A. Contributions

Our contributions are summarized as follows:

New task. We opened up a new challenging task, Sketch-based Video Object Localization (SVOL). We also identified several challenges that the SVOL task setting brings.

New dataset. We presented a new SVOL dataset curated from the video dataset (ImageNet-VID [60]) and three sketch datasets with different styles (Sketchy [63], TU-Berlin [20], QuickDraw [33]) and provide a benchmark with comparison against the frame-level baselines and several variants.

Strong baseline. We proposed a novel framework named SVANet, equipped with SVOL-tailored designs such as Cross-modal Transformer and per-frame set matching, that serves as a strong baseline on the benchmark: SVANet improves mIoU by 29.4%, 17.7%, 16.8% over the strong counterpart, Sketch-DETR [59], using Sketchy, TU-Berlin, QuickDraw sketch dataset, respectively.

Extensive experiments. We thoroughly investigated the effects of model components with comprehensive ablative studies in various aspects. Last but not least, we demonstrated the strong generalizability of SVANet on unseen datasets and novel categories, which makes sketch as query highly practical in real-world scenarios.

B. Related Work

Our work builds on previous work in several areas, including sketch-based applications, query-based localization, and Transformer architecture.

B.1. Sketch-based Applications

Our work builds on the idea of using sketches as a way to query visual data. Sketch is a universal communication tool that is not bound by age, race, language, or national boundaries. Recently, sketch-based applications have grown at an unprecedented rate due to the widespread use of touch-screen devices such as smartphones and tablets that enable acquiring sketch data much easier than ever. Here are some examples of various applications using sketch [1]:

Image retrieval: a user sketches an object or scene and the system retrieves similar image from a database [3, 18, 61, 82].

Image synthesis: a user sketches an image and the system generates a photorealistic version of the image [10, 31, 74];

Image editing: a user sketches desired changes to an image, and the system makes the changes automatically [52, 80].

Robot interface: a user sketches a task for a robot to perform, such as picking up an object and placing it in a specific location [5, 62].

3D modeling: a user sketches a 3D object and the system generates a 3D model of the object [49, 71].

[1]For a more detailed list of sketch-based applications, we recommend referring to [78].

Augmented reality: a user sketches an object, and the system overlays a 3D model of the object in the real-world environment [32, 36].

Additionally, sketch is particularly effective at representing detailed features of an object like its shape, pattern, and pose. This ability to convey such fine-grained information has made sketches a popular tool in a variety of studies, including image [83], scene [44], and video [79] retrieval. For example, in fine-grained image retrieval, sketch are used as queries to retrieve specific objects within images, such as a specific breed of dog or type of car [64, 66]. This allows users to search for images with specific visual characteristics, such as the shape of a dog's ears or a car's grille, so that objects of the same category can be differentiated.

In the SVOL problem, while sketches have the capability to provide fine-grained information, we opt to focus on category-level object localization, *i.e.*, localization is carried out in a shape- and pose-agnostic manner within the same category. This is because it is not natural to match a static sketch that has a specific shape and pose with objects in a video, whose shape and pose dynamically change over time. Additionally, by focusing on category-level localization, we can take advantage of the abstract nature of sketches. We can identify the location of objects in a video by sketching only some key features of the object, such as the headlights and grille of a car.

B.2. Query-based Localization Tasks

SVOL is related to the literature on object detection and tracking in videos, with added constraint of using a sketch as the query. Query-based object localization is similar to object detection [43, 46, 57, 58] (or video object detection [12, 23, 28, 72, 77, 86]) in that they both aim to locate the bounding boxes of objects in an image (or a video). However, query-based localization is grounded on the given query, rather than pre-defined object classes. Query-based localization tasks have been studied using various query types in diverse dimensions.

Query. **Image** queries can be localized based on appearance similarity, allowing themselves to be easily transferred to other objects with just a few image samples. This desirable property opens up a new avenue for research on one/few-shot localization [22, 29, 50, 70]. However, image queries are hard to acquire in some privacy or security-related situations, making their usage in some applications difficult. **Language** queries, on the other hand, are highly useful given that we just need to describe the objects of interest in natural language. However, its universality is limited since the assumed language (English) may not be familiar to some people (non-native English speakers). As such, when the language is re-targeted, the neural network may require extra learning or translation before providing the query. **Sketch** queries differ from image queries in that they lack rich information such as

color, texture, and background information; most free-hand sketches are composed solely of monochromatic lines, with no texture and context. In addition, since the sketches are drawn by envisioning abstract objects, even the same object may be drawn differently by a different person. These characteristics make sketch-based localization challenging. Nevertheless, we argue that this line of research is valuable since it offers the highest degree of freedom among the three query types and can transcend the language barrier (*i.e*., the sketch of 'cat' can be understood whether or not you are a native English speaker). Our work focuses on the emerging role of sketch in the context of query-based localization.

