Doodle Your Keypoints: Sketch-Based Few-Shot Keypoint Detection

Subhajit Maity[1]* Ayan Kumar Bhunia[2] Subhadeep Koley[2] Pinaki Nath Chowdhury[2]
Aneeshan Sain[2] Yi-Zhe Song[2]
[1] Department of Computer Science, University of Central Florida, United States.
[2] SketchX, CVSSP, University of Surrey, United Kingdom.
Subhajit@ucf.edu; {a.bhunia, s.koley, p.chowdhury, a.sain, y.song}@surrey.ac.uk
https://subhajitmaity.me/DYKp

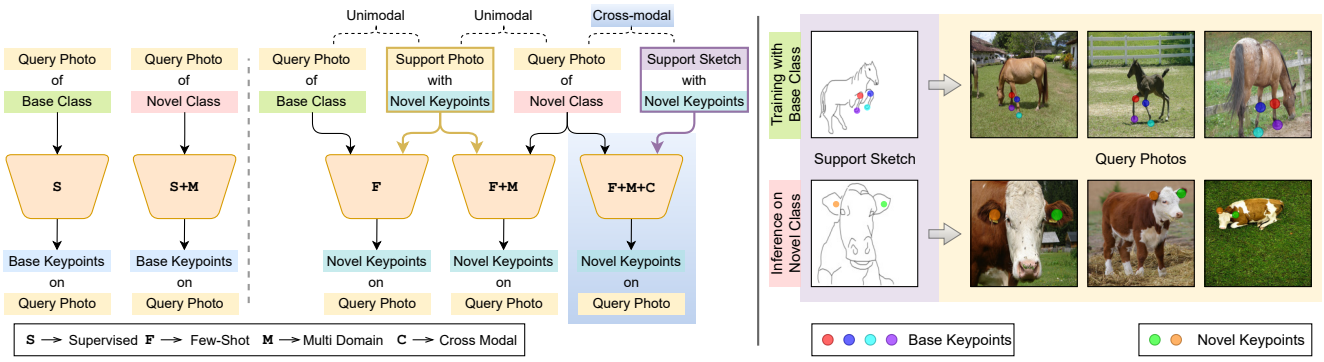

Figure 1. Keypoint detection [17], usually approached in a supervised setting [29], along with cross-domain adaptation [15], can be set up in a few-shot paradigm [123] by typically learning to localize novel keypoints from limited annotated photos [41], also adapting to novel classes [67]. Unlike the existing works, we approach the problem in a cross-modal setup. *(right)* The proposed few-shot framework adapts to localize novel keypoints on photos of unseen classes given a few annotated sketches.

Abstract

Keypoint detection, integral to modern machine perception, faces challenges in few-shot learning, particularly when source data from the same distribution as the query is unavailable. This gap is addressed by leveraging sketches, a popular form of human expression, providing a source-free alternative. However, challenges arise in mastering cross-modal embeddings and handling user-specific sketch styles. Our proposed framework overcomes these hurdles with a prototypical setup, combined with a grid-based locator and prototypical domain adaptation. We also demonstrate success in few-shot convergence across novel keypoints and classes through extensive experiments.

1. Introduction

Keypoint detection, a fundamental component in modern machine perception and visual understanding, has been integral to computer vision tasks since its inception [15, 17, 113, 133]. Initially harnessed as *distinctive* features [4, 38, 66], keypoints have evolved from handcrafted designs to play crucial roles in contemporary pose estimation [17, 61] and landmark detection [116, 123]. Their significance extends to applications such as localizing landmarks in faces [113, 116] and joints [36], contributing to pose [17, 19, 29, 135] and action [64] understanding.

While existing approaches predominantly rely on heatmap regression or direct regression methods dependent on large annotated datasets [17, 18, 29, 78, 117, 135], few-shot keypoint learning remains a challenge. Current strategies, often repurposed from unsupervised or semi-supervised techniques, are constrained to specific image domains, limiting their *applicability* to novel keypoints on unseen objects [67]. The scarcity of research addressing the challenge of unavailable source modality in few-shot keypoint learning, particularly for novel keypoints, highlights a significant gap in the current landscape.

Why choose *sketches*? Sketches [42] serve as a popular form of human expression and offer a viable alternative when the source data is limited in a few-shot setting. Unlike real images, human sketches, easily annotated, provide a practical solution for localizing novel keypoints in unseen classes. While generalized few-shot keypoint detection [67] must handle domain shifts and necessitates a few annotated examples, sketches present the opportunity to create a *source-free* pipeline for practical implementations.

In the pursuit of utilizing sketches as an alternative sup-

*Work done as an intern at SketchX before joining UCF.

port modality for a source-free few-shot keypoint detection approach, several distinct challenges emerge. First, mastering *cross-modal* keypoint embeddings proves formidable due to inherent disparities between these embeddings and joint sketch-photo embedding as per conventional cross-modal learning. Second, encoding images within this joint feature space may lack direct *correspondence* for specific keypoint embeddings, adding intricacy to the task. Third, the untested nature of few-shot capability within this cross-modal feature space introduces additional complexity. Simultaneously, sketches, as expressive forms of human representation rather than precise depictions of photorealism, introduce variability through user-specific styles, necessitating effective generalization across diverse styles to facilitate efficient keypoint learning.

To tackle these challenges, we adopt a prototypical setup [108], constructing keypoint *prototypes* by pooling embeddings from encoded support feature maps. These prototypes are then correlated with query feature maps, generating rich keypoint descriptors. The subsequent use of the Grid Based Locator (GBL) [67] allows for the localization of keypoints from the rich descriptor through direct regression. To bridge the sketch-photo domain gap at the keypoint level, we introduce prototypical domain adaptation [115], which addresses the disparity between support prototypes and query keypoint embeddings pooled from query feature maps. Managing user style diversity is deemed crucial, and therefore, we maximize the similarity between corresponding keypoint embeddings extracted from different sketch versions. Each version represents a single object-level photo and exhibits varying styles.

Our contributions include *(a)* the introduction of a *sketch-based prototypical keypoint detection* framework, *(b)* the pioneering exploration of *source-free* few-shot keypoint detection, and *(c)* extensive experimentation and analysis establishing the proposed method as a successful solution for few-shot convergence on novel keypoints in novel classes within a *source-free cross-modal* setup.

2. Related Works

Sketch for Visual Understanding: Having a surprisingly similar *interpretation* as real objects for human perception [42], sketch [22] and line drawings [56, 60, 109] have been a *popular* mode of expression, driving applications of *hand-drawn sketches* in multiple tasks including retrieval [7, 8, 20, 102, 103, 142], generation [34, 50], inpainting [128], segmentation [44], 3D modeling [143], and augmented reality [132]. In particular, sketch-based image retrieval [24, 28, 100, 101, 107, 138] involving various deep networks [7, 24, 96] to learn a *cross-modal embedding space* has been immensely popular among recent research directions due to its significant commercial importance. Due to the popularity of sketch as a *barrier-free* mode of human interaction across ethnic and social diversity, it has been used to guide object detection [21], saliency exploration [10], video synthesis [59], editing [134], and other computer vision problems [23, 131]. Attributing to the *cognitive* and *creative* potential [32] of humane intelligence, sketch has been used for Pictionary-styled gaming [6]. With all these diverse directions, the research community has explored the exploitation of both representative [119] and discriminative [9, 21] properties of sketch images. A more detailed exploration of sketches for various visual understanding tasks and their comparative analysis with several future directions are given in [5]. In this work, we particularly explore the *representative* property of the sketch in a *localized context* at a cross-modal latent for learning distinct keypoints in real photos.

Few-Shot Learning: Few-shot learning [121] relates to meta-learning [139] closely, delving into strategies promoting *faster convergence* with *fewer training samples* and is approached by a learning-to-learn [51] approach. From an early age meta-learning [124] usually involved models [90], either specifically designed to use external memory buffers [35, 105, 125] and fast weights [76], or strategies [99, 145] designed to imitate and manipulate optimization [74, 88, 91, 136] and gradients [30, 80]. Irrespective of strategies like context adaptation [148], cross-modal setup [85, 106, 130], Bayesian [140], Low-shot [31, 37, 86, 122], or the recent metric learning approaches [2, 12, 62, 82, 141], modern few-shot learning uses a *support* set of a few *unseen* examples to maximize task-level performance on the *query* set. Among metric-based approaches [49, 114, 118], the prototypical network [108] and improvements [53, 58] became popular recently. Few-shot strategies, primarily curated for classification to distinguish *class-specific semantics*, cannot directly cater to keypoint detection due to significantly different task spaces. In this context, prototypes [108], carrying rich *representative* features from *semantically distinct keypoints*, are suitable for the few-shot keypoint learning regimes [67].

