the number of heads in the transformer attention layer, H and W are image height and width at that specific layer in U-Net, and D_l the dimensionality of the layer. Given query and key, the cross-attention map $\mathbf{M}_l \in \mathbb{R}^{(H\times W)\times N}$ is then computed via softmax along the N di-

[1] We choose t=1 where T=50 steps via hyper-parameter tuning.

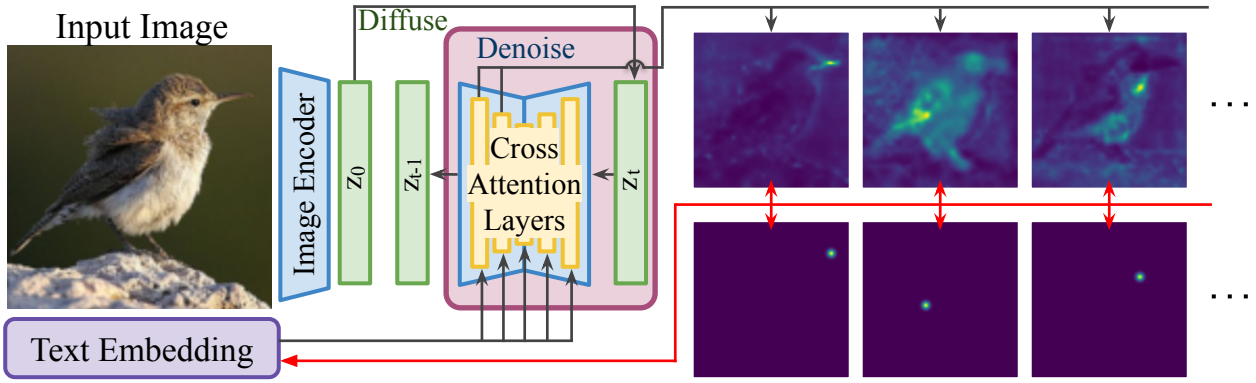


Figure 3. **Overview –** we pass a randomly initialized text embedding into Stable Diffusion [44] and extract the attention maps. We then optimize the text embedding to have localized attention maps, by supervising them to become a single-mode Gaussian distribution, drawn at the location of their maxima. We also enforce attention maps to be transformation equivariant to small affine transformations on images. We repeat this process over a set of training images, which after optimization provides a set of K keypoints.

mension, and average pooling along the C dimension:

$$\mathbf{M}_l(\mathbf{e}, \mathbf{X}) = \mathbb{E}_c \left[\text{softmax}_n \left(\mathbf{Q}_l^c \cdot \mathbf{K}_l / \sqrt{D_l} \right) \right]. \quad (2)$$

As various layers of the U-Net exhibit distinct levels of semantic understanding, following Hedlin et al. [15], we collect this information by average pooling across a *selection* of layers:

$$\tilde{\mathbf{M}} = \mathbb{E}_{l=7..10} \left[\mathbf{M}_l(\mathbf{e}, \mathbf{X}) \right] \in \mathbb{R}^{(H \times W) \times N}. \quad (3)$$

In what follows, to lighten the notation, we drop the attention mask arguments $(\mathbf{e}, \mathbf{X})$ and write the attention map for the n-th token as $\tilde{\mathbf{M}}_n$.

3.2. Optimizing to find the keypoint embeddings

To obtain a text embedding that can be used to locate keypoints, for each of them, we simply optimize for two objectives that respectively encourages localization and equivariance to geometric transformations. We thus write

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{localize}} + \lambda_{\text{equiv}} \mathcal{L}_{\text{equiv}}, \quad (4)$$

where we apply λ_{equiv}=10 to balance the two losses to be in a similar operating range. Equivariance is enforced in the typical form of learning to be invariant to transformations. We first quickly detail $\mathcal{L}_{\text{equiv}}$, and then discuss how we enforce localization, which is the core of our method.

Equivariance – $\mathcal{L}_{\text{equiv}}$. To ensure our model's attention mechanism remains consistent across different geometric transformations $\mathcal{T}$ of the input, we use the typical equivariance loss [25]:

$$\mathcal{L}_{\text{equiv}} = \mathbb{E}_n \, \| \mathcal{T}^{-1}(\mathbf{M}_n(\mathbf{e}, \mathcal{T}(\mathbf{X}))) - \mathbf{M}_n(\mathbf{e}, \mathbf{X}) \|^2 \quad (5)$$

For $\mathcal{T}$ we simply utilize minor affine transformations. We use random rotations between ± 15 degrees, translations between $\pm 0.25 \times W$, and scaling between 100–120% of the original image size.