Dimension. **Temporal** localization aims to identify the temporal span (1D) in which the query object appears in the video. **Spatial** localization attempts to locate all object instances that match the query object within a still image (2D). **Spatio-temporal** localization seeks to locate all object instances that match the query object in every frame of video (3D). Our work belongs to the spatio-temporal category.

Task. The challenging and open-ended nature of the query-based localization problem lends itself to a variety of tasks: image-based localization in natural images [2, 14, 45, 47] and videos [11, 13, 39, 40] (a.k.a., visual object tracking); language-based localization in natural images [15, 16, 24, 30, 35, 81] (a.k.a., visual grounding or referring expression comprehension) and videos [1, 6, 21, 27, 38, 65, 76, 85] (a.k.a., video grounding or natural language video localization); and sketch object localization in natural images [59, 67]. The most relevant tasks to ours are video grounding [21, 38, 65, 85] and sketch object localization in images [59, 67].

Comparison to previous works. In the realm of query-based localization research, various query types and domains have been explored, such as image, language, and video. However, one noticeable gap in the existing literature pertains to the absence of studies focused on sketch queries in the video domain. We seek to address this particular gap in knowledge. To address this research void, we propose a novel task "Sketch-based Video Object Localization". This task is designed to facilitate the precise localization of spatio-temporal object boxes within video content, with the query input provided in the form of a sketch. This novel approach bridges the gap between sketch-based queries and video object localization, opening up new avenues for exploration and advancement in the field.

B.3. Sketch-based Image Object Localization

There are few methods for image object localization using sketch queries [59, 67], and we are the first to propose a sketch-based video object localization approach. We adopt Cross-modal Attention [67] and Sketch-DETR [59] as the image-level baseline in our SVOL benchmark.

Tripathi *et al*., Cross-modal Attention [67] generates object proposals that match the query sketch in an image. This mechanism operates above the off-the-shelf object detection framework, Faster R-CNN [58], in which the key component is region proposal network (RPN). Tripathi *et al*. modifies the RPN structure to integrate the sketch information in order to create object proposals that are relevant to the query sketch. In more detail, feature vectors of different regions in the image feature map are scored using the global sketch representation to determine compatibility. The attention feature is then calculated by multiplying these compatibility scores with image feature maps. These attention feature maps are concatenated with the original feature maps and projected to a lower-dimensional space, which is then input to RPN to yield relevant object proposals. The pooled object proposals are scored using a sketch feature vector to localize the object of interest.

Riba *et al*., Sketch-DETR [59] is built on the DETR [8] architecture. Given a natural image and a query sketch, Sketch-DETR [59] transforms each input with a separate CNN backbone, and generates feature maps for each input modality. They are then fused via concatenation. Specifically, the sketch is inflated by the resolution of the image feature map then projected using a 1×1 convolution operation. The obtained feature map are flattened before being fed into the Transformer encoder-decoder. The final bounding boxes and their respective score are predicted through a shared feed-forward network (FFN).

B.4. Vision and Multimodal Transformers

Transformer [69] is a universal sequence processor with an attention-based architecture that is originally designed for machine translation. The primary components of Transformer are self-attention that captures long-range interactions within a single context and cross-attention that considers token correspondences between two sequences.

Vision Transformers. Beyond natural language processing [7, 17, 54, 55], Transformers have rapidly become the *de facto* standard in a variety of computer vision applications: image recognition [19], object detection [8], panoptic segmentation [73], human object interaction [34], action recognition [37, 75], and object tracking [11]. Among these, it is worth noting that Detection Transformer (DETR) [8] has made a significant breakthrough in the field of object detection by successfully adopting Transformer design and bipartite matching algorithm. By design, DETR eliminates the need for heuristics (*e.g*., non-maximum suppression) in the detection pipeline while leveraging the capability of global relation modeling. Inspired by the recent successes of Transformers, particularly DETR, we view the SVOL task as a set prediction problem and build our model on the Transformer architecture. Furthermore, since videos are a sequence of frames, the Transformer is well-suited to model temporal information of videos.

Multimodal Transformers. Transformers have shown to be particularly effective in multimodal processing due to their ability to selectively attend to relevant information from multiple modalities (*e.g.*, text, image, and audio). This has been demonstrated in various tasks, including image captioning [53], visual question answering [9], natural language video grounding [76], text-to-image synthesis [56], and text-to-speech [41]. Recent studies such as CLIP [53] and DALL-E [56] highlighted the potential of pre-training Transformer-based models on a vast amount of image-text pairs using a contrastive loss. This provides a strong starting point when fine-tuning on downstream tasks, thereby allowing the model to generalize well unseen datasets. Furthermore, Transformers have shown to transfer well across tasks, making them a versatile model for various multimodal processing tasks [42, 48]. These properties of Transformer make itself a strong candidate for the SVOL task, where the intricate relationship between the query sketch and video objects must be modeled to bridge the gap between natural video and sketches with various styles.