Keypoint Detection: Ever since its inception [4, 38, 66], keypoint detection has been one of the major research trends in computer vision. The advent of deep learning [16, 113] has propelled the strategies and the usefulness of keypoint detection. Although being motivated by object detection [13, 95] and semantic segmentation [40] research, keypoint detection has advanced object detection by rethinking it from keypoints [52, 146, 147]. However, it has been traditionally used for identifying landmarks in the face [113, 116] and joints [36], specifically to understand the skeletal structure or pose [17, 61] of both living [15] and non-living objects [133]. Keypoint detection is typically solved either by employing a *heat-based localization* strategy [11, 29, 78, 113, 135], or using *direct regression* [18, 117]. The heatmap-based strategies, specif-

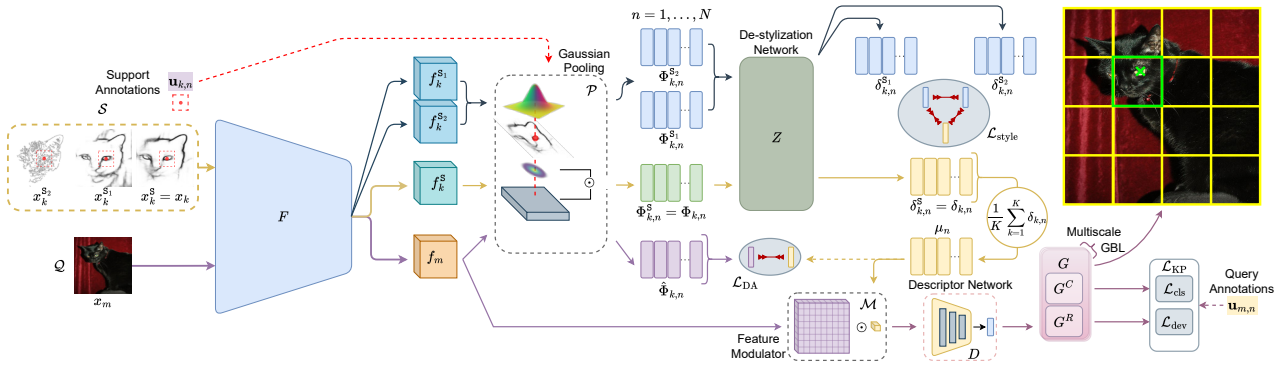

Figure 2. Overview of the proposed few-shot key-point detection framework that processes sketches or edgemaps in the support set and photos in the query set. It employs an encoder to extract deep feature maps followed by the derivation of keypoint embeddings through Gaussian Pooling. Support keypoint prototypes are constructed by averaging keypoint embeddings after disentangling style information through the de-stylization network. Support-query correlation is performed by a point-to-point multiplication of prototype and a query feature map, and subsequently a descriptor network formulates a query descriptor, which is used for localization by the GBL module.

ically the ones for pose estimation, can be further classified into *top-down* approaches [29, 111] and *bottom-up* approaches [17, 19, 72, 78]. Keypoint detection research towards cross-domain adaptation [15], shape deformation aware pose translation [57], self-supervision [47, 81], contrastive learning [3], semi-supervision [27, 43, 73, 75, 87] and unsupervised pipelines [46, 65, 116] were also seen. The initial few-shot approaches [11, 137] explored semi-supervised or few-shot adaptation [123] following self-supervised pre-training, without leaving any room for cross-modal adaptation. While some strategies [33, 68, 69, 112] adapt to novel keypoints, exploiting shape-level [41, 126] or location-level [67] uncertainty, and additional textual guidance [70], they inherently cannot address the domain gap between sketches to learn from and photos to localize on.

Sketch as an Alternate to Photos: Multi-modal learning [120] has been a recent trend to boost performance [63] and perform semi-supervision [8], and cross-modal [54, 144] adaptations. In this regard, sketch [104, 142] became a popular choice for cross-modal [7, 100] counterparts of object detection [21], saliency recognition [10] *etc*. Despite multimodal [45, 83, 130] approaches for *few-shot* learning included sketch [9], keypoints regime remained unexplored. Sketch [42] being easy to achieve using a *few strokes* remains a potential alternative for photos in particular scenarios, *e.g*. when the photo collection is hazardous or the subject is a rare species of bird or animal, or when the photo cannot be used due privacy and ethical constraints or, particularly when variability in object pose and perspective becomes a big challenge by occluding keypoints.

3. Methodology

Overview: In the context of few-shot keypoint detection, we have access to relevant dataset $\mathcal{D} = \{x_i^{\mathrm{P}}, y_i\}$ with keypoint annotations of multiple species or objects where $x_i^{\mathrm{P}} \in \mathcal{I}^{3\times H\times W}$ is a standard RGB image and the annotation $y_i = \{\mathbf{u}_n, \mathbf{v}_n\}_{n=1}^{N}$ where $\mathbf{u} \in \mathbb{R}^2$ denotes coordinate location $\mathtt{x}, \mathtt{y} \in [-1, 1]$ of a keypoint, and visibility $\mathbf{v} \in \{0, 1\}$ specifies if the keypoint is visible in the photo x_i^{P}, for each of the N pre-defined keypoints. Due to the lack of a sketch-photo dataset with keypoint-level annotation, we use off-the-shelf edge detectors [14, 110, 129] to generate multiple versions of edgemaps from each photo x_i^{P}, considering them as equivalent to sketch data. In particular, we consider PiDiNet [110] as the primary edge detector to generate x_i^{S}, and HED [129] and Canny [14] as additional detectors providing *style diversity*, respectively producing x_i^{S1} and x_i^{S2} against any real photo x_i^{P}. All the edgemaps are standard RGB images $\mathcal{I}^{3\times H\times W}$. Thus, it is formulated as an N-way K-shot learning problem to detect N keypoints on query photos using K annotated sketches. The *support* set $\mathcal{S} = \{x_k, y_k\}_{k=1}^{K}$ and *query* set $\mathcal{Q} = \{x_m, y_m\}_{m=1}^{M}$ contains K and M distinct samples $\{x_i, y_i\}$ respectively from $\mathcal{D}$ where $x_k = x_k^{\mathrm{S}}$ and $x_m = x_m^{\mathrm{P}}$.

3.1. Few-Shot Framework

The proposed few-shot framework closely follows Lu *et al*. [67] by building prototypes [108] for all N keypoints from support set $\mathcal{S}$, satisfying the requirements of keypoint localization, which is achieved by direct regression using a Grid-based Locator (GBL) [67] catering it towards sketch data (Sec. 3.1.2). The proposed architecture (Fig. 2) uses an image encoder F to extract feature maps from the support edgemaps and the query photo. Subsequently, given the support keypoint locations, a keypoint extractor $\mathcal{P}$ extracts keypoint embeddings from support features, followed by computing support prototypes, which correlate to the query features using a feature modulator $\mathcal{M}$. Feature descriptors are obtained from correlated feature maps through a descriptor network [67] D and lastly, the GBL modules (grid classifier G^C and regressor G^R) operating in multi-scale localizes the keypoints. The system is trained with keypoints loss $\mathcal{L}_{\mathrm{KP}}$. Additionally, a *domain adaptation* strategy [115] optimizes a domain loss $\mathcal{L}_{\mathrm{DA}}$ discussed in Sec. 3.2 and a novel *de-stylization* network Z utilizing an attention mechanism [25] maximizes information across diverse edgemap

styles using $\mathcal{L}_{\text{style}}$ as elaborated in Sec. 3.3. Moreover, auxiliary keypoints and tasks improve performance using auxiliary losses $\mathcal{L}_{\text{KP-aux}}$, $\mathcal{L}_{\text{DA-aux}}$, and $\mathcal{L}_{\text{style-aux}}$ as given in Sec. 3.4. The total loss is in Eq. (1) where λ_{KP}, λ_{DA} and λ_{style} are corresponding loss scaling factors.

$$\begin{aligned}\mathcal{L}_{\text{total}} = \lambda_{\text{KP}}(\mathcal{L}_{\text{KP}} + \mathcal{L}_{\text{KP-aux}}) + \lambda_{\text{DA}}(\mathcal{L}_{\text{DA}} + \mathcal{L}_{\text{DA-aux}}) \\ + \lambda_{\text{style}}(\mathcal{L}_{\text{style}} + \mathcal{L}_{\text{style-aux}})\end{aligned} \tag{1}$$

3.1.1. Support Prototypes & Query Correlation

Image Encoder: The image encoder $F : \mathcal{I}^{3\times H\times W} \rightarrow \mathbb{R}^{c\times h\times w}$ is an essential component of the framework that takes an input x_i, *i.e.* either a support sketch or edgemap x_k^{S} or a query photo x_m^{P}, and encodes it to convolutional feature map f_i. Meaning, for each of the K edgemaps $x_k \in \mathcal{S}$ and for each of the M photos $x_m \in \mathcal{Q}$, the encoder F generates support feature f_k and query feature f_m respectively.