Encouraging localization – $\mathcal{L}_{\text{localize}}$. We encourage localization by forcing $\tilde{\mathbf{M}}_n$ to be a single-mode Gaussian distribution located at its maximum. In more details, denoting the Gaussian image that shares the same maximum as $\tilde{\mathbf{M}}_n$ as $\mathbf{G}_n$, we write

$$\mathcal{L}_{\text{localize}} = \mathbb{E}_n \| \tilde{\mathbf{M}}_n - \mathbf{G}_n \|^2. \quad (6)$$

To create the Gaussian images $\mathbf{G}_n$, we first identify the spatial location exhibiting the maximal response within the heatmap corresponding to each token n by taking the argmax:

$$\boldsymbol{\mu}_n = \arg\max_{w,h} \; \tilde{\mathbf{M}}_n[h, w]. \quad (7)$$

We then generate a Gaussian image; see Fig. 3:

$$\mathbf{G}_n = \exp\left(-\frac{\| \mathbf{XY}_{\text{coord}} - \boldsymbol{\mu}_n \|_2^2}{2\sigma^2} \right), \quad (8)$$

where $\mathbf{XY}_{\text{coord}}$ is a tensor of image coordinates.

Promoting mutual exclusivity. It is important to note that while $\mathcal{L}_{\text{localize}}$ in (6) at first glance seem to only encourage localization, it also enforces $\tilde{\mathbf{M}}_n$ to be mutually exclusive for different n because of the softmax operation in (2). Should multiple embeddings become similar, their attention responses in (2) $\mathbf{Q}_l^c \cdot \mathbf{K}_l$ will become similar, resulting in the softmax of the attention map being a flat response (i.e. deviating from a Gaussian shape). In other words, (6) naturally enforces exclusivity with the help of (2).

Stabilizing optimization by working with a subset. We noticed in our experiments that attention maps $\mathbf{M}$ for some

tokens can be 'spread' for some images, *e.g.*, due to occlusions, destabilizing optimization. We thus opt for a simple solution of looking into the top-K tokens that are local. Specifically, we apply our losses over $n \in \mathcal{N}(\kappa)$, which returns the $\kappa \in \mathbb{N}$ entries with the most spatially localized heatmap responses [2], as measured by KL divergence:

$$\mathcal{N}(\kappa) = \text{ArgTop}_\kappa \ \{-\text{KL}(\mathbf{G}_n, \tilde{\mathbf{M}}_n)\}_{n=1}^{N}. \tag{9}$$

Final keypoints. While $\mathcal{L}_{\text{localize}}$ naturally enforces exclusivity, it does not guarantee a complete coverage of the object. Thus, after we finish optimizing, we refine the set of keypoints through furthest point sampling using the training images. Specifically, for each image we write:

$$\mathcal{K} = \text{FPS}_K(\{\boldsymbol{\mu}_i \mid i \in \mathcal{N}(\kappa)\}), \tag{10}$$

where K is the desired number of keypoints $K < \kappa$. Then, as the set $\mathcal{K}$ differs in each image, we simply choose K tokens that appeared most frequently in $\mathcal{K}$ within the training image set.

3.3. Implementation details

Test-time ensembling. At inference time, as in common literature [3, 10, 15], rather than employing the attention map of the original image, we average the attention maps across multiple augmentations (we use the same transformations for test time augmentation as in (5)) of the image:

$$\tilde{\mathbf{M}} \longleftarrow \sum_i \mathcal{T}_i^{-1}\left(\mathbf{M}\left(\mathbf{e}, \mathcal{T}_i(\mathbf{X})\right)\right). \tag{11}$$

Upsampling attention maps. The attention maps in (2) are typically of low resolution. Specifically, as we use Stable Diffusion 1.5 [44], depending on the layer we extract the attention maps from, they are either 16×16 or 32×32. We thus opt to upsample the query $\mathbf{Q}$ via bicubic interpolation to achieve a standard resolution of 128×128. We have experimented with other upsampling techniques such as the commonly used bilinear sampling or a learned upsampler that is trained alongside, but a simple bicubic upsample was shown to be effective.