C. Preliminary: Transformer

We build our Cross-modal Transformer (CMT) on top of Transformer design [2], which: (i) densely relates every pair of elements in the input sequence; (ii) captures long-range context with minimal inductive bias (compared to CNNs or RNNs); (iii) effectively models interaction between multi-modal (cross-domain) sequences. These desirable properties of Transformer makes itself well-suited for the CMT design.

The common practice is to use attention mechanism with residual connection, dropout, and layer normalization. Attention in the general $\mathbf{QKV}$ form is a popular yet strong mechanism for neural systems. Given that attention operations are key building blocks of CMT, we first briefly discuss their general form.

C.1. QKV Attention

Given input sequences $\mathbf{q} \in \mathbb{R}^{L_1 \times D}$, $\mathbf{k} \in \mathbb{R}^{L_2 \times D}$ and $\mathbf{v} \in \mathbb{R}^{L_2 \times D}$, we project them into separate embedding spaces. We call the embedded representations as query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$).

$$\mathbf{Q} = (\mathbf{q} + \mathtt{pos}_q)\mathbf{W}_\mathbf{q}, \tag{1}$$

$$\mathbf{K} = (\mathbf{k} + \mathtt{pos}_k)\mathbf{W}_\mathbf{k}, \tag{2}$$

$$\mathbf{V} = \mathbf{v}\mathbf{W}_\mathbf{v}, \tag{3}$$

where $\mathbf{W}_\mathbf{q}, \mathbf{W}_\mathbf{k}, \mathbf{W}_\mathbf{v} \in \mathbb{R}^{D \times D_h}$ are learnable weights. Since the Transformer is inherently permutation-invariant w.r.t input sequence, we add positional encoding $\mathtt{pos}_q \in \mathbb{R}^{L_1 \times D}$ and $\mathtt{pos}_k \in \mathbb{R}^{L_2 \times D}$ (fixed absolute encoding to represent positions using sine and cosine functions of different frequencies) to embedded sequences.

The attention weights $\mathbf{A}_{ij}$ are computed by comparing two elements of the sequence (dot products) to their respective query $\mathbf{Q}_i$ and key $\mathbf{K}_j$ representations, normalized by D_h.

$$\mathbf{A} = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{D_h}}\right), \tag{4}$$

Finally, we calculate a weighted sum over all value representation $\mathbf{V}$.

$$\mathrm{Att}(\mathbf{q}, \mathbf{k}, \mathbf{v}) = \mathbf{A}\mathbf{V}. \tag{5}$$

We call this operation as *Self-Attention* for the special case where $\mathbf{q}$, $\mathbf{k}$, and $\mathbf{v}$ are all the same.

C.2. Multi-Head Attention

Multi-Head Attention (MHA) allows the model to jointly attend to information from different representation subspaces at different positions. It is a simple extension of Attention in which several Attentions heads are executed in parallel followed by a projection of their concatenated outputs. To maintain the computed value and the number of parameters constant when changing the number of heads k, D_h is typically set to D/k.

$$\begin{aligned}&\mathrm{MHA}(\mathbf{q}, \mathbf{k}, \mathbf{v}) = \\ &[\mathrm{Att}_1(\mathbf{q}, \mathbf{k}, \mathbf{v}); \cdots ; \mathrm{Att}_k(\mathbf{q}, \mathbf{k}, \mathbf{v});]\mathbf{W}_{\mathbf{MHA}},\end{aligned} \tag{6}$$

where [;] denotes concatenation on the channel axis and $\mathbf{W}_{\mathbf{MHA}} \in \mathbb{R}^{k \cdot D_h \times D}$ is learnable weight.