Support Prototypes: The keypoint detection task inherently demands the learning of keypoint embeddings $\Phi_{k,n} \in \mathbb{R}^c$. So, we use $\mathcal{P}$ operator for extracting embedding $\Phi_{k,n} = \mathcal{P}(f_k, \mathbf{u}_{k,n})$ for keypoint n in $x_k \in \mathcal{S}$. $\mathcal{P}$ is realized by Gaussian pooling [67] in Eq. (2) to encode *sufficiently discriminative* local context *without* any hard local boundary at the corresponding ground-truth location $\mathbf{u}_{k,n} \in y_k$, where $f_k[\mathbf{x}]$ refers to the $\mathbb{R}^c$ vector from f_k indexed at location $\mathbf{x}$.

$$\mathcal{P}(f_k, \mathbf{u}_{k,n}) = \sum_{\mathbf{x}} \exp\left(\frac{-\|\mathbf{x} - \mathbf{u}_{k,n}\|_2^2}{2\xi^2}\right) \cdot f_k[\mathbf{x}] \tag{2}$$

Each $\Phi_{k,n}$ from support x_k is mapped to a *style-agnostic* embedding $\delta_{k,n} \in \mathbb{R}^c$ enriched with *mutual information across styles* (Sec. 3.3) using de-stylization network Z.

Following the standard prototypical network [108], we compute each support keypoint prototype $\mu_n : n = 1, \ldots, N$ as a mean of the visible de-stylized support keypoint embeddings $\delta_{k,n}$ across the support set $\mathcal{S}$ as per Eq. (3). The visibility of the prototype μ_n is `true` if $\mathbf{v}_{k,n} = 1$ for at least one $x_k \in \mathcal{S}$. The loss uses it to not penalize the model for predicting keypoints absent in $\mathcal{S}$.

$$\mu_n = \frac{1}{K}\sum_{k=1}^{K} \delta_{k,n} \tag{3}$$

Support-Query Correlation: While the original prototypical network [108] used a distance metric to assign class probabilities for the classification task, the keypoint localization is inherently different from it and requires local-level feature correspondence for each keypoint. Hence, support prototypes $\mu_n : n = 1, \ldots, N$ for all the keypoints are separately correlated with query feature f_m encoded from query photo $x_m \in \mathcal{Q}$ with a feature modulator [67] $\mathcal{M}$.

$$\mathcal{A}_{m,n} = \mathcal{M}(f_m, \mu_n) \qquad \mathcal{A}_{m,n}[\mathbf{x}] = f_m[\mathbf{x}] \odot \mu_n \tag{4}$$

$\mathcal{M}$ is defined by elementwise multiplication of prototype μ_n and query feature f_m at every location index $\mathbf{x}$ of f_m for each of the N keypoints, as in Eq. (4) and generates correlated feature $\mathcal{A}_{m,n} \in \mathbb{R}^{c\times h\times w}$ corresponding to x_m.

Following, Lu *et al.* [67] a *descriptor* network $D : \mathcal{A}_{m,n} \rightarrow \Psi_{m,n}$ refines and encodes the positional information in the query photo x_m from the correlated feature maps $\mathcal{A}_{m,n}$ to descriptors $\Psi_{m,n} \in \mathbb{R}^d$ for all N keypoints.

3.1.2. Keypoints Localization

Overview: Grids [77, 94, 98], being used for more than a decade for detection [92–94] tasks, were extended to keypoint localization [67] using a grid regression by modeling keypoint uncertainty. Inspired by the same, we use a direct regression in our GBL without any notion of uncertainty, as it becomes daunting due to the *sparse* nature of sketches.

We dismantle the localization task of GBL into two subproblems, *grid classification* and *local grid offset regression*, to restrict the error margin to a small patch. To be precise, to localize keypoint n in a query photo x_m in multi-scale, a set of S grid-scales $L = \{L_1, \ldots, L_S\}$ is considered and each grid-scale $L_i \in L$ is used to divide x_m into $L_i \times L_i$ patches across height and width, of size $3 \times \frac{H}{L_i} \times \frac{W}{L_i}$ each. The patches are labelled as the set $l_i = \{0, 1, 2, \ldots, ({L_i}^2 - 1)\}$ in *left-to-right* and *top-to-bottom* order. The ground-truth $\mathbf{u}_n$ in the corresponding y_m, in the range $[-1, 1]$, is converted to the grid map label $\mathbf{u}_n^C$ and offset $\mathbf{u}_n^R$ as per Eq. (5). Formally, $\mathbf{t}$ denotes the ground-truth coordinates $\mathtt{x}, \mathtt{y} \in \mathbf{u}_n$ normalized in $[0, L_i]$, and $\mathbf{z}$ refers to the $\mathtt{x}, \mathtt{y}$ indexing of the grids l_i. Next, the grid index $\mathbf{z} = \{\mathbf{z}_\mathtt{x}, \mathbf{z}_\mathtt{y}\}$ is converted to actual grid label $\mathbf{u}_n^C \in l_i$ and the grid offset $\mathbf{u}_n^R$ is the deviation from the center, *i.e.* location inside the grid $\mathbf{u}_n^C$ normalized in $[-1, 1]$.

$$\begin{aligned}\mathbf{t} &= (\mathbf{u}_n/2 + 0.5) * L_i & \mathbf{z} &= \lfloor \max(0, \min(\mathbf{t}, 1)) \rfloor \\ \mathbf{u}_n^C &= \mathbf{z}_\mathtt{y} \cdot L_i + \mathbf{z}_\mathtt{x} & \mathbf{u}_n^R &= 2(\mathbf{t} - \mathbf{z} - 0.5)\end{aligned} \tag{5}$$

The submodules of our version of GBL, the grid classifier G^C and the grid regressor G^R try to predict $\mathbf{u}_n^C$ and $\mathbf{u}_n^R$ respectively for localizing keypoints. The total loss is given $\mathcal{L}_{\text{KP}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{dev}}$ which optimizes a completely different objective compared to Lu *et al.* [67]. All the loss calculations and predictions for keypoint n are considered only if it is visible in x_m and has a visible support prototype μ_n.

Grid Classifier: The grid classifier G^C learns to estimate the probability distribution $\hat{\mathbf{u}}_n^C = p(l|\Psi_{m,n})$ of keypoint n being located in patch $l \in l_i$ from the query *keypoint descriptor* $\Psi_{m,n}$ using `softmax` on a linear layer as described in Eq. (6), with an output space $\mathbb{R}^{{L_i}^2}$ for each keypoint n. This specifically resolves the dependency of grid regressor by providing it with a smaller scope to localize.

$$p(l|\Psi_{m,n}) = \frac{\exp(G^C(\Psi_{m,n})_l)}{\sum_{l' \in l_i} \exp(G^C(\Psi_{m,n})_{l'})} \tag{6}$$

While Lu *et al.* [67] considers grid-level uncertainty using log-likelihood, we opted to train G^C with the much simpler cross-entropy loss $\mathcal{L}_{\text{cls}}$, as given in Eq. (7), leveraging the one-hot encoded ground-truth $\mathbb{1}(\mathbf{u}_n^C)$.

$$\mathcal{L}_{\text{cls}} = -\sum \log(\hat{\mathbf{u}_n^C}) \odot \mathbb{1}(\mathbf{u}_n^C) \tag{7}$$

$$\mathcal{L}_{\text{dev}} = \|\hat{\mathbf{u}_n^R} - \mathbf{u}_n^R\|_1 \tag{8}$$

Grid Regressor: The grid regressor G^R tries to predict the offset values $\hat{\mathbf{u}_n^R} = G^R(\Psi_{m,n})$ against the relevant grid

l (from ground-truth $\mathbf{u}_n^C$ during training, and from grid classifier prediction $\max(\hat{\mathbf{u}}_n^C)$ during inference) for keypoint n from the corresponding query *keypoint descriptor* $\Psi_{m,n}$ using a linear layer with output space $\mathbb{R}^2$. Unlike modeling the offset as a sample of multivariate Gaussian distribution in Lu *et al*. [67], our G^R uses an l_1 loss $\mathcal{L}_{\text{dev}}$ as in Eq. (8).