4. Results

4.1. Experimental setup

We evaluate our method on five standard datasets for unsupervised keypoint evaluation:

- **CelebA** dataset [28]: A dataset of 202,599 facial images of celebrities. We evaluate both the aligned and non-aligned cases following the standard protocol of omitting images with faces occupying less than 30% of the image. The standard metric for this dataset is to measure the average ℓ_2 error normalized by the inter-ocular distance.
- **CUB-200-2011** dataset [58]: This dataset consists of 11,788 bird images. We use both the aligned (CUB-aligned) and non-aligned (CUB-all) variants. For the non-aligned variants, we further look at CUB-001, CUB-002, and CUB-003, which are specific bird subcategories. Notably, these subsets contain only 30 images each—we only use these 30 for training. We follow the standard protocol [6, 30] and normalize the images to be of 256×256. The standard metric for this dataset is the mean ℓ_2 error, normalized by the dimension of the images after normalization.
- **Tai-Chi-HD** dataset [50]: This dataset contains 3049 training videos and 285 test videos of people performing Tai-Chi, which shows more diverse poses compared to the other datasets, and is the most challenging among the human pose-centric datasets that we use. We follow Siarohin et al. [51] and use 500 images for testing and 300 images for training. The standard metric for this dataset is to measure the accumulated ℓ_2 error, with the images standardized to 256×256.
- **DeepFashion** dataset [29]: This dataset contains 53k images of fashion models, mostly standing with a white background. We follow Lorenz et al. [30] and only keep full body images. This leaves 10,604 images for training and 1,179 images for testing. Also following the baselines, we use keypoints generated by AlphaPose [11] as ground truth. The standard metric for this dataset is the percentage of correct keypoints (PCK) with a 6-pixel threshold.
- **Human 3.6M** dataset [17]: This dataset is of humans performing various actions, comprised of 3.6 million images. We follow the standard protocol [67] and focus on six activities: direction, discussion, posing, waiting, greeting, and walking. We utilize subjects 1, 5, 6, 7, 8, and 9 for training, while subject 11 is reserved for testing. This division yields a training dataset comprising 796,648 images and a testing dataset containing 87,975 images. The background for this dataset is also simple, and often masked out with ground-truth masks for evaluation. This dataset is also typically heavily pre-processed and aligned when used for unsupervised keypoint evaluation. We experiment with the standard pre-processing [30, 67] and also a relaxed version of our own. To relax the alignment, we crop a square bounding box such that the margin from the bounding box to the person is 100 pixels, which on average corresponds to the person's height being $2/3$ of the crop. We further add a uniform random translation up to 100 pixels (same as the margin) to remove the central bias. Example crops are visualized in Fig. 4f. The standard metric for this dataset is the ℓ_2 error after normalizing the image resolution to 128×128.

[2] We empirically found that using κ=25 works best in general.

Method	Aligned (K=10)↓	Wild (K=4) ↓	Wild (K=8) ↓
Thewlis et al. [55]	7.95	-	31.30
Zhang et al. [67]	3.46	-	40.82
LatentKeypointGAN [12]	5.85	25.81	21.90
Lorenz et al. [30]	3.24	15.49	11.41
IMM [19]	**3.19**	19.42	8.74
LatentKeypointGAN-tuned [12]	3.31	12.10	5.63
Autolink [13]	3.92	7.72	5.66
Autolink † [13]	3.54	6.11	5.24
Our method	3.60	**5.24**	**4.35**

Table 1. **Quantitave results for the CelebA dataset –** we report results with the standard metrics. Our method performs best for non-aligned cases and is comparable to the state of the art for the aligned case. † symbol represents the thickness-tuned variant.

Note that each dataset comes with its own metric. To make results more comparable across the human pose datasets, we report both their original metrics as well as the ℓ_2 error when normalizing the image resolution to 128×128.

Regressing human-annotated landmarks. To evaluate the quality of unsupervised keypoints, one must relate them with human-annotated landmarks. As in prior research [55], we use linear regression (without bias) to relate between unsupervised keypoints and human-annotated landmarks.

Number of keypoints and hyperparameters. For each method, we use the standard number of unsupervised keypoints defined for each evaluation protocol—we denote them in our Tables. We use the same hyperparameter for all our experiments as introduced in Sec. 3.2, except for the number of optimization iterations. We optimize the embeddings for 10k iterations, except for the human pose datasets, for which we optimize 500 iterations. To find the number of optimization rounds we use a 10% validation subset from the training data. While we observed our results on the validation subset to improve consistently for most datasets, we found 10k to give a reasonable optimization time of two hours on an RTX 3090. For the human pose dataset, we found optimization to have converged already at 500 iterations on our validation split.

4.2. Experimental results

Quantiative results – Tabs. 1 to 3. We report our results for each dataset in Tabs. 1 to 3. As shown, except for the case when data is heavily processed and aligned (CelebA aligned in Tab. 1, CUB-aligned in Tab. 2, and Human 3.6M in Tab. 3), our method significantly outperforms the state of the art. The most visible gains are for the Tai-Chi-HD dataset, the most challenging among human pose datasets, and on CUB unaligned datasets. For the CUB dataset and the Tai-Chi-HD datasets, we outperform even those that have been supervised with silhouettes or saliency maps.