D. SVOL Dataset & Analysis

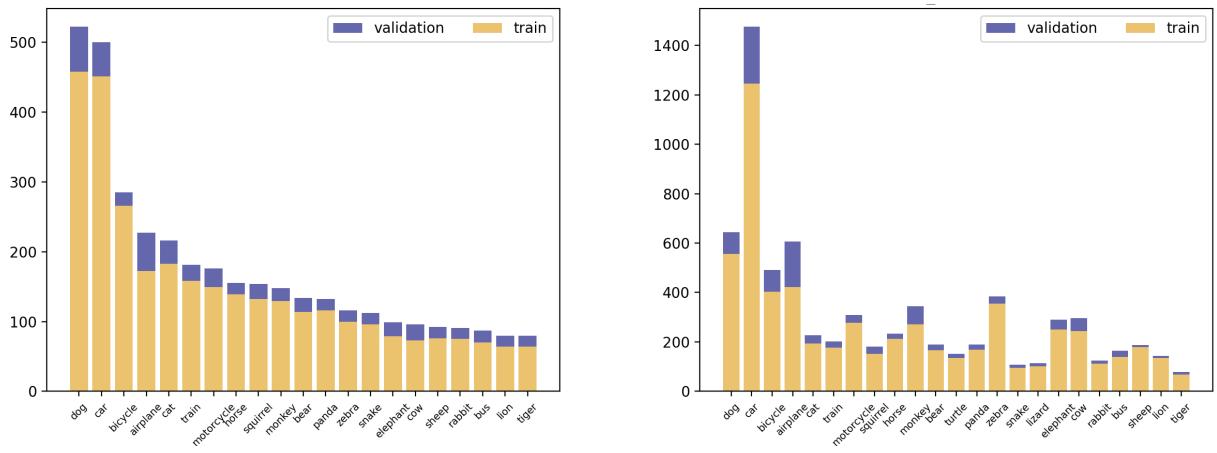
The SVOL dataset is built on multiple datasets [20, 33, 60, 63]. We consider only the categories that intersects between the video dataset [60] and the sketch datasets [20, 33, 63].

D.1. SVOL Dataset

ImageNet-VID [60] is built for video object detection task. It contains 5,354 snippets (train/val/test split is 3,852/555/937) that are annotated with 30 object categories, including vehicles (*e.g.*, airplane, bus, *etc.*) and animals (*e.g.*, bird, dog, *etc.*). Each object instance is annotated in the form of {video name, frame number, class label, instance id, bounding box}. We use a validation set for evaluation since test annotations are not publicly available.

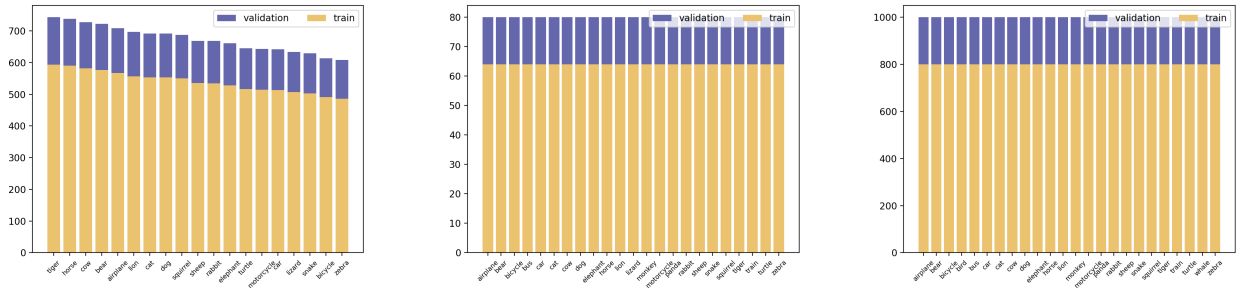
Sketchy [63] is a large-scale collection of sketch-photo pairs that has 75,471 sketches belonging to 125 categories. Drawers are not allowed to directly trace objects; rather, sketches are drawn after seeing specific photographic objects. This forces the drawers to sketch from memory in the

[2] We leave an original paper as a reference [69] for further details of Transformer building blocks.



(a) ID-agnostic (b) ID-specific

Figure 1. **Video dataset [60] class distribution** when objects of the same category are (a) counted as a whole (*i.e.*, *agnostic* to instance ID) or (b) counted individually (*i.e.*, *sensitive* to instance ID). The x-axis denotes the category and y-axis denotes the frequency.



(a) Sketchy (b) TU-Berlin (c) QuickDraw

Figure 2. **Sketch dataset class distribution.** (a) Sketchy [63], (b) TU-Berlin [20], and (c) QuickDraw [33]. The x-axis denotes the class and y-axis denotes the frequency.

