Weak-Shot Keypoint Estimation via Keyness and Correspondence Transfer

Junjie Chen[1] **Zeyu Luo**[1] **Zezheng Liu**[1] **Wenhui Jiang**[1] **Li Niu**[2]* **Yuming Fang**[1]
[1]Jiangxi University of Finance and Economics
[2]Shanghai Jiao Tong University
{chenjunjie,2202320619,2202426552,wenhui}@jxufe.edu.cn,
ustcnewly@sjtu.edu.cn, fangyuming@jxufe.edu.cn

Abstract

Keypoint estimation is a fundamental task in computer vision, but generally requires large-scale annotated data for training. Few-shot and unsupervised keypoint estimation are prevalent economical paradigms, but the former still requires annotations for extensive novel classes while the latter only supports for single class. In this paper, we focus on the task of weak-shot keypoint estimation, where multiple novel classes are learned from unlabeled images with the help of labeled base classes. The key problem is what to transfer from base classes to novel classes, and we propose to transfer keyness and correspondence, which essentially belong to comparing entities and thus are class-agnostic and class-wise transferable. The keyness compares which pixel in the local region is more key, which can guide the keypoints of novel classes to move towards the local maximum (*i.e.*, obtaining precise keypoints). The correspondence compares whether the two pixels belongs to the same semantic part, which can activate the keypoints of novel classes by reinforcing the consistency between two paired images. Extensive experiments and analyses on large-scale benchmark MP-100 demonstrate our effectiveness.

1 Introduction

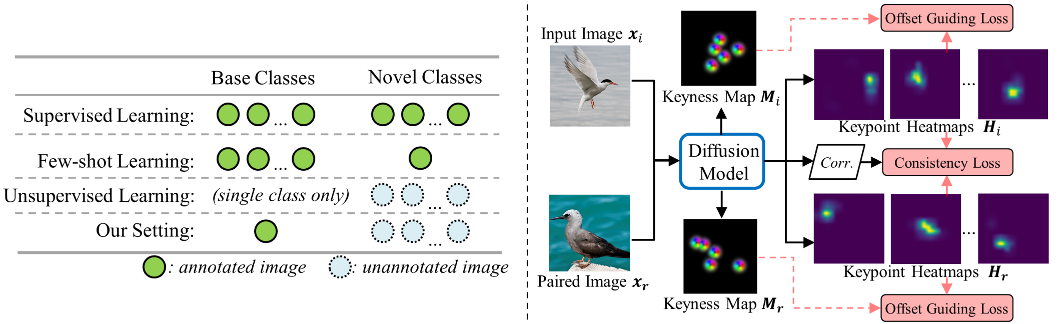
Keypoint estimation is a fundamental computer vision task and has extensive applications in real world, including intelligent interaction [65], behavioural analysis [2] and augmented reality [46]. Although existing keypoint estimation methods [1, 5, 53] have achieved great success, they usually require large-scale annotated data of all classes for fully supervised learning. As a consequence, the expensive annotation dramatically limits category-wise expansion and wider application.

Few-shot learning and unsupervised learning are two prevalent paradigms to economize annotations. As illustrated in Fig. 1 (a), few-shot keypoint estimation [69, 59, 9] greatly reduces the number of labeled images of novel classes (*i.e.*, novel object categories), and unsupervised keypoint estimation [64, 79, 17] reduces the annotations of target class. However, few-shot methods still requires non-negligible annotations for each novel class, and thus the annotation cost can become substantially high with a large number of novel classes. Unsupervised methods only focuses on single class, and is difficult to promise that the discovered keypoints are desired. Thus, it would be more practicable if we can learn desired keypoints from unlabeled images for multiple novel classes.

In this paper, we follow the paradigm of weak-shot learning [7, 38, 8, 26] and explore the task of weak-shot keypoint estimation. Weak-shot learning has achieved promising success in economizing annotations for various tasks, including classification [7], detection [38, 80, 32], segmentation [8, 22, 26, 3, 81], and so on [61]. Specifically in keypoint estimation, we would like to learn keypoints

*Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).



(a) Comparison of various task settings.

(b) Overview of our framework.

Figure 1: (a) Data comparison among fully supervised learning, few-shot learning, unsupervised learning and our weak-shot learning. Our setting is more economical and practicable. (b) Our proposed framework transferring keyness and correspondence from base classes to support the unsupervised learning of multiple novel classes with offset guiding loss and consistency loss.

for novel classes from unlabeled images with the support of labeled images of base classes, as shown in Fig. 1 (a). Intuitively, we leverage the knowledge transferred from base classes to facilitate the unsupervised learning of multiple novel classes. Therefore, our setting integrates few-shot and unsupervised keypoint estimation to alleviate their drawbacks (*i.e.*, annotations on novel classes, single-class discovery, and uncontrollable discovery).

In weak-shot keypoint estimation, the key problem is what to transfer to facilitate the unsupervised learning of novel classes. A representative paradigm [63, 7] in previous works is to transfer the comparing entities (or relation), which is class-agnostic and class-wise transferable. Different from the image-level comparison in [63, 7], here we explore to compare pixels to facilitate keypoint localization. Specifically, we conduct the following two types of comparisons: intra-image comparison (*i.e.*, keyness) and inter-image comparison (*i.e.*, correspondence).

Respectively, for intra-image comparison, we propose to compare the pixels in local regions, since "keypoint" can be understood as the most key point within a local region or the point that are more key compared to nearby ones (dubbed as "keyness"). Given a keypoint Gaussian map, we compute the two-dimensional derivative as its keyness map, where each pixel is a 2-D vector indicating the more key direction. We learn such keyness from the annotations of base classes, and employ the transferred keyness to the keypoints of novel classes to move towards the local maximum (*i.e.*, by offset guiding loss). For inter-image comparison, we propose to compare the pixels in two images belonging to the same class, since "keypoint" should be semantic consistency across images. Given two images from the same class, we compute the bipartite pairwise similarities to derive correspondences as in [44]. Analogous to keyness transfer, we class-wise transfer correspondences to activate the keypoints of novel classes (*i.e.*, by consistency loss).

Based on aforementioned keyness and correspondence transfer, we design a well-tailored and effective model for weak-shot keypoint estimation, as shown in Fig. 1 (b). In particular, we follow [18] and build our model upon pretrained diffusion model [55], because unsupervised keypoint estimation is quite challenge, let alone in the multi-class scenario. The training data contains image pairs from both base classes and novel classes. For the image pair from base classes, we enable the loss terms to learn keyness and correspondence. Otherwise, we switch the loss terms and apply estimated keyness and correspondence to support the unsupervised learning of novel classes.

We conduct comprehensive experiments and in-depth analyses on the large-scale multi-category pose dataset MP-100 [69]. Because the model has never accessed the GT keypoints of novel classes, we adopt the matching-based evaluation [13] and regression-based evaluation [18] for quantitative analysis. Our contributions could be summarized as follows:

(1) We are the first to explore weak-shot keypoint estimation, where we could use the knowledge transferred from base classes to facilitate the unsupervised learning of multiple novel classes.

(2) We propose a well-tailored framework to learn keyness and correspondence from base classes, and transfer them to provide effective supervision for the unsupervised learning of novel classes.

(3) Extensive experiments on the large-scale dataset demonstrate the effectiveness of our method.

2 Related Works

Fully-Supervised Pose Estimation. Most existing methods focus on fully-supervised learning for specific class, including human [1, 29], animals [5, 27], and vehicles [53, 60]. Technically, these methods could be roughly classified into regression-based [49, 28, 35, 36, 15], heatmap-based [75, 10, 20, 21, 24, 47], and transformer-based [45, 70, 33, 58, 74] methods. For example, DEKR [15] employed adaptive convolutions through spatial transformer to activate keypoint regions and learn representations for more accurate keypoints. Above works have achieved great success for estimating pose of specific classes, but they require abundant annotations and are inapplicable for novel classes. In this paper, we focus on estimating poses for multiple novel classes.