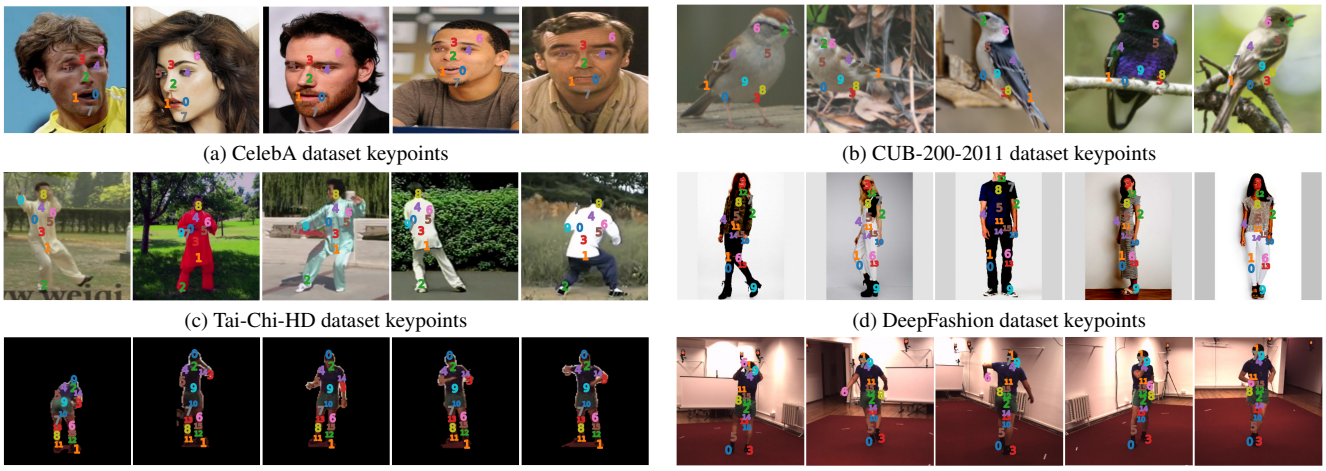
Method	Supervision	CUB-aligned (K=10) ↓	CUB-001 (K=4) ↓	CUB-002 (K=4) ↓	CUB-003 (K=4) ↓	CUB-all (K=4) ↓
SCOPS [16]	GT silhouette	-	18.3	17.7	17.0	12.6
Choudhury et al. [6]	GT silhouette	-	11.3	15.0	10.6	9.2
DFF [9]	testing dataset	-	22.4	21.6	22.0	-
SCOPS [16]	saliency maps	-	18.5	18.8	21.1	-
Lorenz et al. [30]	unsupervised	3.91	-	-	-	-
ULD [55, 67]	unsupervised	-	30.1	29.4	28.2	-
Zhang et al. [67]	unsupervised	5.36	26.9	27.6	27.1	22.4
LatentKeypointGAN [12]	unsupervised	5.21	22.6	29.1	21.2	14.7
GANSeg [14]	unsupervised	**3.23**	22.1	22.3	21.5	12.1
Autolink [13]	unsupervised	4.15	20.6	20.3	19.7	11.6
Autolink † [13]	unsupervised	3.51	20.2	19.2	18.5	11.3
Our method	unsupervised	5.06	**10.5**	**11.1**	**10.3**	**5.4**

Table 2. **Quantitave results for the CUB-200-2011 dataset –** we report results with the standard metrics. Except for the CUB-aligned case, our method performs nearly twice better than the compared methods, even outperforming Choudhury et al. [6], which is supervised with ground-truth silhouettes. † represents the thickness-tuned variant.

Method	Supervision	Human 3.6M (K=16) ℓ_2 standard / unaligned ↓	DeepFashion (K=16) PCK↑ / Rel. ℓ_2 ↓	Tai-Chi-HD (K=10) Cum ℓ_2 ↓ / Rel. ℓ_2 ↓
Newell et al. [20]	paired gt	2.16 / -	-	-
DFF [9]	testing dataset	-	-	494.48 / 14.78
SCOPS [16]	saliency maps	-	-	411.38 / 12.29
Jakab et al. [20]	video*	2.73 / -	-	-
Siarohin et al. [51]	videos	-	-	389.78 / 11.65
Zhang et al. [68]	videos	-	-	343.67 / 10.27
Zhang et al. [67]	videos	4.14 / -	-	-
Schmidtke et al. [47]	video*	3.31 / -	-	-
Sun et al. [53]	videos	2.53 / -	-	-
Thewlis et al. [55]	unsupervised	7.51 / -	-	-
Zhang et al. [67]	unsupervised	4.91 / -	-	-
LatentKeypointGAN [12]	unsupervised	-	49%	437.69 / 13.08
Lorenz et al. [30]	unsupervised	2.79 / -	57%	-
GANSeg [14]	unsupervised	-	59%	417.17 / 12.47
autolink [13]	unsupervised	2.81 / 7.59	65%	337.50 / 10.08
autolink † [13]	unsupervised	**2.76** / -	66%	316.10 / 9.45
Our method	unsupervised	4.45 / **5.77**	**70%**/6.46	**234.89 / 7.02**