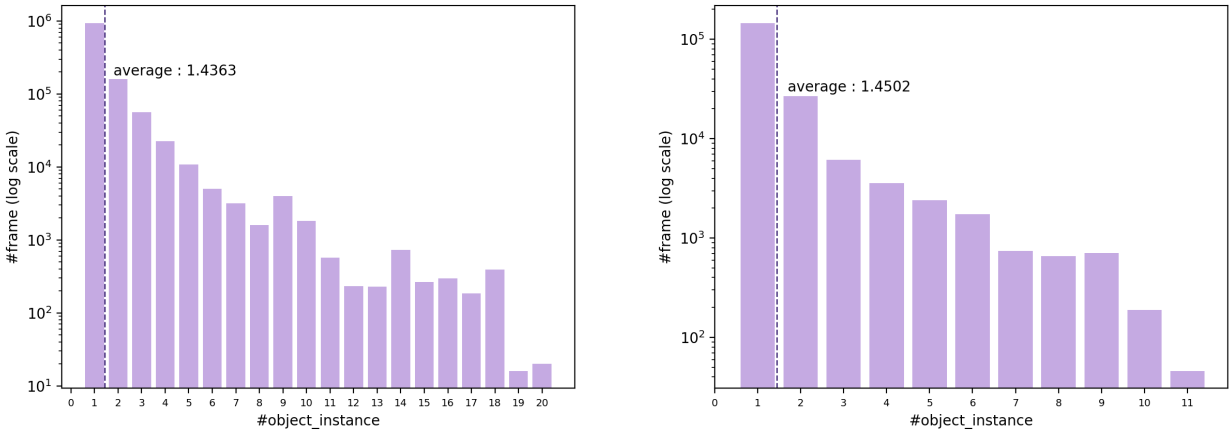
same way that a user of sketch-based image retrieval systems [18, 61, 82] would draw from a mental image of the desired object. 23 categories overlap with ImageNet-VID: airplane, bear, bicycle, car, cat, cow, dog, elephant, horse, lion, lizard, motorcycle, rabbit, sheep, snake, squirrel, tiger, turtle, zebra.

TU-Berlin [20] is a crowd-sourced sketch dataset composed of 20,000 unique sketches with 250 categories. The sketches are uniformly distributed over 250 object categories, which exhaustively cover the vast majority of objects seen in daily life. The median drawing time for each sketch is 86 seconds. Due to the low quality of some sketches, humans correctly identify just 73% of these hand-drawings. 21 categories overlap with ImageNet-VID: airplane, bear, bicycle, bus, car, cat, cow, dog, elephant, horse, lion, monkey, motorcycle, panda, rabbit, sheep, snake, squirrel, tiger, train, zebra.

QuickDraw [33] is a huge collection of 50 million sketches organized into 345 categories. Over 15 million players have contributed millions of sketches playing "Quick, Draw!" game [3], where a neural network tries to guess the sketches. The players are asked to draw a sketch of a given category in 20 seconds while the computer attempts to classify them. The way the sketches are collected results in a high degree of variety in the dataset, although most sketches are of low quality due to time limit. 24 categories overlap with ImageNet-VID.

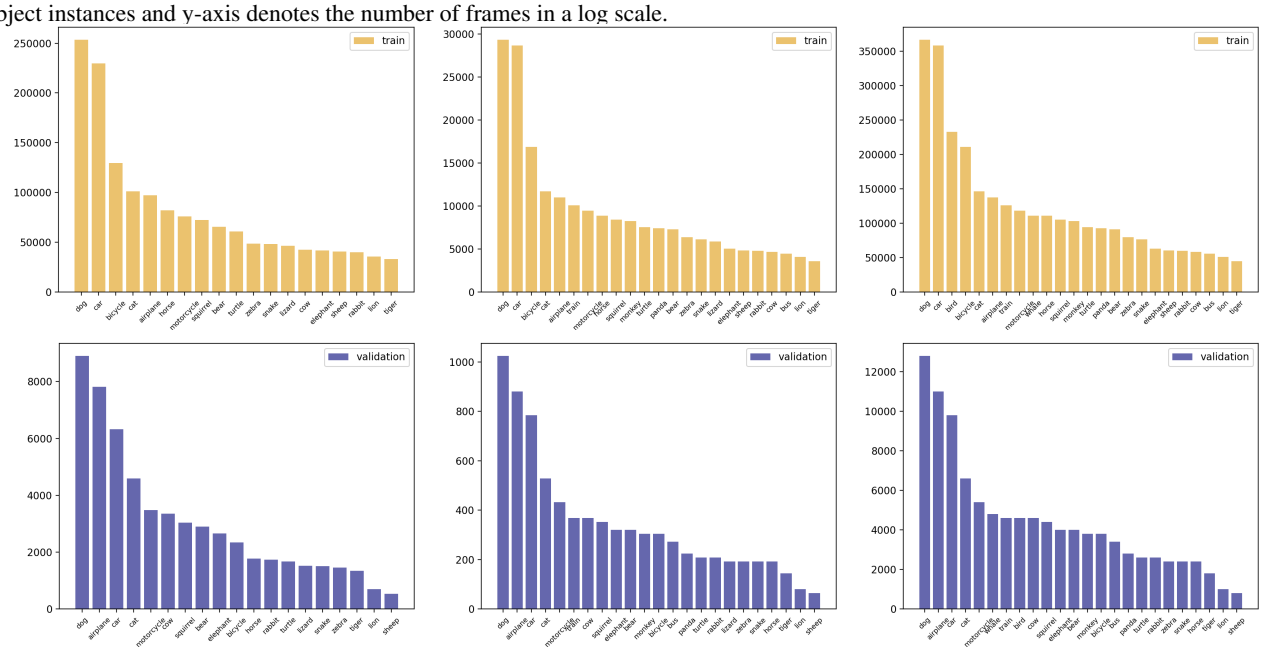
Video ∩ Sketch	Categories
ImageNet-VID ∩ Sketchy (19 classes)	airplane, bear, bicycle, car, cat, cow, dog, elephant, horse, lion, lizard, motorcycle, rabbit, sheep, snake, squirrel, tiger, turtle, zebra
ImageNet-VID ∩ TU-Berlin (21 classes)	airplane, bear, bicycle, bus, car, cat, cow, dog, elephant, horse, lion, monkey, motorcycle, panda, rabbit, sheep, snake, squirrel, tiger, train, zebra
ImageNet-VID ∩ QuickDraw (24 classes)	airplane, bear, bicycle, bird, bus, car, cat, cow, dog, elephant, horse, lion, monkey, motorcycle, panda, rabbit, sheep, snake, squirrel, tiger, train, turtle, whale, zebra