Zero-Shot and Few-Shot Keypoint Estimation. Zero-shot keypoint estimation locates keypoints for novel classes using descriptions, and existing methods generally learn a mapping from description to keypoints [71, 72, 56, 42, 76, 77, 78]. Our task setting is closer to few-shot keypoint estimation, which locates keypoints for novel classes using a few examples. Early methods focus on specific domains, *e.g.*, facial images [4, 67], clothing images [14], or animal images [62]. Recently, POMNet [69] introduced the task of category-agnostic keypoint estimation, and proposed a matching framework to retrieve results based on few-shot. Later, CapeFormer [59] further enhanced the similarity modeling and proposed to refine the coarse keypoints via a transformer decoder. Meanwhile, Lu *et al.* [40] explored a more flexible few-shot scenario to learn novel/base classes and novel/base keypoints. Although existing methods [43, 41, 19, 48, 30, 31, 54, 52] have great promoted few-shot keypoint estimation, they requires non-negligible annotations for novel classes. In contrast, we propose to harness diffusion models and learn novel classes from unannotated images.

Unsupervised Keypoint Estimation. Unsupervised keypoint estimation only require unlabeled images to discovery keypoints for single class. Overall, the high-level idea of most existing method is to design keypoint bottleneck in image reconstruction [64, 17, 16, 39] or video reconstruction [23, 57]. Recently, Hedlin *et al.* [18] proposed to employ attention maps of text embeddings in pretrained diffusion models as keypoint heatmaps, and optimize text embeddings with localization loss and equivariance loss for unsupervised keypoint estimation. Although above methods could discover keypoints from unlabeled images, most of them are difficult to discover desired keypoint for multiple classes. In this work, we take the inspiration of [18] and propose a well-tailored framework to transfer keypoint prompts and correspondences for weak-shot keypoint estimation.

Weak-Shot Learning Deep learning methods generally require large-scale labeled data, and thus the demand of reducing annotations is extensive in various tasks. To this end, weak-shot learning, *i.e.*, learning weakly labeled novel classes with the support of fully labeled base classes, has been explored in classification [7, 51, 50], detection [38, 80], semantic segmentation [8, 81], instance segmentation [22, 26], and so on [61]. To name a few, SimTrans [7] proposed to transfer the pair-wise similarity to de-noise the web data of novel classes. RETAB [81] proposed to transfer affinity and boundary to expand seed mask to semantic mask. Here we explore weak-shot keypoint estimation, and propose to transfer keyness and correspondence to support multiple novel classes.

3 Method

3.1 Task Setting of Weak-Shot Keypoint Estimation

In weak-shot keypoint estimation, we would like to learn keypoints for novel classes from unlabeled images with the support of labeled images of base classes. Specifically, in the training stage, there are labeled samples from base classes available, and the i-th labeled sample consists of image $\boldsymbol{x}_i \in \mathbb{R}^{3\times H_{full}\times W_{full}}$ and GT keypoints $\boldsymbol{P}_i^* \in \mathbb{R}^{K_i^*\times 2}$. Note that, K_i^* is the keypoint number, and the images from different classes could have different keypoint numbers. For ease of process, we pad all keypoints to a unified number K with visibilities, *i.e.*, $[\boldsymbol{P}_i^*; \boldsymbol{V}_i^*] \in \mathbb{R}^{K\times(2+1)}$. There are also unlabeled samples for novel classes available, which have no overlap with base classes. In the test stage, the model should locate the keypoints for images from multiple novel classes.

3.2 The Forward Pipeline of Our Framework

As shown in Fig. 2, we propose a compact framework for learning and transferring keyness and correspondence. In the training stage, each mini-batch contains images randomly sampled from

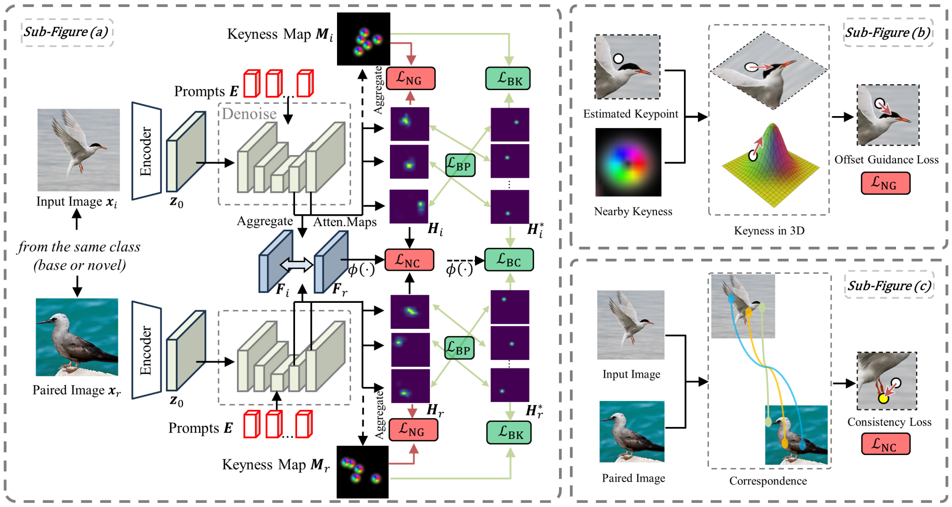


Figure 2: (a) The detailed illustration of our framework in the training stage. Given image pairs from the same base or novel class, we employ keypoint prompts to estimate keypoint heatmaps and aggregate feature maps to estimate keyness and extract correspondences. For labeled image pairs from base classes, we learn keypoint prompts, keyness and correspondence via $\mathcal{L}_{\mathrm{BP}}$, $\mathcal{L}_{\mathrm{BK}}$ and $\mathcal{L}_{\mathrm{BC}}$. For unlabeled image pairs from novel classes, we transfer them to learn valid keypoints via $\mathcal{L}_{\mathrm{NG}}$, $\mathcal{L}_{\mathrm{NC}}$ and $\mathcal{L}_{\mathrm{NU}}$. The sub-figures (b) and (c) illustrate the insight of $\mathcal{L}_{\mathrm{NG}}$ and $\mathcal{L}_{\mathrm{NC}}$.

both base classes and novel classes. Additionally, we pair each training image $\boldsymbol{x}_i$ with a reference image $\boldsymbol{x}_r$ (i.e., paired image) from the same class, and apply respective supervisions on image pairs from base or novel classes. To estimate keypoints, we follow [18] and maintain N learnable keypoint prompts $\boldsymbol{E} \in \mathbb{R}^{N\times D}$, which correspond to text/query embeddings and set as $N = 200$ by default. Different from the single-class scenario in [18], our keypoint prompts are shared by all images and all classes, and thus could learn transferrable knowledge under our match-based supervision.

Firstly, for each image in the input pair, we obtain keypoint heatmaps according to keypoint prompts $\boldsymbol{E}$. We denote the $\Phi_l^c(\cdot)$ and $\Psi_l^c(\cdot)$ as the c-th head and the l-th linear layers of the U-Net in the transformer part of pretrained Diffusion model. We compute the query as $\boldsymbol{Q}_l^c = \Phi_l^c(\boldsymbol{z}_{t=1}) \in \mathbb{R}^{H\times W\times D_l}$, where $\boldsymbol{z}$ is the latent embedding mapped from image $\boldsymbol{x}$, and we use $t = 1$ as suggested by [18]. We compute the key from keypoint prompts $\boldsymbol{K}_l^c = \Psi_l^c(\boldsymbol{E}) \in \mathbb{R}^{N\times D_l}$. Then we obtain keypoint heatmaps $\boldsymbol{H} \in \mathbb{R}^{H\times W\times N}$ by collecting cross-attention maps from various layers:

$$\boldsymbol{H} = \mathbb{E}_{l=7..10}\left[\mathbb{E}_c\left[\mathrm{softmax}_c(\boldsymbol{Q}_l^c \cdot \boldsymbol{K}_l^c/\sqrt{D_l})\right]\right]. \tag{1}$$

In this way, we can obtain the estimated keypoint heatmaps $\boldsymbol{H}_i \in \mathbb{R}^{H\times W\times N}$ and $\boldsymbol{H}_r \in \mathbb{R}^{H\times W\times N}$ for images $\boldsymbol{x}_i$ and $\boldsymbol{x}_r$ using keypoint prompts $\boldsymbol{E}$. More details could be found in [18].