Prediction: The final prediction for keypoint n in query x_m is computed from the predicted grid label probability $\hat{\mathbf{u}}_n^C$ and grid offset $\hat{\mathbf{u}}_n^R$. Considering the predicted grid label as $\max(\hat{\mathbf{u}}_n^C)$, for any grid scale, the prediction is computed by adding $\hat{\mathbf{u}}_n^R$ to its center. The final prediction is given as a mean across all S grids in L.

3.2. Optimizing Sketch-Photo Domain Gap

From the problem definition, the sketch-photo domain gap remains the biggest challenge as no sketch $x_k \in \mathcal{S}$ has *content-level pairing* with any of the photo $x_m \in \mathcal{Q}$. Considering this, we seek guidance from modern domain adaptation [115] that particularly implements a prototypical framework [108] to optimize *transport loss* pulling source prototypes and target samples closer. However, as keypoint detection differs significantly from the classification at the task level, the transport loss [115] cannot be used naively. Moreover, the unsupervised setting [115] practically ignores the keypoint-level class information available.

Bi-directional transport loss [115] uses source and target data along with a common prototype in an unsupervised setting. As the cross-modal task space needs domain adaptation at the *keypoint-level* instead of *feature-level*, we consider support keypoint embeddings $\phi_{k,n}$ and extract similar keypoint embeddings $\hat{\Phi}_{m,n}$ from query photo x_m.

Transport loss [115] uses a *point-to-point* cost $\mathbf{C}(\mu_n, \hat{\Phi}_{m,n}) = \|\mu_n - \hat{\Phi}_{m,n}\|_2$ between prototypes and query keypoints which is realized by l_2 distance. It also relies on a cosine similarity score between μ_n and $\hat{\Phi}_{m,n}$ as an *unnormalized class likelihood* measure to adjust the cost $\mathbf{C}$, which is adapted to a normalized distance-based similarity measure $\mathbf{Sim}(\mu_n, \hat{\Phi}_{m,n}) = \exp\left(-\|\mu_n - \hat{\Phi}_{m,n}\|_2^2\right)$ signifying the likelihood of a query keypoint embedding $\hat{\Phi}_{m,n}$ relating to corresponding support keypoint prototype μ_n, to encourage transport guided by keypoint-level *representative* information instead of *discriminative* class information. See §Suppl.

$$p(\hat{\mu}_n) = \exp\left(-\|\hat{\mu}_n - \mu_n\|_2^2\right) \qquad \hat{\mu}_n = \frac{1}{M}\sum_{m=1}^{M} \hat{\Phi}_{m,n} \tag{9}$$

While iteratively updating the target distribution $p(\hat{\mu}_n)$ was necessary for an unsupervised setting, we have access to the information of query keypoint correspondence to the prototypes and thus, we convert this domain adaptation [115] paradigm to a supervised one using Eq. (9). As the class-level probability distribution to relate the prototype μ_n to query embedding $\hat{\Phi}_{m,n}$ is not coherent with our task, we ignore the `softmax` operation from both the transport losses,

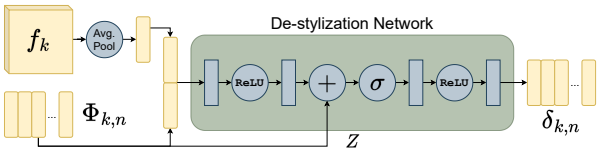

Figure 3. The design of the de-stylization network Z to disentangle the styles fusing global context to local keypoint embeddings.

resulting in reduction of two transport losses [115] to one:

$$\mathcal{L}_{\text{DA}} = \sum p(\hat{\mu}_n)\mathbf{C}(\mu_n, \hat{\Phi}_{m,n})\mathbf{Sim}(\mu_n, \hat{\Phi}_{m,n}) \tag{10}$$

3.3. Learning Style-Agnostic Keypoints

Sketch largely depends on the user's *expressive intent and imagination* as opposed to *photorealism* and has significant user-specific variance in terms of sparsity, information content, and style [100]. This style diversity impacts performance as the framework eventually learns dense information at the keypoint level, irrespective of sketch sparsity. So, we design a de-stylization network Z that de-stylizes any support keypoint embedding $\Phi_{k,n}$ extracted from support edgemap x_k to $\delta_{k,n} \in \mathbb{R}^c$. The off-the-shelf edge detectors [14, 110, 129] generate *different instances* of edgemap x_k^{S}, x_k^{S1} and x_k^{S2} for the *same* photo $x_k^{\text{P}} \in \mathcal{S}$ not only providing the style diversity but also aiding the encoder to learn dense features to tackle inherent sparsity of the sketch data. The design of the de-stylization network Z illustrated in Fig. 3 is inspired by the *multi-scale channel attention* module [25] that aggregates local and global context. (Design detailed in §Suppl.) The *disentanglement* being targetted at the keypoint embeddings with the local context, Z is devised to take the encoded feature f_k^{S}, f_k^{S1} and f_k^{S2} respectively, providing the notion of sparsity in a global scale along with keypoint embeddings $\Phi_{k,n}^{\text{S}}$, $\Phi_{k,n}^{\text{S1}}$ and $\Phi_{k,n}^{\text{S2}}$ for mapping to $\delta_{k,n}^{\text{S}}$, $\delta_{k,n}^{\text{S1}}$ and $\delta_{k,n}^{\text{S2}}$ respectively using an attention mechanism [25]. Considering all the versions of edgemap x_k^{S}, x_k^{S1} and x_k^{S2} as supplements to various user styles with different level of abstraction and information content, Z ideally should disentangle the same and map any keypoint *similar* to each other. Taking x_k^{S} as primary sketch equivalent for training, we have $f_k = f_k^{\text{S}}$, $\Phi_{k,n} = \Phi_{k,n}^{\text{S}}$ and $\delta_{k,n} = \delta_{k,n}^{\text{S}}$. Thus $\mathcal{L}_{\text{style}}$ is derived as:

$$\mathcal{L}_{\text{style}} = \sum_n (\|\delta_{k,n} - \delta_{k,n}^{\text{S1}}\|_2 + \|\delta_{k,n} - \delta_{k,n}^{\text{S2}}\|_2 + \|\delta_{k,n}^{\text{S1}} - \delta_{k,n}^{\text{S2}}\|_2) \tag{11}$$

3.4. Auxiliary Keypoints Improve Performance

Along with the main keypoints given by the annotations, we strategically *interpolate* auxiliary keypoints [67] for the episodic training of our framework. Precisely, given two visible annotated main keypoints n_1 and n_2, and a relative position $t \in (0, 1)$, interpolation $\mathcal{T}(t, (\mathbf{u}_{n_1}, \mathbf{u}_{n_2}))$ obtains an auxiliary keypoint at t times the length on the line running from n_1 to n_2. Simply, auxiliary keypoints [67] are considered on the line interpolating a particular keypoint to another when both are visible. An off-the-shelf saliency detector [127] is used to see if an auxiliary keypoint is inside the saliency region of the object, determining its visibility.

While the generation strategy $\mathcal{T}$ is inspired by Lu *et al*. [67], our auxiliary keypoints serve a completely different set of objectives. Unlike using auxiliary keypoints, particularly to aid the uncertainty learning, we do not model any interaction between main and auxiliary keypoints. Instead, we use these additional keypoints to provide additional *augmented* data for the components in our framework. To this end, we extend the existing objectives of localization, domain adaptation, and de-stylization to T auxiliary keypoints by creating an auxiliary task of T way keypoint learning and calculating $\mathcal{L}_{\text{KP-aux}}$, $\mathcal{L}_{\text{DA-aux}}$, and $\mathcal{L}_{\text{style-aux}}$ accordingly.

4. Experiments

Dataset: For the experiments, we have used the Animal Pose dataset [15] consisting of 5 different species of animals (cat, cow, dog, horse, sheep) with a total of 4,666 images and 6,117 instances. The annotations contain 20 keypoints for each species, from which we used 17 (11 as *base* for training and 6 as *novel* for evaluation). All the ablation experiments are also done on this dataset.