[3] https://quickdraw.withgoogle.com/



(a) Train (b) Val

Figure 3. **Number of object instances per frame** in ImageNet-VID [60] (a) train and (b) validation split. The x-axis denotes the number of object instances and y-axis denotes the number of frames in a log scale.



(a) **VID-Sketchy** (b) **VID-TU-Berlin** (c) **VID-QuickDraw**

Figure 4. **Class distribution of video-sketch pair.** (a) VID-Sketchy, (b) VID-TU-Berlin, and (c) VID-QuickDraw. The x-axis denotes the class and y-axis denotes the frequency.

D.2. SVOL Data Analysis

Dataset	Split	min	max	mean	median
ImageNet-VID	Train	6	5492	290.6	180
	Val	11	2898	317.3	232

D.2.1 Frame Length Distribution

We use the video dataset from the ImageNet-VID dataset [60]. The train split has 3,862 videos that are fully annotated with the 30 object categories, yielding 866,870 bounding boxes for 1,122,397 frames. In validation split, 555 videos are fully annotated with the 30 object categories, resulting in 135,949 bounding box annotations for 176,126 frames. We summarize the statistics for the frame length distribution of ImageNet-VID dataset below.

D.2.2 Class Histogram

Video dataset. We show the class histogram of ImageNet-VID [60] dataset in Fig. 1. Here, *ID-specific* refers to taking into account the identity (ID) of an object instance when counting the number, whereas *ID-agnostic* refers to not taking it into account. In *ID-agnostic*, the number is counted only once even if multiple object instances belonging to the same category appear in a video. For example, in the case

of "car", the number is around 500 without considering the track-id, but exceeds 1,400 with considering the track-id. This indicates that there are many scenes in the video in which multiple "car" object instances appear concurrently. We count only the object categories that are common in both the video and sketch datasets. The statistics for the class distribution of the SVOL video dataset are summarized below.

Dataset	Track-ID	Split	min	max	mean	median
ImageNet-VID	id-specific	Train	67	1246	276.8	194
		Val	9	229	47.7	29
	id-agnostic	Train	56	458	151.8	118
		Val	4	64	21.8	19

Sketch dataset. We show the class histograms of Sketchy [63] and TU-Berlin [20] datasets in Fig. 2. The QuickDraw [33] dataset has 1,000 sketch images per class are uniformly distributed and all train/val splits are 800/200. The number of sketches per class is relatively evenly distributed in the Sketchy dataset, however there is an imbalance between classes in the TU-Berlin dataset. The following table summarizes the statistics for the class distribution of SVOL sketch datasets.

Dataset	Split	min	max	mean	median
Sketchy	Train	486	594	539.5	535
	Val	122	149	135.4	134
TU-Berlin	Train	64	458	150.7	116
	Val	16	64	24.7	19
QuickDraw	Train	800	800	800	800
	Val	200	200	200	200

Number of object instances. Fig. 3 shows the distribution of the number of object instances per frame in the SVOL video dataset. The average object instances per video is 1.4363 in the train split and 1.4502 in the validation split.