Secondly, we aggregate the feature maps from above diffusion process and obtain hyperfeatures $\boldsymbol{F} \in \mathbb{R}^{H\times W\times D}$ for estimating keyness. Concretely, $\boldsymbol{F} = \sum_s^S \sum_l^L w_{l,s} \cdot f_l(F_{l,s})$, where $f_l(F_{l,s})$ is squeezed feature map, S is the number of subsampled timesteps selected from the diffusion timesteps, and $w_{l,s}$ denotes the mixing weights for the l-th layer at the i-th timestep, and we use the architecture of [44]. Since keyness compares the pixels in local regions, we intuitively estimate it via conv. layers with act. functions. We denote the estimated keyness map as $\boldsymbol{M} \in \mathbb{R}^{2\times H\times W}$, where each pixel $\boldsymbol{M}[x, y]$ is a 2D vector indicating the direction to the most key pixel.

Finally, we use the hyperfeatures of the paired images to compute the semantic correspondences, denoted as $\boldsymbol{F}_i$ and $\boldsymbol{F}_r$ respectively for images $\boldsymbol{x}_i$ and $\boldsymbol{x}_r$. Specifically, we compute the mapping $\phi(\cdot)$ from keypoint to its corresponding keypoint in the other image as:

$$\phi(\boldsymbol{p}_i)_{\to r} = \arg\max_{\boldsymbol{p}_r} S(\boldsymbol{F}_i[\boldsymbol{p}_i], \boldsymbol{F}_r[\boldsymbol{p}_r]), \tag{2}$$

where $\boldsymbol{p}_i$ and $\boldsymbol{p}_r$ are points on two paired images and $S(\cdot,\cdot)$ denotes cosine similarity. As in [44], we only need to compute the similarity between all keypoints on $\boldsymbol{x}_i$ and all pixels on $\boldsymbol{x}_r$, and thus the computation is relatively efficient and affordable.

Given above estimated keypoint heatmaps $(\boldsymbol{H}_i, \boldsymbol{H}_r)$, keyness maps $(\boldsymbol{M}_i, \boldsymbol{M}_r)$, and correspondence $\phi(\cdot)$ for image pair $(\boldsymbol{x}_i, \boldsymbol{x}_r)$, we switch the loss terms for learning from base classes and transferring to novel classes. For labeled image pairs from base classes, we learn keypoint prompts, keyness and correspondence via $\mathcal{L}_{\text{BP}}$, $\mathcal{L}_{\text{BK}}$ and $\mathcal{L}_{\text{BC}}$. For unlabeled image pairs from novel classes, we transfer them to learn valid keypoints via $\mathcal{L}_{\text{NG}}$, $\mathcal{L}_{\text{NC}}$ and $\mathcal{L}_{\text{NU}}$. We introduce the loss terms as following.

3.3 Learning from Base Classes

To learn keypoint prompts from base classes, we apply following loss over the keypoint heatmaps:

$$\mathcal{L}_{\text{BP}} = \frac{1}{K}\sum_{k}^{K} \left\| \boldsymbol{H}_i[\delta(k)] - \boldsymbol{H}_i^*[k] \right\|^2 + \left\| \boldsymbol{H}_r[\delta(k)] - \boldsymbol{H}_r^*[k] \right\|^2, \tag{3}$$

where $\boldsymbol{H}_i^*[k]$ is the k-th GT keypoint Gaussian map generated as in [69]. Besides, the max value for $\boldsymbol{H}_i^*[k]$ is 1 if the i-th keypoint is visible, otherwise 0. $\delta(\cdot)$ is the matching function solved as in [6, 11, 9]. Although it is feasible to directly apply the matching-loss to each image respectively as in [9, 11, 6], we propose to conduct joint matching over two paired images in light of the semantic consistency of keypoints. Specifically, we calculate the cost $\boldsymbol{C} \in \mathbb{R}^{N\times K}$ by considering the cost over paired two images *i.e.*, $\boldsymbol{C}[n,k] = \|\boldsymbol{H}_i[n] - \boldsymbol{H}_i^*[k]\|^2 + \|\boldsymbol{H}_r[n] - \boldsymbol{H}_r^*[k]\|^2$. Such loss can adaptively embed knowledge into the same matched prompt, leading to transferability.

To learn valid keyness, we apply the loss analogous to Eqn. 3 due to the same "map" representation:

$$\mathcal{L}_{\text{BK}} = \left\| \boldsymbol{M}_i - \boldsymbol{M}_i^* \right\|^2 + \left\| \boldsymbol{M}_r - \boldsymbol{M}_r^* \right\|^2, \tag{4}$$

where $\boldsymbol{M}^* \in \mathbb{R}^{2\times H\times W}$ indicate the GT keyness map of input image or paired image, where each pixel $\boldsymbol{M}[x,y]$ is a 2D vector indicating the direction to the most key pixels. We compute the derivative of GT keypoint Gaussian map to obtain GT keyness map, *i.e.*, the GT of $\boldsymbol{M}[x,y]$ is the derivative of the GT keypoint Gaussian map at $[x,y]$. To learn valid correspondences, we compute cosine similarity between every possible pair of points and apply a symmetric cross entropy loss according to GT keypoints as [44], which is denoted as $\mathcal{L}_{\text{BC}}$.

Note that, the names of above loss terms (*i.e.*, $\mathcal{L}_{\text{BP}}$, $\mathcal{L}_{\text{BK}}$ and $\mathcal{L}_{\text{BC}}$) all begin with **B**, indicating that they are enabled only when the image pair comes from **B**ase classes.

3.4 Transferring to Novel Classes

For the unlabeled image pair $\boldsymbol{x}_i$ and $\boldsymbol{x}_r$ belonging to novel classes, we transfer and reuse the keypoint prompts $\boldsymbol{E}$ to estimate keypoint heatmaps $\boldsymbol{H}_i$ and $\boldsymbol{H}_r$. Note that such estimations without accessing any data of novel classes satisfyingly locate keypoints for multiple novel classes, which will be demonstrated in Sec. 4.3. Nevertheless, the domain gap between base classes and novel classes inevitably matters. To bridge the domain gap, we propose to apply offset guiding loss and consistency loss according to the transferred keyness and correspondences.

Specifically, we firstly convert the heatmaps $\boldsymbol{H}_i$ and $\boldsymbol{H}_r$ to the keypoint coordinates and visibilities $[\boldsymbol{P}_i; \boldsymbol{V}_i]$ and $[\boldsymbol{P}_r; \boldsymbol{V}_r] \in \mathbb{R}^{N\times(2+1)}$ via the soft argmax and max operators in [59]. Given the transferred keyness map, we guide the coordinates with following loss:

$$\mathcal{L}_{\text{NG}} = \frac{1}{N}\sum_{k}^{N} \left\| \boldsymbol{P}_i[k] - dt(\boldsymbol{P}_i[k] + \boldsymbol{M}_i[\boldsymbol{P}_i[k]]) \right\|^2 + \left\| \boldsymbol{P}_r[k] - dt(\boldsymbol{P}_r[k] + \boldsymbol{M}_r[\boldsymbol{P}_r[k]]) \right\|^2, \tag{5}$$

where $\boldsymbol{M}_i[\boldsymbol{P}_i[k]]$ is an offset vector indicating the direction for $\boldsymbol{P}_i[k]$ to be more key, and $dt(\cdot)$ detaches and blocks the gradient to provide stable supervision. By such offset guiding loss, the estimated keypoints could be supervised to gradually offset to more key positions.