Table 3. **Quantitative results for human pose datasets –** We report results for the Human 3.6M dataset, Deep Fashion dataset, and the Challenging Tai-Chi-HD datasets. We report both standard metrics for each dataset and the relative ℓ_2 error after normalizing images to 128 × 128. Our method, except for the Human 3.6M dataset that is heavily pre-processed, outperforms all baselines. This includes, for the challenging Tai-Chi-HD dataset, supervised ones. * denotes additional supervision (Jakab et al. [20] uses unpaired ground truth, and Schmidtke *et al*. [47] use the T-pose). The † symbol represents the thickness-tuned variant for Autolink.

We note that our primary focus is on unaligned cases, as we argue that they represent more how keypoints would be used in real-world applications—most real-world datasets are unaligned except for specific classes of objects. Moreover, methods focusing on aligned settings use strong locational priors, and as shown by their results in the unaligned setup—CelebA in the wild, non-aligned cases of Human 3.6M and CUB-200-2011, and Tai-Chi-HD—may perform significantly worse once this alignment prior is broken. Given that the performance of our method, even in the aligned case, is not too far off from methods that utilize alignment, we suspect a more in-depth tuning of our method may make our method outperform these methods,



(a) CelebA dataset keypoints

(b) CUB-200-2011 dataset keypoints

(c) Tai-Chi-HD dataset keypoints

(d) DeepFashion dataset keypoints

(e) Human 3.6M dataset keypoints

(f) Unaligned Human 3.6M dataset keypoints

Figure 4. **Qualitative examples of unsupervised keypoints –** we show our learned keypoints for the CelebA, CUB-200-2011, Tai-Chi-HD, DeepFashion, and Human 3.6M datasets (both for cropped and masked as well as our relaxed version). Note how our keypoints are consistent despite the variability. Our method significantly outperforms other baselines, especially for the challenging Tai-Chi-HD dataset and the CUB subsets.

Variant	Normalized ℓ_2
Full (Our method)	5.4
Without test time ensembling	5.6
Without furthest point sampling	6.4
Without upsampling the query $\mathbf{Q}$	8.0
Without equivariance	22.2

Table 4. **Ablation results –** we report the effect of each of our design choices can be seen on the CUB-all dataset. All components contribute to the final performance.

but we leave this as future work.

Finally, also note that for CUB-001, CUB-002, and CUB-003, these datasets are small. These datasets are non-aligned, have a large variability between individual images, and only contain 30 images each in the training set. Our method, *just from 30 images*, successfully identifies keypoints. These results highlight the potential of leveraging emergent (prior) knowledge within Stable Diffusion [44].

Qualitative results – Fig. 4. We provide example visualizations of our unsupervised keypoints in Fig. 4. As shown, our method discovers keypoints that are consistently localized across the dataset, despite the wide appearance variety.