Additionally, we consider Animal Kingdom dataset [79] with 33,099 photos of 850 different species of broadly 5 superclasses (mammal, amphibian, reptile, bird, fish), which we treat as image classes for our purpose. It contains 23 keypoint annotations for each photo, and keeping similarity, we use only 11 *base* keypoints and 7 *novel* keypoints.

Implementation Details: The entire framework is implemented in Python with PyTorch [84] and is trained on a 6GB Nvidia RTX 3060 mobile GPU. The encoder F uses a Imagenet [26] pre-trained ResNet50 [39] backbone. The descriptor network D uses 3 `convolution` layers with `ReLU` [1] in sequence. The de-stylization network Z follows Fig. 3. The GBL components, G^C and G^R use a `linear` layer each, with an additional `softmax` for G^C.

We set $K = 1$ shot and $M = 5$ with $N = 11$ base keypoints for training and $N = 6$ novel keypoints for evaluation. The ablations also use the same setup unless otherwise specified. The input images x_i, for both sketch or edgemap x_k and photo x_m are resized to the height and width $H = W = 384$ and thus the resulting feature map f_i has height and width $h = w = 12$. According to the encoder F configuration, the feature map f_i has $c = 2048$ channels and the descriptor network D allows it to produce query keypoint descriptors of dimension $d = 4096$. The auxiliary keypoints use 6 pre-defined pairs from 11 base keypoints and set $t = \{0.25, 0.5, 0.75\}$ as interpolation points; meaning a maximum of 18 auxiliary keypoints can possibly exist for any x_i during training. The multi-scale aspect of GBL is realized using 3 grid scales, $L = \{8, 12, 16\}$. The empirically determined hyper-parameters are $\xi = 14$, $\lambda_{\text{KP}} = 0.5$, $\lambda_{\text{DA}} = 0.001$ and $\lambda_{\text{style}} = 0.001$. The entire setup is trained for 80,000 iterations of episodes with Adam [48] optimizer with a learning rate of 0.0001.

Figure 4. Sample support ($\mathcal{S}$) and query ($\mathcal{Q}$) for training (*top*) and evaluation (*bottom*) for all evaluation settings.

Evaluation Protocol: We use percentage correct keypoints (PCK) [81] with a margin $\tau = 0.1$ as metric. The prediction is correct if within a range τ times the larger side of the object bounding box. We consider four different settings (see Fig. 4) to analyze the performance of our framework. From the definition of few-shot keypoint detection, we evaluate performance with novel keypoints along with the base keypoints used in training for the base classes it is trained on, as well as never-before-seen classes. All the ablation studies are done for *novel keypoints on unseen classes* of Animal Pose [15] dataset unless otherwise specified. For the evaluation of a novel class, we trained the framework with episodes with data from the rest of the classes, and for seen classes, we used a random 7 : 3 train-test split.

Competitors: Due to the lack of existing works on *cross-modal* few-shot keypoint learning, we repurposed the next best alternative FSKD [67] to use edgemaps [110] as support and photos as query, providing a comparable strategy, thanks to its robustness in detecting novel keypoints across significant differences in object classes.

We establish a vanilla baseline **B-Vanilla** using only the base framework (Sec. 3.1). For B-Vanilla, Z is replaced by an identity function such that we have $\Phi_{k,n} = \delta_{k,n}$ and the additional edgemaps $x_k^{S_1}$ and $x_k^{S_2}$ are not used, setting $\mathcal{L}_{\text{style}} = 0$. We further establish a few incremental baselines. **B-DA** represents the usage of sketch-photo domain adaptation [115] on top of B-Vanilla. **B-Style** uses only style-agnostic keypoint learning presented in Sec. 3.3 on B-Vanilla. As **B-Full** uses both domain adaptation (Sec. 3.2) and de-stylization (Sec. 3.3), it provides the proposed model when used with auxiliary keypoints (Sec. 3.4).

Class	Keypoints	Methods	PCK@0.1 on Animal Pose Dataset [15]						PCK@0.1 on Animal Kingdom Dataset [79]					
			Cat	Cow	Dog	Horse	Sheep	**Mean**	Mammal	Amphibian	Reptile	Bird	Fish	**Mean**
Seen	Base	B-Vanilla	54.12	39.27	44.65	45.58	37.17	44.16	22.32	21.07	18.94	21.77	18.19	20.46
		FSKD [67]	58.95	44.61	49.53	50.21	40.47	48.75	25.52	24.42	21.81	25.73	22.24	23.94
		Proposed	**67.34**	**49.89**	**56.28**	**56.35**	**45.65**	**55.10**	**31.31**	**29.93**	**28.41**	**30.88**	**27.87**	**29.68**
	Novel	B-Vanilla	24.70	15.62	19.08	12.45	18.44	18.06	9.67	7.24	6.55	8.41	4.96	7.37
		FSKD [67]	47.70	35.44	39.81	35.42	31.59	37.99	18.89	17.39	16.72	19.23	15.94	17.63
		Proposed	**55.69**	**43.09**	**46.58**	**43.94**	**36.39**	**45.14**	**25.05**	**23.27**	**22.56**	**24.04**	**21.78**	**23.34**
Unseen	Base	B-Vanilla	43.03	40.31	36.28	44.72	38.03	40.47	12.83	14.12	12.57	13.68	12.41	13.12
		FSKD [67]	41.54	38.10	33.72	41.02	36.30	38.14	14.86	14.28	13.79	15.65	12.16	14.15
		Proposed	**47.36**	**42.97**	**38.30**	**46.17**	**41.03**	**43.17**	**21.98**	**20.15**	**18.96**	**21.52**	**17.19**	**19.96**
	Novel	B-Vanilla	22.71	15.56	16.92	13.58	18.18	17.39	7.26	5.12	3.93	5.69	4.08	5.22
		FSKD [67]	36.75	35.76	32.84	32.66	31.58	33.92	10.96	9.34	9.68	11.45	8.86	10.06
		Proposed	**44.42**	**40.13**	**36.91**	**37.77**	**35.77**	**39.00**	**16.48**	**14.62**	**13.76**	**15.91**	**11.33**	**14.42**

Table 1. A quantitative comparison of the baseline and state-of-the-art strategies with the proposed framework in $K = 1$ shot setting delineating the superiority of the proposed method in overall performance on Animal Pose [15] and Animal Kingdom [79] datasets.

Figure 5. Visualising detection (✖) and ground-truth (●) for novel keypoints for base classes (*top*) and unseen classes (*bottom*).

4.1. Performance Analysis

A detailed performance comparison of the *naive* baseline B-Vanilla and FSKD [67] (repurposed for the cross-modal paradigm) with the proposed framework is presented in Tab. 1. We evaluate in four aforementioned settings for both datasets [15, 79], and our findings are: *(a)* B-Vanilla performs poorly in the majority of the settings as it lacks robustness in keypoint learning due to the sketch-photo domain gap, and absence of auxiliary keypoints. The significant performance gap comes from the lack of auxiliary keypoints during training, discussed further in Sec. 4.4. *(b)* FSKD [67] performs much better than the baseline due to its clever use of localization uncertainty. The huge margin over B-Vanilla can be attributed to uncertainty modeling with main and auxiliary keypoints. However, it falls short of the proposed model due to the sketch-photo domain gap. *(c)* The proposed framework significantly outperforms all its counterparts for all evaluation settings and datasets [15, 79]. This is due to the careful consideration of the sketch-photo domain gap and information maximization across diverse styles. *(d)* All methods have the poorest performance on novel keypoints for unseen classes, which is expected due to the two *levels* of *domain shift* defined by the evaluation protocol. However, our work accounts for the third domain gap between sketches or edgemaps and photos. *(e)* Our method outperforms FSKD [67] by ≈ 5 and ≈ 4 PCK in Animal Pose [15] and Animal Kingdom [79], respectively, for novel keypoints on unseen classes. A few sample predictions with

Figure 6. Inference (✖) with ground-truth (●) for base (*top*) and novel (*bottom*) keypoints with annotated sketch prompts.

ground truth from Animal Pose [15] are presented in Fig. 5. It also has superior performance over FSKD [67], maintaining a margin of $\approx 4 - 8$ in all evaluation settings. *(f)* B-Vanilla performs significantly lower for novel keypoints than base keypoints, showing its poor few-shot capability.