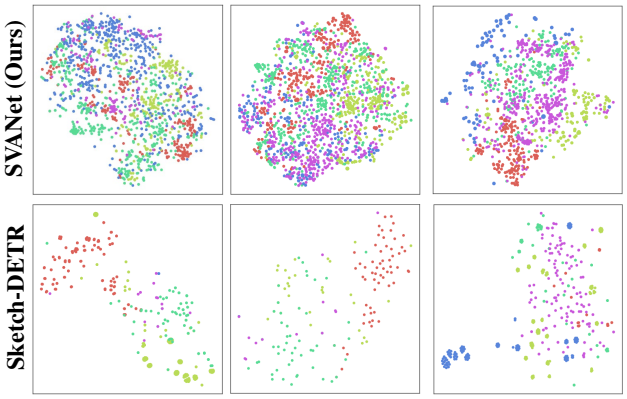
D.3. Data Curation

The SVOL dataset is made up of a combination of video dataset and sketch datasets. To ensure that the SVOL evaluation set remain unseen in the training phase, we split videos and sketches into training and evaluation sets, respectively, and then construct a training set of SVOL with a combination of both training sets, and an evaluation set of SVOL with a combination of both evaluation sets. This guarantees the models to be evaluated on video-sketch pairs that are totally unseen throughout the training phase. While this setting is most closest to the actual environment in which the SVOL system operates, it requires the model to be generalized to both videos and sketches.

Formally, let $\{\mathcal{V}_{tr}, \mathcal{V}_{ev}\}$ train/eval video datasets, $\{\mathcal{S}_{tr}, \mathcal{S}_{ev}\}$ train/eval sketch datasets, and $\{\mathbb{C}_{\mathcal{V}}, \mathbb{C}_{\mathcal{S}}\}$ video/sketch category sets. For all categories that are common for video and sketch datasets, *i.e.*, $\forall c \in \mathbb{C}_{\mathcal{V}} \cap \mathbb{C}_{\mathcal{S}}$, we construct SVOL train set by pairing $\mathcal{V}_{tr}$ and $\mathcal{S}_{tr}$, and SVOL eval set with $\mathcal{V}_{ev}$ and $\mathcal{S}_{ev}$. Only video and sketch are paired when they have the same class label. The number of pairs generated for each video-sketch datasets are summarized in the table below (Fig. 4 shows the per-category distribution).

Split	VID-Sketchy	VID-TU-Berlin	VID-QuickDraw
Train	1,545,801	215,040	2,958,400
Eval	57,660	7,952	10,6400

In practice, we only use videos that contain at least one query sketch object within 32 frames uniformly sampled from the video, and bounding box annotations that correspond to the sketch object are regarded as the ground truths for that pairing. A video can be paired more than once since it can contain multiple objects. We note that the class label is only a means for pairing and is not considered in actual training.



(a) S→T (b) S→Q (c) SC→USC (w/ S)

Figure 5. **Feature distribution of SVANet *vs*. Sketch-DETR [59]** when transfer is performed for the cases of (a) Sketchy→TU-Berlin, (b) Sketchy→QuickDraw, and (c) Seen→Unseen Categories with the Sketchy dataset. Each data point represents the last hidden state of the CMT, and the color indicates the category it belongs to. We plot samples of 5 random categories with a confidence score higher than 0.9.

E. Additional Qualitative Results

Feature distribution in transfer evaluation. In Fig. 5, we perform transfer evaluation and visualize the feature distribution of SVANet and that of Sketch-DETR using t-SNE [68]. Here, (a) and (b) depict the results in dataset-level transfer, whereas (c) represents the result of category-level transfer. In other words, for (a) and (b), models trained on the Sketchy dataset (TU-Berlin dataset for (b)) were employed to map the feature distributions of samples from the TU-Berlin dataset (QuickDraw dataset for (b)). On the other hand, for (c), models trained on certain categories were utilized to visualize the feature distributions of samples from previously unseen categories. Compared to Sketch-DETR, SVANet appears to be nicely clustered when transferred to unseen datasets, which implies that SVANet effectively captures class-discriminative representations. When transferred to unseen categories, SVANet embeds the same category into a similar subspace, demonstrating that learned sketch-video mapping can generalize well. Moreover, SVANet shows