For the consistency loss, we reuse the aggregation network to obtain pseudo-points $\widetilde{\boldsymbol{P}}_r \in \mathbb{R}^{N\times 2}$ by corresponding $\boldsymbol{P}_i$ from image $\boldsymbol{x}_i$ to image $\boldsymbol{x}_r$, *i.e.*, $\widetilde{\boldsymbol{P}}_r = \phi(\boldsymbol{P}_i)_{\to r}$. We obtain pseudo-points $\widetilde{\boldsymbol{P}}_i$ similarly. Then, we use below consistency loss to facilitate the keypoint learning of novel classes:

$$\mathcal{L}_{\text{NC}} = \sum_{n}^{N} \boldsymbol{V}_r[n] \cdot \left\| \boldsymbol{H}_i[n] - \mathcal{H}(\widetilde{\boldsymbol{P}}_i[n]) \right\|^2 + \boldsymbol{V}_i[n] \cdot \left\| \boldsymbol{H}_r[n] - \mathcal{H}(\widetilde{\boldsymbol{P}}_r[n]) \right\|^2, \tag{6}$$

where $\boldsymbol{V}_r[n]$ is the visibility of the n-th pseudo-point blocking wrong supervision from invisible pseudo-point, and $\mathcal{H}(\cdot)$ is a function mapping keypoint coordinates to Gaussian heatmap in [69]. By our consistency loss, the keypoint heatmaps estimated on the unlabeled images of novel classes are regularized to approximate to the corresponding keypoints on the paired images.

Besides, we apply the equivariance and localization loss to complement our unsupervised learning on novel classes, which are proposed in [18] for single-class unsupervised learning. We use the same configuration and denote as $\mathcal{L}_{\text{NU}}$. Note that, the names of above losses (*i.e.*, $\mathcal{L}_{\text{NG}}$, $\mathcal{L}_{\text{NC}}$ and $\mathcal{L}_{\text{NU}}$) all begin with **N**, indicating that they are enabled when the image pair comes from **N**ovel classes.

3.5 Summary of Training and Inference

Our framework is end-to-end learnable, which can jointly learn keypoint prompts, keyness and correspondence from base classes and transfer them to novel classes. In the training stage, our full loss for the i-th image $\boldsymbol{x}_i$ can be summarized as:

$$\mathcal{L}_{\text{FULL}} = \mathbf{1}_{\boldsymbol{x}_i \in \mathcal{B}} \cdot (\mathcal{L}_{\text{BP}} + \mathcal{L}_{\text{BK}} + \alpha \cdot \mathcal{L}_{\text{BC}}) + \mathbf{1}_{\boldsymbol{x}_i \in \mathcal{N}} \cdot (\beta \cdot \mathcal{L}_{\text{NG}} + \gamma \cdot \mathcal{L}_{\text{NC}} + \lambda \cdot \mathcal{L}_{\text{NU}}). \tag{7}$$

where $\mathbf{1}_{(\cdot)}$ is the indicator function, $\mathcal{B}$ (*resp.*, $\mathcal{N}$) denotes the image set of base classes (*resp.*, novel classes). Hence, we respectively apply the supervisions for the base images and novel images. As for the loss balancing, we find that $\mathcal{L}_{\text{BP}}$ balance well with $\mathcal{L}_{\text{BK}}$, may due to the same map representation. Besides, $\alpha = 0.1$, $\beta = 0.2$, $\gamma = 0.1$ and $\lambda = 0.5$ are hyper-parameters for balancing the other losses.

In the test stage, we remove the additional modules estimating keyness and correspondence, which are only used to provide effective supervisions for learning novel classes without annotations. Compared with the related work [18] in unsupervised keypoint estimation, our model keeps the same architecture for inference and has extra capacity to estimate keypoints for multiple unlabeled classes.

4 Experiments

4.1 Dataset and Implementation Details

We follow the related works [69, 59, 9] to conduct our experiments on MP-100 dataset [69], which is the most prevalent benchmark dataset and covers 100 classes within 8 super-classes. MP-100 dataset contains keypoint numbers ranging from 8 to 68 across different classes. We follow the dataset splits in [69], *i.e.*, all classes are split into non-overlapping train/val/test sets with the ratio of $70 : 10 : 20$, and there are five random splits (S1-S5) to reduce the impact of randomness. Considering that our model is built on pretrained diffusion model, we focus on learning from a small-scale training set in this paper, similar to large model adaption [66]. Unless otherwise stated, we use 3 labeled images per base class and 30 unlabeled images per novel class by default.

In practice, the keypoint orders of different classes matter transfer learning. Specifically, the default orders in MP-100 are related, *e.g.*, the first keypoint of base/novel classes always corresponds to "left-eye". On the one hand, such relation requires that all class annotations keep explicit relation and thus severely limits class-wise extension. On the other hand, models may class-wise transfer by overfitting such unpractical relation. Therefore, we use the shuffled orders of each class in experiments, which has less limitation and thus more practical.

4.2 Evaluation and Metric

Similar to unsupervised keypoint estimation [18], our model have never accessed the GT annotations of novel classes, and thus we cannot directly compare the estimated keypoints with GT keypoints. Thus, we employ the following two kinds of evaluations.

Matching-based evaluation Considering that the number of GT keypoints varies across classes, we follow unsupervised semantic segmentation [13] to compare the estimation with GT based on matching. We employ Average Precision (AP) to compute the metric, because unmatched keypoints will reduce recall rate and thus reduce AP. Specifically, for the test image $\boldsymbol{x}_i$ with padded GT keypoints $[\boldsymbol{P}_i^*; \boldsymbol{V}_i^*] \in \mathbb{R}^{K \times (2+1)}$ come from novel class, the model produces padded keypoints $[\boldsymbol{P}_i; \boldsymbol{V}_i] \in \mathbb{R}^{N \times (2+1)}$. Then, we compute the matching cost over all test images in each class by:

$$\mathcal{C}_{n,k} = \sum_i \boldsymbol{V}_i^*[k] \cdot \mathbf{1}_{(|\boldsymbol{P}_i[n] - \boldsymbol{P}_i^*[k]| < T_p)} \cdot \mathbf{1}_{(\boldsymbol{V}_i[n] > T_v)}, \tag{8}$$

Table 1: Comparison with prior works. We report results on 5 dataset splits with matching-based and regression-based evaluation.

Method	MP-100 Split1		MP-100 Split2		MP-100 Split3		MP-100 Split4		MP-100 Split5		AVG	
	mAP ↑	L2 ↓	mAP ↑	L2 ↓	mAP ↑	L2 ↓	mAP ↑	L2 ↓	mAP ↑	L2 ↓	mAP ↑	L2 ↓
Sim.Base. [68]	14.4	60.8	12.9	65.3	13.3	63.5	13.1	64.4	14.2	61.3	13.6	63.1
GroupPose [37]	17.4	55.7	16.7	58.2	17.0	57.1	15.9	58.4	18.2	52.1	17.0	56.3
He *et al.* [16]	18.3	51.5	16.2	56.6	18.6	51.7	17.8	54.3	19.5	50.2	18.1	52.9
Hedlin *et al.* [18]	27.0	38.3	22.1	50.3	23.5	38.6	22.1	44.2	24.1	47.1	23.8	43.7
MetaPoint [9]	31.5	33.5	36.2	31.7	30.1	22.4	26.8	25.9	36.9	39.6	32.3	30.6
Ours	**70.4**	**18.4**	**61.3**	**26.2**	**60.7**	**19.6**	**61.7**	**21.3**	**58.9**	**24.3**	**62.6**	**22.0**
Oracle*	86.1	14.4	78.3	21.6	72.8	16.9	69.6	19.4	76.0	20.3	76.6	18.5