4.2. Generalization on Free-Hand Sketches

The proposed framework is trained on *synthetic* sketches or edgemaps [110] of the photos in the Animal Pose dataset [15]. Thus, we curate a small subset of *real* sketches of the same classes from the Sketchy database [104] to evaluate how well the model generalizes on real free-hand sketches. The evaluation is performed on query photos in the training dataset [15] using a total of 30 real sketches [104] with manual keypoint prompts in $K = 1$ shot setting. The PCK@0.1 values reported for base and novel keypoints are $42.40\%(\downarrow 0.77)$ and $38.49\%(\downarrow 0.51)$ for unseen classes, and $53.29\%(\downarrow 1.81)$ and $44.99\%(\downarrow 0.15)$ for seen classes respectively, showing the trained model easily adapting to free-hand sketches [104]. See §Suppl. (Tab. 5). The inference on a few samples is shown

Method	N (only Main)	$N+T$ (+ Auxiliary)
B-Vanilla	17.39	29.98
B-DA	18.31	31.76
B-Style	18.97	32.51
B-Full	19.03	**39.00**

Table 2. Performances of incremental baselines.

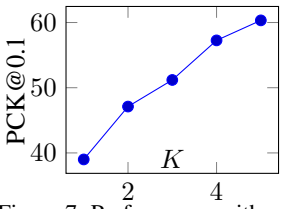

Figure 7. Performance with varying shots.

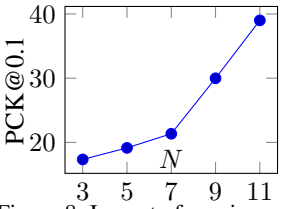

Figure 8. Impact of varying main keypoints.

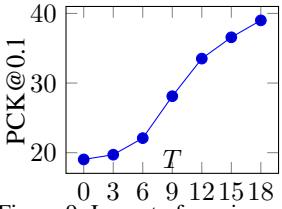

Figure 9. Impact of varying auxiliary keypoints.

in Fig. 6. The framework, designed for few-shot settings, possesses a high capability for adapting to novel data distributions, which helps it generalize well on real sketches. Sketches are contour-level depiction [42] of real photos and are represented closely to edgemaps due to de-stylization (Sec. 3.3), facilitating easy adaptation.

4.3. Experiments on Multi-modality

Considering *novel keypoints on unseen classes*, our framework establishes its superiority with a PCK@0.1 of 39.00% (↑ 5.08) (Tab. 1) compared to FSKD [67] (33.92%). While it is impeccable with sketches, we experiment to see how it fares when being trained with photos instead. FSKD [67] being a state-of-the-art on photos with 44.75% PCK@0.1, our model proves to be robust enough to achieve 43.72% (↓ 1.03) when trained with photos. However, jointly training our model with sketch and photo results in 46.54% outperforming FSKD [67] by 1.79. Evidently, mutual information from both modalities helps in having a more generalized keypoint understanding. Details of the *multi-modal* training and results are in §Suppl.

4.4. Ablation Study

Number of Shots: Typically, increasing the number of shots in a few-shot setting boosts performance [67]. We present the empirical data in Fig. 7 that shows performance increase with increasing K, although the rate of increment gradually diminishes, denoting *information saturation*.

Significance of Loss Objectives: Justifying the independent contribution of the loss objectives, we present a *strip-down-styled* performance analysis of the incremental baselines at $K = 1$ shot in Tab. 2. B-Vanilla is a naive baseline with the poorest performance. B-DA uses domain adaptation [115], B-Style accounts for style diversity, improving $\approx 1 - 1.5$ over B-Vanilla. B-Full devises both modules (Secs. 3.2 and 3.3), has superior performance. Evidently, auxiliary keypoints [67] (Sec. 3.4) significantly boost performance by enhancing the capability of other modules.

Number of Keypoints: Increasing the number of keypoints manifests better *keypoint diversity*, resulting in better generalization with improved performance (Figs. 8 and 9). This is also coherent with auxiliary keypoints that, being synthetic in nature, provide robustness to the keypoints learning regime in a semi-supervised setup. Unlike Lu *et al*. [67], we augment keypoints for the enhancement of the sub-module performances, resulting in improvements of $\approx 12 - 20$ for all the baselines (Tab. 2). Clearly, better-performing baselines improve more than others.

Viability of Edgemaps: As we have used edgemaps as synthetic sketch data to train our framework, to justify the choice of edge detectors, *i.e*. PiDiNet [110], HED [129], and Canny [14], we experimented with a large number of algorithms, and this particular combination proved to have the best performance. Moreover, our framework is flexible to incorporate superior algorithms from future research.

Further Insights: *(a)* Albeit the support keypoint representations $\Phi_{k,n}$ are disentangled to $\delta_{k,n}$ by Z, the query keypoints $\hat{\Phi}_{m,n}$ from photos do not have style or sparsity associated and thus do not require de-stylization. Alongside learning robust support keypoint representations $\delta_{k,n}$, this implicitly poses a support-query embedding alignment restricting $\mathcal{L}_{\text{DA}}$ from collapsing. *(b)* Computation of $\hat{\mu}_n$ for domain adaptation [115] (Sec. 3.2) is detached from the computation graph to prevent a collapse caused by simultaneous gradients through the query and support prototypes. *(c)* The architecture of the de-stylization network (Sec. 3.3) Z is non-trivial as complex models lead to unstable training, and simpler ones are incapable of style disentanglement. (See §Suppl.) *(d)* The additional edgemaps $x_i^{\text{S}_1}$ and $x_i^{\text{S}_2}$ are only used in de-stylization loss $\mathcal{L}_{\text{style}}$, but not in localization loss $\mathcal{L}_{\text{KP}}$ due to significant performance degradation.

5. Conclusion

The few-shot approach to keypoint localization being fairly new in the research community, we explore the scope of *cross-modal* few-shot keypoint detection, adapting quickly to novel keypoints as well as unknown image classes. To this end, we propose a framework to detect novel keypoints on query photos from a few annotated sketches, achieving a *source-free* setup in this paradigm. Particularly, we carefully address the sketch-photo domain gap using domain adaptation, besides catering to the style and abstraction diversity of sketches. Despite being trained on synthetic sketch-like data, our method proves to be adapting well for free-hand sketches. Furthermore, we show that multi-modal training helps generalize better on query photos. To the best of our knowledge, this is amongst the first of the works on cross-modal few-shot keypoint learning, and we hope our work inspires future research in this context.

References

[1] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018. 6

[2] Kelsey R. Allen, Evan Shelhamer, Hanul Shin, and Joshua B. Tenenbaum. Infinite mixture prototypes for few-shot learning. In *ICML*, 2019. 2

[3] Yutong Bai, Angtian Wang, Adam Kortylewski, and Alan Yuille. Coke: Contrastive learning for robust keypoint detection. In *WACV*, 2023. 3

[4] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *ECCV*, 2006. 1, 2

[5] Ayan Kumar Bhunia. *Towards Practicality of Sketch-Based Visual Understanding*. PhD thesis, University of Surrey, 2022. 2

[6] Ayan Kumar Bhunia, Ayan Das, Umar Riaz Muhammad, Yongxin Yang, Timothy M Hospedales, Tao Xiang, Yulia Gryaditskaya, and Yi-Zhe Song. Pixelor: A competitive sketching ai agent. so you think you can sketch? *Siggraph Asia*, 2020. 2

[7] Ayan Kumar Bhunia, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Sketch less for more: On-the-fly fine-grained sketch-based image retrieval. In *CVPR*, 2020. 2, 3, 17

[8] Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Aneeshan Sain, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. More photos are all you need: Semi-supervised learning for fine-grained sketch based image retrieval. In *CVPR*, 2021. 2, 3, 17

[9] Ayan Kumar Bhunia, Viswanatha Reddy Gajjala, Subhadeep Koley, Rohit Kundu, Aneeshan Sain, Tao Xiang, and Yi-Zhe Song. Doodle it yourself: Class incremental learning by drawing a few sketches. In *CVPR*, 2022. 2, 3

[10] Ayan Kumar Bhunia, Subhadeep Koley, Amandeep Kumar, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Sketch2saliency: Learning to detect salient objects from human drawings. In *CVPR*, 2023. 2, 3

[11] Bjorn Browatzki and Christian Wallraven. 3fabrec: Fast few-shot face alignment by reconstruction. In *CVPR*, 2020. 2, 3

[12] Qi Cai, Yingwei Pan, Ting Yao, Chenggang Yan, and Tao Mei. Memory matching networks for one-shot image recognition. In *CVPR*, 2018. 2

[13] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, 2018. 2

[14] John Canny. A computational approach to edge detection. *IEEE TPAMI*, 1986. 3, 5, 8, 16