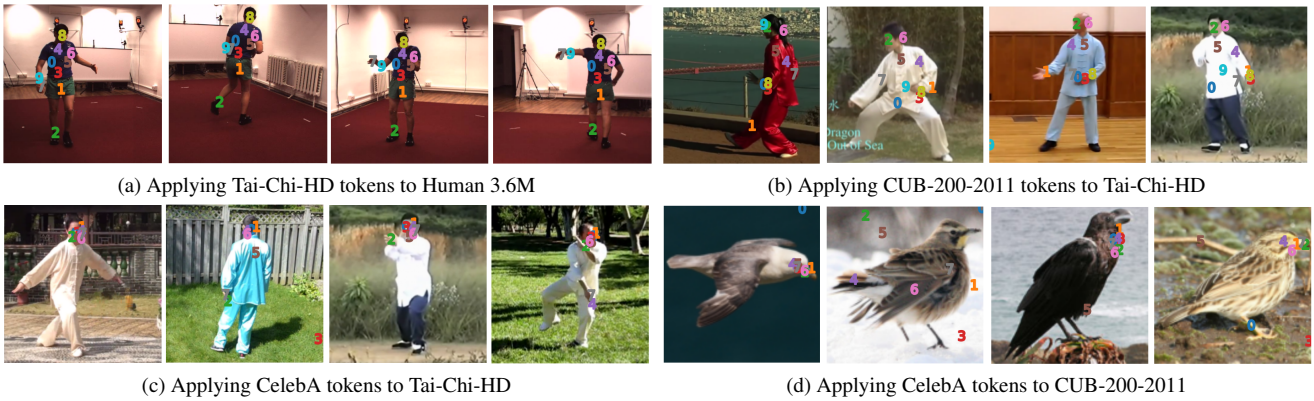
4.3. Ablation study

We perform an ablation study for various design choices of our method on the CUB-all dataset. We report the performance of our method with different components disabled in Tab. 4. As shown, all components contribute to the final performance. Test-time ensembling enhances performance, but the computation cost linearly scales. We choose, ten augmentations, which provide a good compromise between computation time and accuracy. To remove furthest point sampling we set $\kappa{=}K$, which then makes furthest point sampling select all samples. While this causes points to be more grouped, it still provides reasonable performance. To remove upsampling we instead upsample $\mathbf{M}$ to the size of the target image, effectively having the attention map build at lower resolutions, sometimes as low as $16{\times}16$. This results in significant degradation in performance. $\mathcal{L}_{\text{equiv}}$ is essential, as without it, the tokens can 'cheat' and simply opt to learn fixed positions on the image.

Number of training images.. Inspired by our results for the small subsets of CUB-200-2011 dataset, we investigate the impact that the number of images that we use to find keypoints has on our results. We thus optimized our keypoints only with 100 images for CelebA non-aligned setup. Surprisingly, we achieve 5.33 $K{=}8$, which is comparable to the state of the art. This demonstrates once more how our method is able to leverage information that is already learned in Stable Diffusion [44] to find keypoints.

4.4. Generalization

We further test the generalization capacity of our learned keypoints. As they are effectively text embeddings, we can simply apply them to any image, including those completely outside of the training domain. We quantitatively evaluate our method and the previous best-performing method Autolink [13]. We find that even in these generalization experiments, our keypoints reach performance comparable to data-specific keypoints. Applying Tai-Chi-HD tokens to un-



(a) Applying Tai-Chi-HD tokens to Human 3.6M

(b) Applying CUB-200-2011 tokens to Tai-Chi-HD

(c) Applying CelebA tokens to Tai-Chi-HD

(d) Applying CelebA tokens to CUB-200-2011

Figure 5. **Generalization –** we apply our learned text tokens (keypoints) to images from other datasets, including those that are of completely different domains. Our tokens generalize well for data of similar type, and surprisingly well even for some extreme cases.

aligned Human 3.6M achieves state-of-the-art performance despite using fewer keypoints (K=10 vs K=16). We also perform on par with the previous state of the art when we apply CUB-200-2011 tokens to Tai-Chi-HD—a case where the dataset gap is not only about the appearance but also beyond classes. Despite the drastic gap, our keypoints perform extremely well, leveraging the generalization power of large pre-trained diffusion models.

We show qualitative examples in Fig. 5. As shown, even when applied to different datasets, they look reasonable. For example, in Fig. 5a, when applying Tai-Chi-HD tokens to Human 3.6M, the tokens respond to the same locations on the human body as in Tai-Chi-HD. Surprising was when we applied CUB-200-2011 tokens to Tai-Chi-HD in Fig. 5b—they still responded to the body of the human being, reasonably consistently, although these tokens were trained to respond to birds. Of note are tokens two and six, which correspond to the front and back of the bird heads in Fig. 4—they also reply to the front and back of human heads. Applying CelebA tokens to Tai-Chi-HD in Fig. 5c also shows interesting outcomes, as tokens generally respond to human faces, despite the scale being drastically different between the two datasets. Finally, applying CelebA tokens to the CUB-200-2011 dataset in Fig. 5d shows mixed results—when it is 'successful' it focuses also on the faces of the bird, when it fails, it fails completely. These results hint that the keypoints (tokens) we have learned carry semantic meanings, as expected. We note that none of the baselines that we compare against are able to generalize beyond the dataset they were trained for.

5. Conclusions

We have proposed a novel method to find unsupervised keypoints using pre-trained text-to-image diffusion models. Given a set of images of a certain object, we propose to optimize the text embeddings (tokens) such that the cross-attention maps within diffusion models become localized as Gaussians with a small standard deviation. By doing so, we find text tokens that can be used to extract keypoints by extracting the maxima of the attention maps. We have shown that our method, on multiple datasets, under the challenging un-aligned setup, significantly outperforms the state of the art. We have further demonstrated that these tokens are also generalizable.