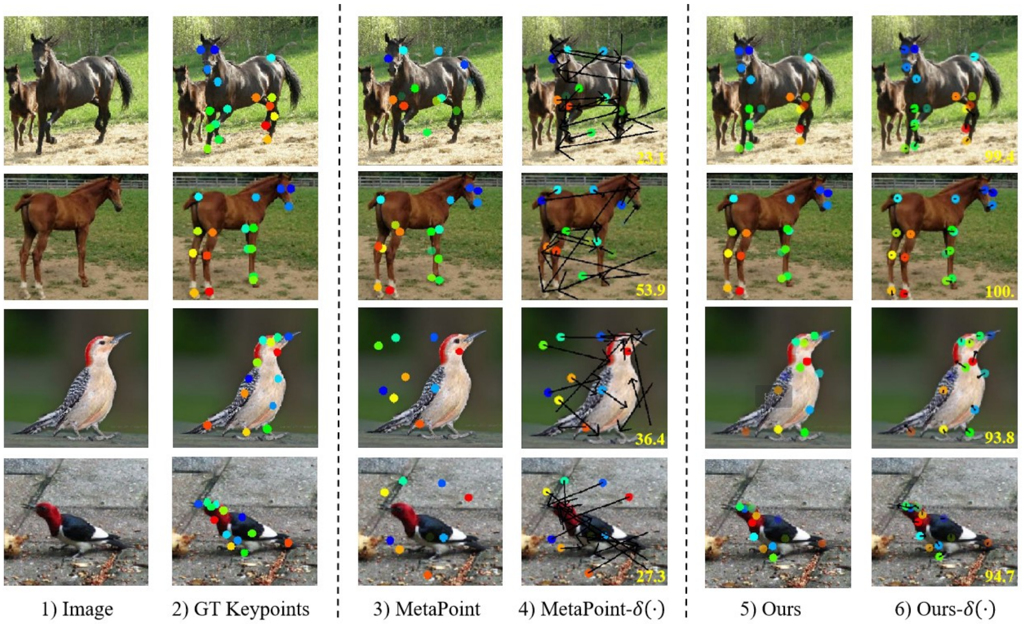


Figure 3: The qualitative comparison against the most competitive baseline. The first two columns show the image and GT keypoints. The mid two columns display the localized keypoints and matched keypoints of MetaPoint[9]. Note that we use transparency to show the visibility, the right-bottom number indicates the F1-score of keypoint of this sample, and the black arrows show the deviations to matched GT. The right two columns display the same results of our method.

where $\mathbf{1}_{(\cdot)}$ is the indicator function, T_p is a distance threshold as in PCK [73] and T_v is a score threshold for computing AP. We follow [69] to adopt $T_p = 0.2$, and use 9 scores from 0.1 to 0.9 to calculate AP. Finally, we obtain our overall metric by averaging the AP of all classes, *i.e.*, mAP↑.

Regression-based evaluation We also follow related work [18, 17, 79, 23, 39] to compare estimation with GT based on linear regression for better comprehensiveness. Specifically, a linear regression model is firstly trained to map estimated keypoints to padded GT keypoints on val images. Afterwards, in the formal evaluation, all the estimated keypoints on test images are mapped by above regressor, and the averaged L2↓ error to GT is employed as overall metric.

4.3 Comparison with Prior Works

Comparable baselines. As far as we know, there is no method designed for weak-shot keypoint estimation, and thus we adapt representative methods to our baselines, including Sim.Base. [68], GroupPose [37], He *et al.* [16], Hedlin *et al.* [18] and MetaPoint [9]. Specifically, Sim.Base. [68], He *et al.* [16], and Hedlin *et al.* [18] estimate the heatmaps of padded GT keypoints directly. GroupPose [37] regresses the coordinates of padded GT keypoints directly. MetaPoint [9] contains the first stage in [9], which learn potential keypoints with matching-based supervision. We supervise all methods with heatmap loss or coordinates loss on labeled images of base classes, and complement all methods with unsupervised loss (*i.e.*, equivariance loss and localization loss as in [18]) on unlabeled images of novel classes. We also set a baseline named Oracle* to show our upper-bound by training our model with all labeled images from both base class and novel classes.

Quantitative comparison. We summarize the results of above baselines and our method on five dataset splits in Tab. 1. Firstly, we could find that our method outperforms baselines by large margins in both evaluation metrics. Although Hedlin *et al.* [18] could utilize pretrained diffusion model, it cannot adaptively learn transferrable keypoint prompts and thus only achieves dissatisfactory performances. MetaPoint [9] can learn transferrable keypoints, but employs common matching-based loss on single image and also requires abundant annotated images for training. By contrast, our model can leverage pretrained diffusion model to adaptively learn and transfer keypoint prompts and correspondences, resulting in preferable performances. Besides, our model can reach about 75% of the upper-bound mAP represented by Oracle*, showing promising potential. Overall, our proposed model has relatively robust mAPs on different splits using both evaluation protocols.

Qualitative comparison. We select the most competitive baseline (*i.e.*, MetaPoint [9]) for qualitative comparison in Fig. 3. With threshold $T_v = 0.5$, we show the F1-score (harmonic mean of precision and recall) of matched keypoints in the right-bottom of sub-figures. In the top two rows, the baseline may locate right keypoints for horse, but fails to keep semantic consistency across images (by comparing two rows). In the bottom two rows, the baseline seems to fail to locate keypoints for bird, may due to the challenging domain gap between base and novel classes. Overall, the keypoints located by our model could better fit the object structures and keep semantic consistency across images, which is a general requirement in keypoint estimation. Therefore, our method could learn preferable keypoints in weak-shot keypoint estimation.

4.4 Comparison with Other Settings

Our weak-shot keypoint estimation is related with two prevalent settings (*i.e.*, few-shot and unsupervised setting) as shown in Fig. 1. Here we conduct setting-wise comparison to investigate our potential.

Comparison with few-shot setting. We set representative few-shot baselines including PoseAnything [19], MetaPoint [9] and SCAPE [34]. Considering that Peng *et al.* [52] requires test-time optimization, we don't include it as baseline. As shown in Tab. 2, few-shot methods [19, 9, 34] generally achieve inferior performances in our experiments, due to the limited training data (3 labeled images per base class described in Sec. 4.1). Although few-shot methods may outperform our method by using extra labeled support images from novel classes, the support images are not always available for all novel classes and will have expensive cost if we annotate for a large number of novel classes. Therefore, our method has the unique advantage on transferring knowledge upon diffusion models to facilitate the unsupervised learning of multiple novel classes, which is irreplaceable against few-shot methods.

Comparison with unsupervised setting. We set representative unsupervised baselines including AutoLink [17] and Hedlin *et al.* [18] . Considering that existing unsupervised methods [17, 18] mostly focus on single-class scenario, we conduct the comparison using two splits. In the split "Novel Bird", the novel classes are all the bird classes in MP-100, where the unsupervised methods [17, 18] could be well applied. In the "MP-100 Split1", the novel classes contain multiple classes, *e.g.*, Bed and Lion. As shown in Tab. 3, the unsupervised methods Hedlin *et al.* [18] dramatically degrades when changing the split from "Novel Bird" to "MP-100 Split1", probably because that they require to pre-define a keypoint number and thus are not intuitive to tackle multiple classes having different keypoint numbers. Our method outperforms unsupervised method Hedlin *et al.* [18] dramatically in both splits, because our method could simultaneously tackle multiple novel classes and leverage the transferred knowledge.

4.5 Method Analysis

Analysis on keyness transfer. We learn keyness from labeled base classes and transfer to unlabeled novel classes, and thus the transferability of keyness is a key factor of our method. We quantitatively evaluate the estimated keyness map on base classes and novel classes to investigate its transferability. Tab. 4 summarizes the averaged L2$\downarrow$ distances (after multiplying 100 as in [18] for better readability)

Table 2: Comparison with few-shot baselines.

Method	Extra annotated Novel Classes	MP-100 Split1 mAP↑	MP-100 Split1 L2↓	MP-100 Split2 mAP↑	MP-100 Split2 L2↓
PoseAnything [19]	√	41.7	27.7	32.7	36.0
MetaPoint [9]	√	42.3	27.2	33.1	35.8
SCAPE [34]	√	42.7	27.1	33.8	34.4
Ours	×	70.4	18.4	61.3	26.2

Table 3: Comparison with unsupervised baselines.