[15] Jinkun Cao, Hongyang Tang, Hao-Shu Fang, Xiaoyong Shen, Cewu Lu, and Yu-Wing Tai. Cross-domain adaptation for animal pose estimation. In *ICCV*, 2019. 1, 2, 3, 6, 7, 15, 16, 17, 18

[16] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 2

[17] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE TPAMI*, 2021. 1, 2, 3

[18] Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback. In *CVPR*, 2016. 1, 2

[19] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *CVPR*, 2020. 1, 3

[20] Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Viswanatha Reddy Gajjala, Aneeshan Sain, Tao Xiang, and Yi-Zhe Song. Partially does it: Towards scene-level fg-sbir with partial input. In *CVPR*, 2022. 2, 17

[21] Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Aneeshan Sain, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. What can human sketches do for object detection? In *CVPR*, 2023. 2, 3

[22] Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Aneeshan Sain, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. Democratising 2d sketch to 3d shape retrieval through pivoting. In *ICCV*, 2023. 2

[23] Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Aneeshan Sain, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. Scenetrilogy: On human scene-sketch and its complementarity with photo and text. In *CVPR*, 2023. 2

[24] John Collomosse, Tu Bui, and Hailin Jin. Livesketch: Query perturbations for guided sketch-based visual search. In *CVPR*, 2019. 2

[25] Yimian Dai, Fabian Gieseke, Stefan Oehmcke, Yiquan Wu, and Kobus Barnard. Attentional feature fusion. In *WACV*, 2021. 3, 5, 15

[26] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

[27] Xuanyi Dong and Yi Yang. Teacher supervises students how to learn from partially labeled images for facial landmark detection. In *ICCV*, 2019. 3

[28] Anjan Dutta and Zeynep Akata. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In *CVPR*, 2019. 2

[29] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *ICCV*, 2017. 1, 2, 3

[30] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017. 2

[31] Hang Gao, Zheng Shou, Alireza Zareian, Hanwang Zhang, and Shih-Fu Chang. Low-shot learning via covariance-preserving adversarial augmentation networks. In *NeurIPS*, 2018. 2

[32] Songwei Ge, Vedanuj Goswami, Larry Zitnick, and Devi Parikh. Creative sketch generation. In *ICLR*, 2020. 2

[33] Yuying Ge, Ruimao Zhang, and Ping Luo. Metacloth: Learning unseen tasks of dense fashion landmark detection from a few samples. *IEEE TIP*, 2021. 3, 18

[34] Arnab Ghosh, Richard Zhang, Puneet K Dokania, Oliver Wang, Alexei A Efros, Philip HS Torr, and Eli Shechtman. Interactive sketch & fill: Multiclass sketch-to-image translation. In *ICCV*, 2019. 2

[35] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014. 2

[36] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In *CVPR*, 2022. 1, 2

[37] Bharath Hariharan and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *ICCV*, 2017. 2

[38] Chris Harris, Mike Stephens, et al. A combined corner and edge detector. In *Alvey vision conference*, 1988. 1, 2

[39] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[40] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 2

[41] Xingzhe He, Gaurav Bharaj, David Ferman, Helge Rhodin, and Pablo Garrido. Few-shot geometry-aware keypoint localization. In *CVPR*, 2023. 1, 3, 18

[42] Aaron Hertzmann. Why do line drawings work? a realism hypothesis. *Perception*, 2020. 1, 2, 3, 8

[43] Sina Honari, Pavlo Molchanov, Stephen Tyree, Pascal Vincent, Christopher Pal, and Jan Kautz. Improving landmark localization with semi-supervised learning. In *CVPR*, 2018. 3

[44] Conghui Hu, Da Li, Yongxin Yang, Timothy M Hospedales, and Yi-Zhe Song. Sketch-a-segmenter: Sketch-based photo segmenter generation. *IEEE TIP*, 2020. 2

[45] Yan Huang and Liang Wang. Acmm: Aligned cross-modal memory for few-shot image and sentence matching. In *ICCV*, 2019. 3

[46] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. In *NeurIPS*, 2018. 3

[47] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Self-supervised learning of interpretable keypoints from unlabelled videos. In *CVPR*, 2020. 3

[48] DP Kingma. Adam: a method for stochastic optimization. In *ICLR*, 2014. 6

[49] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, 2015. 2

[50] Subhadeep Koley, Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Picture that sketch: Photorealistic image generation from abstract sketches. In *CVPR*, 2023. 2

[51] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 2015. 2

[52] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *ECCV*, 2018. 2

[53] Aoxue Li, Tiange Luo, Tao Xiang, Weiran Huang, and Liwei Wang. Few-shot learning with global class representations. In *ICCV*, 2019. 2

[54] Chao Li, Shangqian Gao, Cheng Deng, De Xie, and Wei Liu. Cross-modal learning with adversarial samples. In *NeurIPS*, 2019. 3

[55] Ke Li, Kaiyue Pang, and Yi-Zhe Song. Photo pre-training, but for sketch. In *CVPR*, 2023. 17

[56] Mengtian Li, Zhe Lin, Radomir Mech, Ersin Yumer, and Deva Ramanan. Photo-sketching: Inferring contour drawings from images. In *WACV*, 2019. 2, 16

[57] Siyuan Li, Semih Gunel, Mirela Ostrek, Pavan Ramdya, Pascal Fua, and Helge Rhodin. Deformation-aware unpaired image translation for pose estimation on laboratory animals. In *CVPR*, 2020. 3

[58] Wenbin Li, Lei Wang, Jinglin Xu, Jing Huo, Yang Gao, and Jiebo Luo. Revisiting local descriptor based image-to-class measure for few-shot learning. In *CVPR*, 2019. 2

[59] Xiaoyu Li, Bo Zhang, Jing Liao, and Pedro V Sander. Deep sketch-guided cartoon video inbetweening. *IEEE TVCG*, 2021. 2

[60] Yijun Li, Chen Fang, Aaron Hertzmann, Eli Shechtman, and Ming-Hsuan Yang. Im2pencil: Controllable pencil illustration from photographs. In *CVPR*, 2019. 2, 16

[61] 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 *ICCV*, 2021. 1, 2

[62] Yann Lifchitz, Yannis Avrithis, Sylvaine Picard, and Andrei Bursuc. Dense classification and implanting for few-shot learning. In *CVPR*, 2019. 2

[63] Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramanan. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. In *CVPR*, 2023. 3

[64] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *CVPR*, 2020. 1

[65] Dominik Lorenz, Leonard Bereska, Timo Milbich, and Bjorn Ommer. Unsupervised part-based disentangling of object shape and appearance. In *CVPR*, 2019. 3

[66] David G Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999. 1, 2

[67] Changsheng Lu and Piotr Koniusz. Few-shot keypoint detection with uncertainty learning for unseen species. In *CVPR*, 2022. 1, 2, 3, 4, 5, 6, 7, 8, 14, 15, 16, 17, 18

[68] Changsheng Lu and Piotr Koniusz. Detect any keypoints: An efficient light-weight few-shot keypoint detector. In *AAAI*, 2024. 3

[69] 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. 3, 18

[70] Changsheng Lu, Zheyuan Liu, and Piotr Koniusz. Openkd: Opening prompt diversity for zero-and few-shot keypoint detection. In *ECCV*, 2024. 3, 17, 18

[71] Xuanchen Lu, Xiaolong Wang, and Judith E Fan. Learning dense correspondences between photos and sketches. In *ICML*, 2023. 17

[72] Zhengxiong Luo, Zhicheng Wang, Yan Huang, Liang Wang, Tieniu Tan, and Erjin Zhou. Rethinking the heatmap regression for bottom-up human pose estimation. In *CVPR*, 2021. 3

[73] Alexander Mathis, Pranav Mamidanna, Kevin M Cury, Taiga Abe, Venkatesh N Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 2018. 3

[74] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. In *ICLR*, 2018. 2

[75] Olga Moskvyak, Frederic Maire, Feras Dayoub, and Mahsa Baktashmotlagh. Semi-supervised keypoint localization. In *ICLR*, 2020. 3

[76] Tsendsuren Munkhdalai and Hong Yu. Meta networks. In *ICML*, 2017. 2

[77] Mahyar Najibi, Mohammad Rastegari, and Larry S Davis. G-cnn: an iterative grid based object detector. In *CVPR*, 2016. 4

[78] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 1, 2, 3

[79] Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *CVPR*, 2022. 6, 7, 16

[80] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018. 2

[81] David Novotny, Samuel Albanie, Diane Larlus, and Andrea Vedaldi. Self-supervised learning of geometrically stable features through probabilistic introspection. In *CVPR*, 2018. 3, 6, 18