	Tai-Chi-HD →unaligned Human3.6m (K=10)↓	CUB-200-2011 →Tai-Chi-HD (K=10)↓ Cum ℓ_2 ↓ / Rel. ℓ_2 ↓	CelebA →Tai-Chi-HD (K=8)↓ Cum ℓ_2 ↓ / Rel. ℓ_2 ↓	CelebA →CUB-200-2011 (K=8)↓
Ours	**4.88**	**317.94 / 9.50**	- / **8.6**	**18.60**
Autolink [13]	16.92	535.61 / 16.00	- / 28.2	22.56

Table 5. **Generalization –** we quantitatively evaluate the performance of our keypoints on other datasets. Our Tai-Chi-HD keypoints applied to the unaligned Human3.6m setting reach state-of-the-art performance. Interestingly, our CUB keypoints applied to Tai-Chi-HD are on par with the previous state of the art, despite the differences between these datasets.

6. Acknowledgments

The authors would like to thank Cristina Vasconcelos for her constructive feedback during the preparation of this manuscript. Additionally, we extend our gratitude to David Fleet for his approval and support of this work.

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant, NSERC Collaborative Research and Development Grant, Google, Digital Research Alliance of Canada, and Advanced Research Computing at the University of British Columbia.

References

[1] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from

diffusion models improves imagenet classification. *Transactions on Machine Learning Research*, 2023. 2, 3

[2] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. *International Conference on Learning Representations*, 2021. 2, 3

[3] 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/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2, 5

[4] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. *Proceedings of the IEEE International Conference on Computer Vision*, 2022. 2, 3

[5] Weiya Chen, Chenchen Yu, Chenyu Tu, Zehua Lyu, Jing Tang, Shiqi Ou, Yan Fu, and Zhidong Xue. A survey on hand pose estimation with wearable sensors and computer-vision-based methods. *Sensors*, 2020. 2

[6] Subhabrata Choudhury, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Unsupervised part discovery from contrastive reconstruction. *Advances in Neural Information Processing Systems*, 2021. 5, 6

[7] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv Preprint*, 2022. 2

[8] Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero-shot classifiers. *International Conference on Learning Representations*, 2023. 2, 3

[9] Edo Collins, Radhakrishna Achanta, and Sabine Susstrunk. Deep feature factorization for concept discovery. In *Proceedings of the European Conference on Computer Vision*, 2018. 6

[10] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018. 5

[11] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time, 2022. 5

[12] Xingzhe He, Bastian Wandt, and Helge Rhodin. Latentkeypointgan: Controlling gans via latent keypoints. *International Conference on Learning Representations*, 2021. 1, 2, 6

[13] Xingzhe He, Bastian Wandt, and Helge Rhodin. Autolink: Self-supervised learning of human skeletons and object outlines by linking keypoints. In *Advances in Neural Information Processing Systems*, 2022. 1, 2, 3, 6, 7, 8

[14] 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. 2, 6

[15] Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised semantic correspondence using stable diffusion. *Advances in Neural Information Processing Systems*, 2023. 2, 3, 4, 5

[16] Wei-Chih Hung, Varun Jampani, Sifei Liu, Pavlo Molchanov, Ming-Hsuan Yang, and Jan Kautz. Scops: Self-supervised co-part segmentation. In *Conference on Computer Vision and Pattern Recognition*, 2019. 2, 6

[17] 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*, 2013. 2, 5

[18] Marwa Jabberi, Ali Wali, Bidyut Baran Chaudhuri, and Adel M Alimi. 68 landmarks are efficient for 3d face alignment: what about more? 3d face alignment method applied to face recognition. *Multimedia Tools and Applications*, 2023. 1

[19] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. *Advances in Neural Information Processing Systems*, 2018. 1, 2, 6

[20] 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*, 2020. 3, 6

[21] You-Yi Jau, Rui Zhu, Hao Su, and Manmohan Chandraker. Deep keypoint-based camera pose estimation with geometric constraints. In *International Conference on Intelligent Robots and Systems*, 2020. 2

[22] Le Jiang, Caleb Lee, Divyang Teotia, and Sarah Ostadabbas. Animal pose estimation: A closer look at the state-of-the-art, existing gaps and opportunities. *Computer Vision and Image Understanding*, 2022. 2

[23] Yuhe Jin, Weiwei Sun, Jan Hosang, Eduard Trulls, and Kwang Moo Yi. Tusk: Task-agnostic unsupervised keypoints. *Advances in Neural Information Processing Systems*, 2022. 2

[24] Aliasghar Khani, Saeid Asgari Taghanaki, Aditya Sanghi, Ali Mahdavi Amiri, and Ghassan Hamarneh. Slime: Segment like me. *arXiv Preprint*, 2023. 2, 3