Method	Off-the-shelf Base Labels	Novel Bird mAP↑	Novel Bird L2↓	MP-100 Split1 mAP↑	MP-100 Split1 L2↓
AutoLink [17]	×	14.5	60.7	8.2	70.1
Hedlin *et al.* [18]	×	32.6	37.4	24.9	41.7
Ours	√	56.1	22.1	70.4	18.4

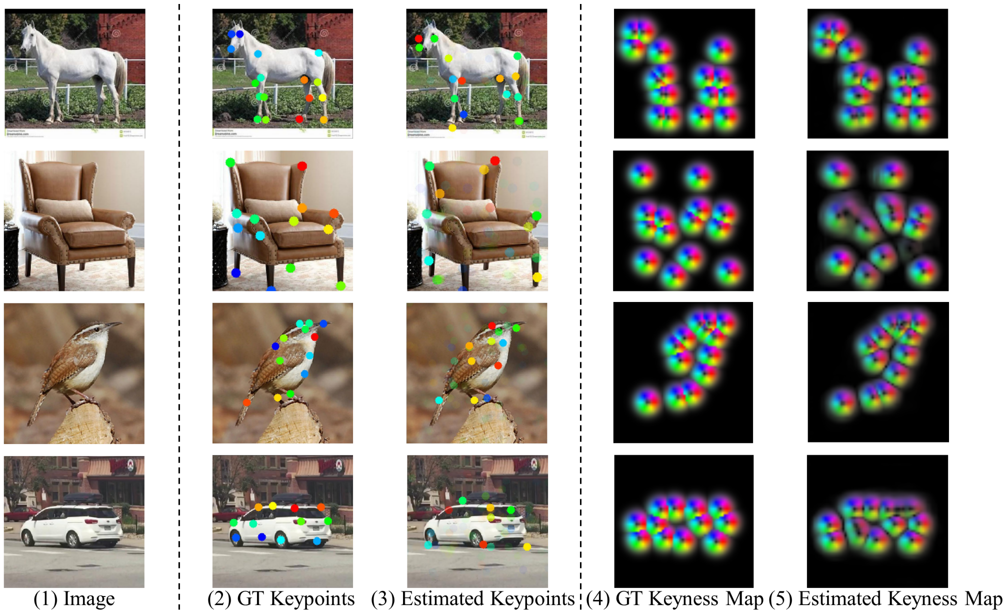

(1) Image (2) GT Keypoints (3) Estimated Keypoints (4) GT Keyness Map (5) Estimated Keyness Map

Figure 4: Qualitative analysis of estimated keypoints and keyness. The col (2,3) show the GT and estimated keypoints, while col (4,5) show the GT and estimated keyness. The keyness is shown by colors, where the hue indicates the direction angle and the saturation denotes the magnitude.

between estimated keyness maps and GT keyness maps for both base classes and novel classes on different splits. The averaged L2 distance before training is about 80 (*i.e.*, the lower bound of performance), and we can see that the L2 distances of base classes are relatively lower because of the full supervision. The L2 distances of novel classes without full supervision satisfactorily approximate to the results of base classes, indicating that the keyness are class-wise transferrable.

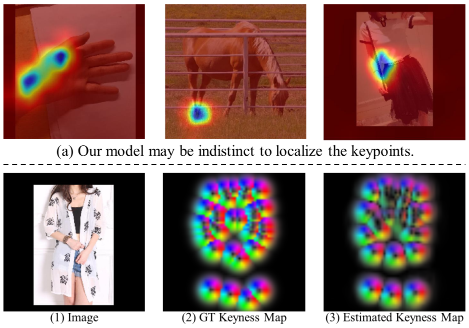
To further explore the transferability of keyness, we visualize the estimated keypoints and keyness map in Fig. 4 in the intermediate training stage to see whether the keyness could provide beneficial guidance. Generally, the estimated keyness maps are more approximate to their GT than the estimated keypoints. Therefore, some biased keypoints (*e.g.*, the right eye of horse in the top row) could be guided to offset to the most key pixels, leading to precise keypoints of novel classes. Although some cases (*e.g.*, the sofa in the 2-nd row) are still imperfect, we are the first to explore unsupervised learning of multiple classes, which is quite challenging and has never been explored before. Overall, the transferred keyness maps are beneficial, and provide valid supervision to learn precise keypoints.

Analysis on correspondence transfer. The learned and transferred correspondences provide beneficial regularization for the unsupervised learning of novel classes. We evaluate their PCKs according to the GT keypoints on $\boldsymbol{x}_i$ and the GT keypoints corresponded from $\boldsymbol{x}_r$ to $\boldsymbol{x}_i$. As shown in Tab. 4, the PCKs of base classes are high enough due to full supervision. And the PCKs of novel classes are relatively satisfactory, indicating that the transferred correspondences can provide valid regularization.

Ablation study. To study the contributions of our loss terms, we evaluate the combinations of our basic model, $\mathcal{L}_{\mathrm{NU}}$,$\mathcal{L}_{\mathrm{NC}}$ and $\mathcal{L}_{\mathrm{NG}}$. As shown in Tab. 5, enabling the unsupervised loss $\mathcal{L}_{\mathrm{NU}}$ could improve the mAP dramatically, indicating the primitively gains of using unannotated images. Solely enabling the consistency loss $\mathcal{L}_{\mathrm{NC}}$ promotes the performance more significantly. Furthermore, incor-

Table 4: The performance of keyness (L2↓) and correspondence (PCKs↑) evaluated for base classes and novel classes on five splits.

Method	Split	S1	S2	S3	S4	S5
Keyness	Base	47.2	43.6	45.3	42.2	44.8
	Novel	53.0	49.6	51.6	48.7	50.9
Correspondence	Base	89.1	94.0	93.3	94.8	93.5
	Novel	73.5	74.8	69.8	69.9	71.2


(a) Our model may be indistinct to localize the keypoints.

(1) Image

(2) GT Keyness Map

(3) Estimated Keyness Map

(b) Our model may fail to estimate fine-grained keyness.

Figure 5: The illustrations of two limitations.

Table 5: Ablation results of our proposed framework on two splits.

Basic	$\mathcal{L}_{\mathrm{NU}}$	$\mathcal{L}_{\mathrm{NC}}$	$\mathcal{L}_{\mathrm{NG}}$	MP-100 Split1		MP-100 Split2	
				mAP↑	L2↓	mAP↑	L2↓
√				60.2	22.4	51.5	29.0
√	√			65.8	20.1	56.7	27.5
√		√		66.5	20.2	57.6	27.9
√			√	67.7	19.5	58.2	27.3
√	√	√	√	70.4	18.4	61.3	26.2

Table 6: The statistical analysis of our method against the strongest baseline (*i.e.*, MetaPoint[9]). We summarize the "mean ± std" of mAP↑ and report the p-values against the baseline.

Dataset Split	MetaPoint [9]	Ours	P-Value
S1	31.5±0.61	70.4±0.53	0.0
S2	36.2±0.72	61.3±0.89	0.0
S3	30.1±0.73	60.9±0.78	0.0
S4	26.8±0.69	61.7±0.62	0.0
S5	36.9±0.71	58.9±0.56	0.0

porating $\mathcal{L}_{\mathrm{NG}}$ alone also demonstrates notable performance improvements. With all losses enabled, our method achieves the best results. Thus, our loss terms are effective and complementary.

4.6 Limitation Discussion

Our weak-shot keypoint estimation is prone to suffer from various issues due to insufficient annotations of multiple novel classes. We summarize three major problems found in practice. Firstly, our model is sometimes difficult to precisely localize keypoints via argmax on heatmaps, as shown in Fig. 5 (a). For example, in the first sub-figure, there are two maximums, and thus we may need better strategies to localize keypoints from heatmaps. Secondly, our estimated keyness map may fail to provide the directions to key directions, as shown in Fig. 5 (b). Our model is difficult to estimate precise or fine-grained keyness map in the chest area of the "long_sleeved_outwear", probably because the GT keypoints are too dense and crowded. Thirdly, we method may suffer from the imbalance between the image numbers of base classes and novel classes. Nevertheless, we are the first to explore such challenging cases, and we would like to address above limitations in future works.