[82] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *NeurIPS*, 2018. 2

[83] Frederik Pahde, Mihai Puscas, Tassilo Klein, and Moin Nabi. Multimodal prototypical networks for few-shot learning. In *WACV*, 2021. 3

[84] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 6

[85] Zhimao Peng, Zechao Li, Junge Zhang, Yan Li, Guo-Jun Qi, and Jinhui Tang. Few-shot image recognition with knowledge transfer. In *ICCV*, 2019. 2

[86] Hang Qi, Matthew Brown, and David G Lowe. Low-shot learning with imprinted weights. In *CVPR*, 2018. 2

[87] Shengju Qian, Keqiang Sun, Wayne Wu, Chen Qian, and Jiaya Jia. Aggregation via separation: Boosting facial landmark detector with semi-supervised style translation. In *ICCV*, 2019. 3

[88] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L. Yuille. Few-shot image recognition by predicting parameters from activations. In *CVPR*, 2018. 2

[89] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 18

[90] Tiago Ramalho and Marta Garnelo. Adaptive posterior learning: few-shot learning with a surprise-based memory module. In *ICLR*, 2019. 2

[91] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2017. 2

[92] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. In *CVPR*, 2017. 4

[93] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[94] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 4

[95] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 2

[96] Leo Sampaio Ferraz Ribeiro, Tu Bui, John Collomosse, and Moacir Ponti. Sketchformer: Transformer-based representation for sketched structure. In *CVPR*, 2020. 2

[97] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: A stylegan encoder for image-to-image translation. In *CVPR*, 2021. 17

[98] Peter M Roth, Sabine Sternig, Helmut Grabner, and Horst Bischof. Classifier grids for robust adaptive object detection. In *CVPR*, 2009. 4

[99] Andrei A. Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. In *ICLR*, 2019. 2

[100] Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Stylemeup: Towards style-agnostic sketch-based image retrieval. In *CVPR*, 2021. 2, 3, 5, 17

[101] Aneeshan Sain, Ayan Kumar Bhunia, Vaishnav Potlapalli, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Sketch3t: Test-time training for zero-shot sbir. In *CVPR*, 2022. 2

[102] Aneeshan Sain, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. Clip for all things zero-shot sketch-based image retrieval, fine-grained or not. In *CVPR*, 2023. 2, 17

[103] Aneeshan Sain, Ayan Kumar Bhunia, Subhadeep Koley, Pinaki Nath Chowdhury, Soumitri Chattopadhyay, Tao Xiang, and Yi-Zhe Song. Exploiting unlabelled photos for stronger fine-grained sbir. In *CVPR*, 2023. 2, 17

[104] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM TOG*, 2016. 3, 7, 16, 17, 18

[105] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *ICML*, 2016. 2

[106] Eli Schwartz, Leonid Karlinsky, Rogério Schmidt Feris, Raja Giryes, and Alex M. Bronstein. Baby steps towards few-shot learning with multiple semantics. *arXiv preprint arXiv:1906.01905*, 2019. 2

[107] Yuming Shen, Li Liu, Fumin Shen, and Ling Shao. Zero-shot sketch-image hashing. In *CVPR*, 2018. 2

[108] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017. 2, 3, 4, 5, 14

[109] X. Soria, E. Riba, and A. Sappa. Dense extreme inception network: Towards a robust cnn model for edge detection. In *WACV*, 2020. 2, 16

[110] Zhuo Su, Wenzhe Liu, Zitong Yu, Dewen Hu, Qing Liao, Qi Tian, Matti Pietikäinen, and Li Liu. Pixel difference networks for efficient edge detection. In *ICCV*, 2021. 3, 5, 6, 7, 8, 16, 17

[111] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 3

[112] 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 *AAAI*, 2024. 3, 18

[113] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *CVPR*, 2013. 1, 2

[114] 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 *CVPR*, 2018. 2

[115] Korawat Tanwisuth, Xinjie Fan, Huangjie Zheng, Shujian Zhang, Hao Zhang, Bo Chen, and Mingyuan Zhou. A prototype-oriented framework for unsupervised domain adaptation. In *NeurIPS*, 2021. 2, 3, 5, 6, 8, 14, 17

[116] James Thewlis, Samuel Albanie, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of landmarks by descriptor vector exchange. In *ICCV*, 2019. 1, 2, 3

[117] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 2014. 1, 2

[118] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NeurIPS*, 2016. 2

[119] Alexander Wang, Mengye Ren, and Richard Zemel. Sketchembednet: Learning novel concepts by imitating drawings. In *ICML*, 2021. 2

[120] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *CVPR*, 2020. 3

[121] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM CSur*, 2020. 2

[122] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In *CVPR*, 2018. 2

[123] 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. 1, 3

[124] Lilian Weng. Meta-learning: Learning to learn fast. *lilianweng.github.io*, 2018. 2

[125] Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014. 2

[126] Thomas Wimmer, Peter Wonka, and Maks Ovsjanikov. Back to 3d: Few-shot 3d keypoint detection with back-projected 2d features. In *CVPR*, 2024. 3

[127] Zhe Wu, Li Su, and Qingming Huang. Stacked cross refinement network for edge-aware salient object detection. In *ICCV*, 2019. 5

[128] Minshan Xie, Menghan Xia, and Tien-Tsin Wong. Exploiting aliasing for manga restoration. In *CVPR*, 2021. 2

[129] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *ICCV*, 2015. 3, 5, 8, 16

[130] Chen Xing, Negar Rostamzadeh, Boris N. Oreshkin, and Pedro H. O. Pinheiro. Adaptive cross-modal few-shot learning. In *NeurIPS*, 2019. 2, 3

[131] Peng Xu, Timothy M Hospedales, Qiyue Yin, Yi-Zhe Song, Tao Xiang, and Liang Wang. Deep learning for free-hand sketch: A survey. *IEEE TPAMI*, 2022. 2

[132] Guowei Yan, Zhili Chen, Jimei Yang, and Huamin Wang. Interactive liquid splash modeling by user sketches. *ACM TOG*, 2020. 2

[133] Heng Yang and Marco Pavone. Object pose estimation with statistical guarantees: Conformal keypoint detection and geometric uncertainty propagation. In *CVPR*, 2023. 1, 2

[134] Shuai Yang, Zhangyang Wang, Jiaying Liu, and Zongming Guo. Deep plastic surgery: Robust and controllable image editing with human-drawn sketches. In *ECCV*, 2020. 2

[135] Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. Transpose: Keypoint localization via transformer. In *ICCV*, 2021. 1, 2

[136] Huaxiu Yao, Xian Wu, Zhiqiang Tao, Yaliang Li, Bolin Ding, Rui rui Li, and Zhenhui Li. Automated relational meta-learning. In *ICLR*, 2020. 2

[137] Qingsong Yao, Quan Quan, Li Xiao, and S Kevin Zhou. One-shot medical landmark detection. In *MICCAI*, 2021. 3

[138] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *ECCV*, 2018. 2

[139] Chengxiang Yin, Jian Tang, Zhiyuan Xu, and Yanzhi Wang. Adversarial meta-learning. *arXiv preprint arXiv:1806.03316*, 2018. 2

[140] Jaesik Yoon, Taesup Kim, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn. Bayesian model-agnostic meta-learning. In *NeurIPS*, 2018. 2

[141] Sung Whan Yoon, Jun Seo, and Jaekyun Moon. Tapnet: Neural network augmented with task-adaptive projection for few-shot learning. In *ICML*, 2019. 2

[142] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In *CVPR*, 2016. 2, 3, 17

[143] Song-Hai Zhang, Yuan-Chen Guo, and Qing-Wen Gu. Sketch2model: View-aware 3d modeling from single free-hand sketches. In *CVPR*, 2021. 2

[144] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. Deep supervised cross-modal retrieval. In *CVPR*, 2019. 3

[145] Linjun Zhou, Peng Cui, Shiqiang Yang, Wenwu Zhu, and Qi Tian. Learning to learn image classifiers with visual analogy. In *CVPR*, 2019. 2

[146] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 2

[147] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krahenbuhl. Bottom-up object detection by grouping extreme and center points. In *CVPR*, 2019. 2

[148] Luisa M. Zintgraf, Kyriacos Shiarlis, Vitaly Kurin, Katja Hofmann, and Shimon Whiteson. Fast context adaptation via meta-learning. In *ICML*, 2019. 2