[25] Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015. 4

[26] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 3

[27] Wu Liu, Qian Bao, Yu Sun, and Tao Mei. Recent advances of monocular 2d and 3d human pose estimation: A deep learning perspective. *ACM Computing Surveys*, 2022. 1

[28] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015. 2, 5

[29] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 2, 5

[30] Dominik Lorenz, Leonard Bereska, Timo Milbich, and Bjorn Ommer. Unsupervised part-based disentangling of object shape and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 1, 2, 5, 6

[31] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 2004. 1

[32] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. *arXiv Preprint*, 2023. 1

[33] Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems*, 2023. 2, 3

[34] Giorgia Marullo, Leonardo Tanzi, Pietro Piazzolla, and Enrico Vezzetti. 6d object position estimation from 2d images: a literature review. *Multimedia Tools and Applications*, 2023. 2

[35] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 3

[36] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3

[37] Aditya Nandy, Chenru Duan, and Heather J Kulik. Audacity of huge: overcoming challenges of data scarcity and data quality for machine learning in computational materials discovery. *Current Opinion in Chemical Engineering*, 2022. 2

[38] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 2021. 3

[39] OpenAI. Gpt-4 technical report, 2023. 2

[40] 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*, 2018. 3

[41] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *International Conference on Learning Representations*, 2022. 2, 3

[42] Linhao Qu, Siyu Liu, Xiaoyu Liu, Manning Wang, and Zhijian Song. Towards label-efficient automatic diagnosis and analysis: a comprehensive survey of advanced deep learning-based weakly-supervised, semi-supervised and self-supervised techniques in histopathological image analysis. *Physics in Medicine & Biology*, 2022. 2

[43] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv Preprint*, 2022. 2, 3

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2, 3, 4, 5, 7

[45] Helena Russello, Rik van der Tol, and Gert Kootstra. T-leap: Occlusion-robust pose estimation of walking cows using temporal information. *Computers and Electronics in Agriculture*, 2022. 2

[46] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 2022. 2, 3

[47] 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*, 2021. 6

[48] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 2022. 2

[49] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *Conference on Computer Vision and Pattern Recognition*, 2019. 2

[50] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Advances in Neural Information Processing Systems*, 2019. 2, 5

[51] Aliaksandr Siarohin, Subhankar Roy, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Motion-supervised co-part segmentation. In *International Conference on Pattern Recognition*, 2021. 5, 6

[52] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 1

[53] Jennifer J Sun, Serim Ryou, Roni H Goldshmid, Brandon Weissbourd, John O Dabiri, David J Anderson, Ann Kennedy, Yisong Yue, and Pietro Perona. Self-supervised keypoint discovery in behavioral videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 6

[54] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 2023. 2, 3

[55] 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*, 2017. 1, 2, 6

[56] Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse, attend, and segment: Unsupervised zero-shot segmentation using stable diffusion. *arXiv Preprint*, 2023. 2, 3

[57] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv Preprint*, 2023. 2

[58] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, California Institute of Technology, 2011. 2, 5

[59] Jinglong Wang, Xiawei Li, Jing Zhang, Qingyuan Xu, Qin Zhou, Qian Yu, Lu Sheng, and Dong Xu. Diffusion model is secretly a training-free open vocabulary semantic segmenter. *arXiv Preprint*, 2023. 2, 3

[60] Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. Tracking everything everywhere all at once. *International Conference on Computer Vision*, 2023. 1

[61] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. *arXiv Preprint*, 2023. 2, 3

[62] Yue Wu and Qiang Ji. Facial landmark detection: A literature survey. *International Journal of Computer Vision*, 2019. 2

[63] Changming Xiao, Qi Yang, Feng Zhou, and Changshui Zhang. From text to mask: Localizing entities using the attention of text-to-image diffusion models. *arXiv Preprint*, 2023. 2, 3

[64] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 3

[65] Lumin Xu, Sheng Jin, Wentao Liu, Chen Qian, Wanli Ouyang, Ping Luo, and Xiaogang Wang. Zoomnas: searching for whole-body human pose estimation in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[66] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 2023. 2, 3

[67] 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/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 5, 6

[68] Yanping Zhang, Qiaokang Liang, Kunlin Zou, Zhengwei Li, Wei Sun, and Yaonan Wang. Self-supervised part segmentation via motion imitation. *Image and Vision Computing*, 2022. 6

[69] Ce Zheng, Wenhan Wu, Chen Chen, Taojiannan Yang, Sijie Zhu, Ju Shen, Nasser Kehtarnavaz, and Mubarak Shah. Deep learning-based human pose estimation: A survey. *ACM Computing Surveys*, 2023. 2