4.7 Significance Test

In this section, we statistically analyse our framework against the most competitive baseline MetaPoint [9] on 5 splits of MP-100 dataset using 3 labeled images per base class and 30 unlabeled images per novel class. With random seeds ranging from 1 to 10, we run both methods for 10 times. Under the significance level 0.05, we conduct the significance test to show that our method is superior than MetaPoint [9]. We summarize the "mean ± std" of mAPs↑ results and p-values in various dataset splits in Tab. 6, where we could see that the p-values are all far below the significance level 0.05. Therefore, this statistical analysis demonstrates that the improvements of our proposed framework is statistically significant.

5 Conclusion

In this paper, we have explored weak-shot keypoint estimation, where multiple novel classes are learned from unlabeled images with the support of labeled base classes. We have proposed a novel framework transferring keyness and correspondence to facilitate the unsupervised learning of novel classes. By transferring keyness and correspondence, our framework has achieved favourable performance for weak-shot keypoint estimation. We have conducted comprehensive experiments on MP-100 dataset to demonstrate our effectiveness.

Acknowledgements

This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFE0210700, in part by the National Natural Science Foundation of China under Grants 62441203, 62311530101, 62132006, 62471287, 62402201, 62161013, in part by the Natural Science Foundation of Jiangxi Province of China under Grants 20252BAC240197, in part by the China Postdoctoral Science Foundation under Grant 2025M771495, and in part by the Early-Career Young Scientists and Technologists Project of Jiangxi Province under Grant 20244BCE52070.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, pages 3686–3693, 2014.

[2] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: an open source facial behavior analysis toolkit. In *WACV*, pages 1–10, 2016.

[3] Vighnesh Birodkar, Zhichao Lu, Siyang Li, Vivek Rathod, and Jonathan Huang. The surprising impact of mask-head architecture on novel class segmentation. In *ICCV*, 2021.

[4] Bjorn Browatzki and Christian Wallraven. 3fabrec: Fast few-shot face alignment by reconstruction. In *CVPR*, pages 6110–6120, 2020.

[5] Jinkun Cao, Hongyang Tang, Hao-Shu Fang, Xiaoyong Shen, Cewu Lu, and Yu-Wing Tai. Cross-domain adaptation for animal pose estimation. In *ICCV*, pages 9498–9507, 2019.

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020.

[7] Junjie Chen, Li Niu, Liu Liu, and Liqing Zhang. Weak-shot fine-grained classification via similarity transfer. In *NeurIPS*, 2021.

[8] Junjie Chen, Li Niu, Siyuan Zhou, Jianlou Si, Chen Qian, and Liqing Zhang. Weak-shot semantic segmentation via dual similarity transfer. In *NeurIPS*, pages 32525–32536, 2022.

[9] Junjie Chen, Jiebin Yan, Yuming Fang, and Li Niu. Meta-point learning and refining for category-agnostic pose estimation. In *CVPR*, 2024.

[10] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *CVPR*, pages 7103–7112, 2018.

[11] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *NeurIPS*, 34:17864–17875, 2021.

[12] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. *https://github. com/open-mmlab/mmpose*, 2020.

[13] Shanghua Gao, Zhong-Yu Li, Ming-Hsuan Yang, Ming-Ming Cheng, Junwei Han, and Philip Torr. Large-scale unsupervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[14] Yuying Ge, Ruimao Zhang, and Ping Luo. Metacloth: Learning unseen tasks of dense fashion landmark detection from a few samples. *IEEE Transactions on Image Processing*, 31:1120–1133, 2021.

[15] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-up human pose estimation via disentangled keypoint regression. In *CVPR*, pages 14676–14686, 2021.

[16] Xingzhe He, Gaurav Bharaj, David Ferman, Helge Rhodin, and Pablo Garrido. Few-shot geometry-aware keypoint localization. In *CVPR*, pages 21337–21348, 2023.

[17] Xingzhe He, Bastian Wandt, and Helge Rhodin. Autolink: Self-supervised learning of human skeletons and object outlines by linking keypoints. In *NeurIPS*, pages 36123–36141, 2022.

[18] Eric Hedlin, Gopal Sharma, Shweta Mahajan, Xingzhe He, Hossam Isack, Abhishek Kar Helge Rhodin, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised keypoints from pretrained diffusion models. In *CVPR*, 2024.

[19] Or Hirschorn and Shai Avidan. Pose anything: A graph-based approach for category-agnostic pose estimation. In *ECCV*, 2024.

[20] Jian Hu, Jiayi Lin, Shaogang Gong, and Weitong Cai. Relax image-specific prompt requirement in sam: A single generic prompt for segmenting camouflaged objects. In *AAAI*, 2024.

[21] Jian Hu, Jiayi Lin, Junchi Yan, and Shaogang Gong. Leveraging hallucinations to reduce manual prompt dependency in promptable segmentation. *NeurIPS*, 37, 2024.

[22] Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, and Ross Girshick. Learning to segment every thing. In *CVPR*, 2018.

[23] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. In *NeurIPS*, 2018.

[24] Sheng Jin, Wentao Liu, Wanli Ouyang, and Chen Qian. Multi-person articulated tracking with spatial and temporal embeddings. In *CVPR*, pages 5664–5673, 2019.

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

[26] Weicheng Kuo, Anelia Angelova, Jitendra Malik, and Tsung-Yi Lin. Shapemask: Learning to segment novel objects by refining shape priors. In *ICCV*, 2019.

[27] Rollyn Labuguen, Jumpei Matsumoto, Salvador Blanco Negrete, Hiroshi Nishimaru, Hisao Nishijo, Masahiko Takada, Yasuhiro Go, Ken-ichi Inoue, and Tomohiro Shibata. Macaquepose: a novel in the wild macaque monkey pose dataset for markerless motion capture. *Frontiers in behavioral neuroscience*, 14:581154, 2021.

[28] Jiefeng Li, Siyuan Bian, Ailing Zeng, Can Wang, Bo Pang, Wentao Liu, and Cewu Lu. Human pose regression with residual log-likelihood estimation. In *ICCV*, pages 11025–11034, 2021.

[29] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *CVPR*, pages 10863–10872, 2019.

[30] Shuo Li, Fang Liu, Zehua Hao, Xinyi Wang, Lingling Li, Xu Liu, Puhua Chen, and Wenping Ma. Logits deconfusion with clip for few-shot learning. In *CVPR*, 2025.

[31] Shuo Li, Fang Liu, Zehua Hao, Kaibo Zhao, and Licheng Jiao. Unsupervised few-shot image classification by learning features into clustering space. In *ECCV*, 2022.

[32] Yan Li, Junge Zhang, Kaiqi Huang, and Jianguo Zhang. Mixed supervised object detection with robust objectness transfer. *IEEE transactions on pattern analysis and machine intelligence*, 41(3):639–653, 2018.

[33] Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In *CVPR*, pages 11313–11322, 2021.

[34] Yujia Liang, Zixuan Ye, Wenze Liu, and Hao Lu. Scape: A simple and strong category-agnostic pose estimator. In *ECCV*, pages 342–360, 2024.

[35] Zhaohe Liao, Jiangtong Li, Li Niu, and Liqing Zhang. Align and aggregate: Compositional reasoning with video alignment and answer aggregation for video question-answering. In *CVPR*, 2024.

[36] Zhaohe Liao, Jiangtong Li, Siyu Sun, Qingyang Liu, Fengshun Xiao, Tianjiao Li, Qiang Zhang, Guang Chen, Li Niu, Changjun Jiang, et al. Divide and conquer: Exploring language-centric tree reasoning for video question-answering. In *ICML*, 2025.

[37] Huan Liu, Qiang Chen, Zichang Tan, Jiang-Jiang Liu, Jian Wang, Xiangbo Su, Xiaolong Li, Kun Yao, Junyu Han, Errui Ding, et al. Group pose: A simple baseline for end-to-end multi-person pose estimation. In *ICCV*, pages 15029–15038, 2023.

[38] Yan Liu, Zhijie Zhang, Li Niu, Junjie Chen, and Liqing Zhang. Mixed supervised object detection by transferring mask prior and semantic similarity. In *NeurIPS*, 2021.

[39] Dominik Lorenz, Leonard Bereska, Timo Milbich, and Bjorn Ommer. Unsupervised part-based disentangling of object shape and appearance. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10955–10964, 2019.

[40] Changsheng Lu and Piotr Koniusz. Few-shot keypoint detection with uncertainty learning for unseen species. In *CVPR*, pages 19416–19426, 2022.

[41] Changsheng Lu and Piotr Koniusz. Detect any keypoints: An efficient light-weight few-shot keypoint detector. In *AAAI*, pages 3882–3890, 2024.

[42] Changsheng Lu, Zheyuan Liu, and Piotr Koniusz. Openkd: Opening prompt diversity for zero-and few-shot keypoint detection. In *ECCV*, 2024.

[43] Changsheng Lu, Hao Zhu, and Piotr Koniusz. From saliency to dino: Saliency-guided vision transformer for few-shot keypoint detection. *arXiv preprint arXiv:2304.03140*, 2023.

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

[45] Weian Mao, Yongtao Ge, Chunhua Shen, Zhi Tian, Xinlong Wang, Zhibin Wang, and Anton van den Hengel. Poseur: Direct human pose regression with transformers. In *ECCV*, pages 72–88. Springer, 2022.

[46] Eric Marchand, Hideaki Uchiyama, and Fabien Spindler. Pose estimation for augmented reality: a hands-on survey. *IEEE transactions on visualization and computer graphics*, 22(12):2633–2651, 2015.

[47] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499. Springer, 2016.

[48] Khoi Duc Nguyen, Chen Li, and Gim Hee Lee. Escape: Encoding super-keypoints for category-agnostic pose estimation. In *CVPR*, pages 23491–23500, 2024.

[49] Xuecheng Nie, Jiashi Feng, Jianfeng Zhang, and Shuicheng Yan. Single-stage multi-person pose machines. In *ICCV*, pages 6951–6960, 2019.

[50] Li Niu, Wen Li, and Dong Xu. Visual recognition by learning from web data: A weakly supervised domain generalization approach. In *CVPR*, 2015.

[51] Li Niu, Ashok Veeraraghavan, and Ashutosh Sabharwal. Webly supervised learning meets zero-shot learning: A hybrid approach for fine-grained classification. In *CVPR*, 2018.

[52] Duo Peng, Zhengbo Zhang, Ping Hu, Qiuhong Ke, David KY Yau, and Jun Liu. Harnessing text-to-image diffusion models for category-agnostic pose estimation. In *ECCV*, pages 342–360, 2024.

[53] N Dinesh Reddy, Minh Vo, and Srinivasa G Narasimhan. Carfusion: Combining point tracking and part detection for dynamic 3d reconstruction of vehicles. In *CVPR*, pages 1906–1915, 2018.

[54] Pengfei Ren, Yuanyuan Gao, Haifeng Sun, Qi Qi, Jingyu Wang, and Jianxin Liao. Dynamic support information mining for category-agnostic pose estimation. In *CVPR*, pages 1921–1930, 2024.

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

[56] Matan Rusanovsky, Or Hirschorn, and Shai Avidan. Capex: Category-agnostic pose estimation from textual point explanation. *arXiv preprint arXiv:2406.00384*, 2024.

[57] Luca Schmidtke, Athanasios Vlontzos, Simon Ellershaw, Anna Lukens, Tomoki Arichi, and Bernhard Kainz. Unsupervised human pose estimation through transforming shape templates. In *CVPR*, pages 2484–2494, 2021.

[58] Dahu Shi, Xing Wei, Liangqi Li, Ye Ren, and Wenming Tan. End-to-end multi-person pose estimation with transformers. In *CVPR*, pages 11069–11078, 2022.

[59] Min Shi, Zihao Huang, Xianzheng Ma, Xiaowei Hu, and Zhiguo Cao. Matching is not enough: A two-stage framework for category-agnostic pose estimation. In *CVPR*, pages 7308–7317, 2023.

[60] Xibin Song, Peng Wang, Dingfu Zhou, Rui Zhu, Chenye Guan, Yuchao Dai, Hao Su, Hongdong Li, and Ruigang Yang. Apollocar3d: A large 3d car instance understanding benchmark for autonomous driving. In *CVPR*, pages 5452–5462, 2019.

[61] Jinghan Sun, Dong Wei, Liansheng Wang, and Yefeng Zheng. Lesion guided explainable few weak-shot medical report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 615–625. Springer, 2022.

[62] Meiqi Sun, Zhonghan Zhao, Wenhao Chai, Hanjun Luo, Shidong Cao, Yanting Zhang, Jenq-Neng Hwang, and Gaoang Wang. Uniap: Towards universal animal perception in vision via few-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5008–5016, 2024.

[63] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018.

[64] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *ICCV*, pages 5916–5925, 2017.

[65] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *ICCV*, pages 9469–9478, 2019.

[66] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 139–149, 2022.

[67] Zhen Wei, Bingkun Liu, Weinong Wang, and Yu-Wing Tai. Few-shot model adaptation for customized facial landmark detection, segmentation, stylization and shadow removal. *arXiv preprint arXiv:2104.09457*, 2021.

[68] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, pages 466–481, 2018.

[69] Lumin Xu, Sheng Jin, Wang Zeng, Wentao Liu, Chen Qian, Wanli Ouyang, Ping Luo, and Xiaogang Wang. Pose for everything: Towards category-agnostic pose estimation. In *ECCV*, pages 398–416, 2022.

[70] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *NeurIPS*, 35:38571–38584, 2022.

[71] Jie Yang, Ailing Zeng, Ruimao Zhang, and Lei Zhang. X-pose: Detecting any keypoints. In *European Conference on Computer Vision*, pages 249–268. Springer, 2025.

[72] Jie Yang, Wang Zeng, Sheng Jin, Lumin Xu, Wentao Liu, Chen Qian, and Ruimao Zhang. Kptllm: Unveiling the power of large language model for keypoint comprehension. *arXiv preprint arXiv:2411.01846*, 2024.

[73] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2878–2890, 2012.

[74] Wang Zeng, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In *CVPR*, pages 11101–11111, 2022.

[75] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *CVPR*, pages 7093–7102, 2020.

[76] Hao Zhang, Lumin Xu, Shenqi Lai, Wenqi Shao, Nanning Zheng, Ping Luo, Yu Qiao, and Kaipeng Zhang. Open-vocabulary animal keypoint detection with semantic-feature matching. *International Journal of Computer Vision*, pages 1–18, 2024.

[77] Hao Zhang, Kaipeng Zhang, Lumin Xu, Shenqi Lai, Wenqi Shao, Naning Zheng, Ping Luo, and Yu Qiao. Language-driven open-vocabulary keypoint detection for animal body and face. *arXiv preprint arXiv:2310.05056*, 2023.

[78] Xu Zhang, Wen Wang, Zhe Chen, Yufei Xu, Jing Zhang, and Dacheng Tao. Clamp: Prompt-based contrastive learning for connecting language and animal pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23272–23281, 2023.

[79] 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.

[80] Yuanyi Zhong, Jianfeng Wang, Jian Peng, and Lei Zhang. Boosting weakly supervised object detection with progressive knowledge transfer. In *ECCV*, 2020.

[81] Siyuan Zhou, Li Niu, Jianlou Si, Chen Qian, and Liqing Zhang. Weak-shot semantic segmentation by transferring semantic affinity and boundary. *BMVC*, 2022.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (12 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our abstract makes the main claims and our introduction summarize our contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We discuss the limitations in appendix.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: Our paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We describe our method clearly and fully in the main text, and we conduct experiment using public dataset. We will release the code after the acceptance of paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We conduct experiment using public dataset, and will release code after the acceptance of paper.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so No is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify all the training and test details in the main text and appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We conduct significance test in appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the details of computer resources in appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in our paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the potential positive societal impacts and negative societal impacts in appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We conduct experiment using public dataset, and will release model after the accpetance of paper.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite all the original papers that produced the code package or dataset.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We conduct experiment using public dataset, and will release code after the accpetance of paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: Our core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.