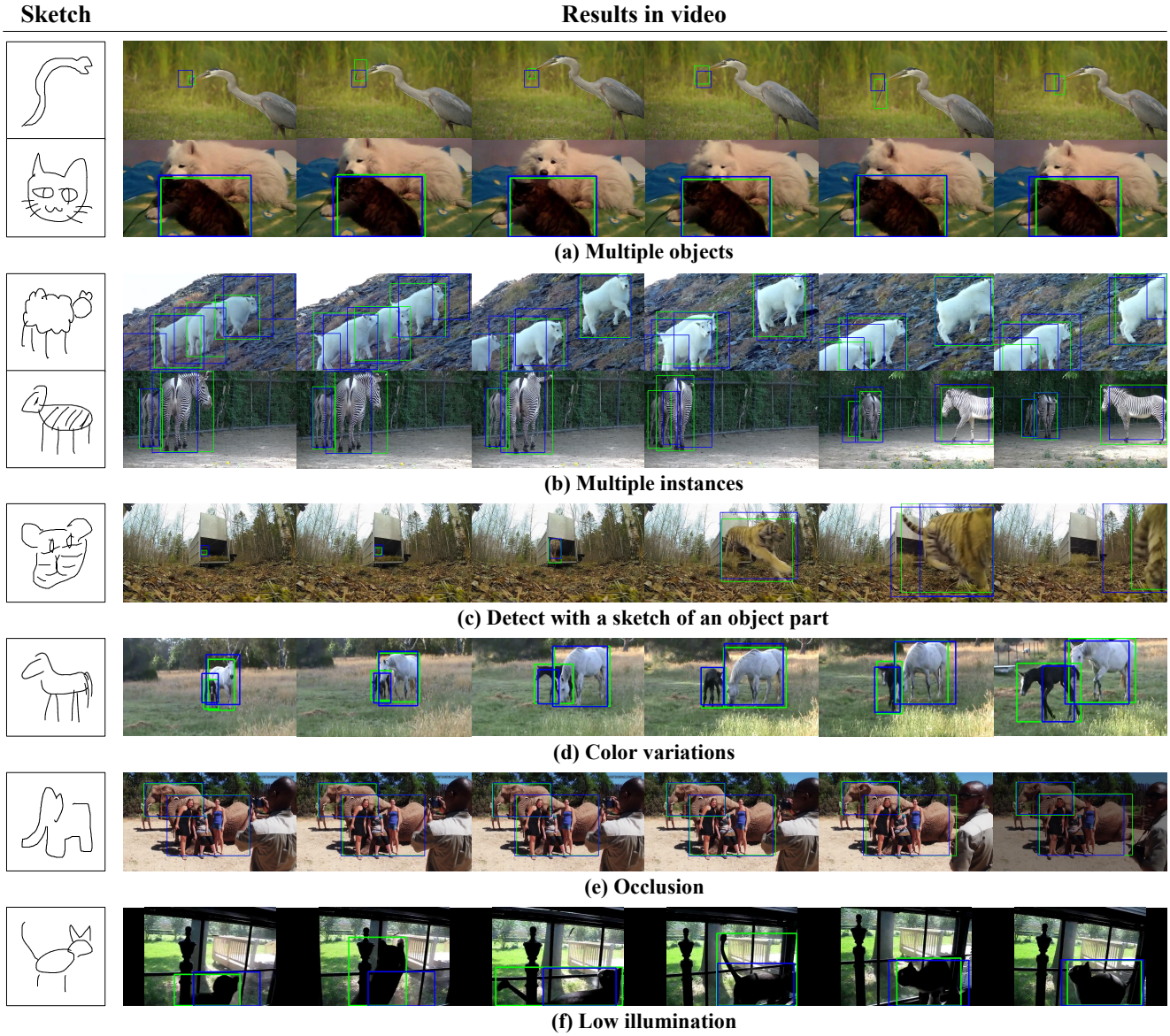


Figure 6. **Success cases** of SVANet on QuickDraw dataset. Green and blue boxes represent ground truths and predictions, respectively.

denser distribution than Sketch-DETR, *i.e.*, only a few data points reach the threshold confidence 0.9 in Sketch-DETR, indicating that our SVANet produces more reliable predictions.

SVOL results: success cases. Fig. 6 shows success cases of SVANet. Our system successfully recognizes the objects that correspond to the query sketch and accurately localizes their bounding boxes in a variety of challenging conditions: (a) various objects appears in a video; (b) multiple object instances with different pose and shape appear in a video; (c) only sketch of a part (face) of object is given; (d) the target objects have different colors; (e) the target objects are occluded by other objects; (f) bad illumination condition.

SVOL results: failure cases. Fig. 7 shows failure cases of SVANet. SVANet suffers particularly when the target object: (a) appears for a very short time (almost 1 or 2 frames out of 32 frames); (b) is too small, and there are numerous distracting factors; (c) is small and moves quickly; (d) is non-salient (here, the target object is a car, but a motorcycle, is detected); (e) is similar to the background.

F. Discussion

F.1. Why Sketch Query?

Sketch query can be more flexible and efficient than language or image query as it allows for more natural and intuitive user input. With sketch query, users can quickly and easily provide a rough sketch of the object they are

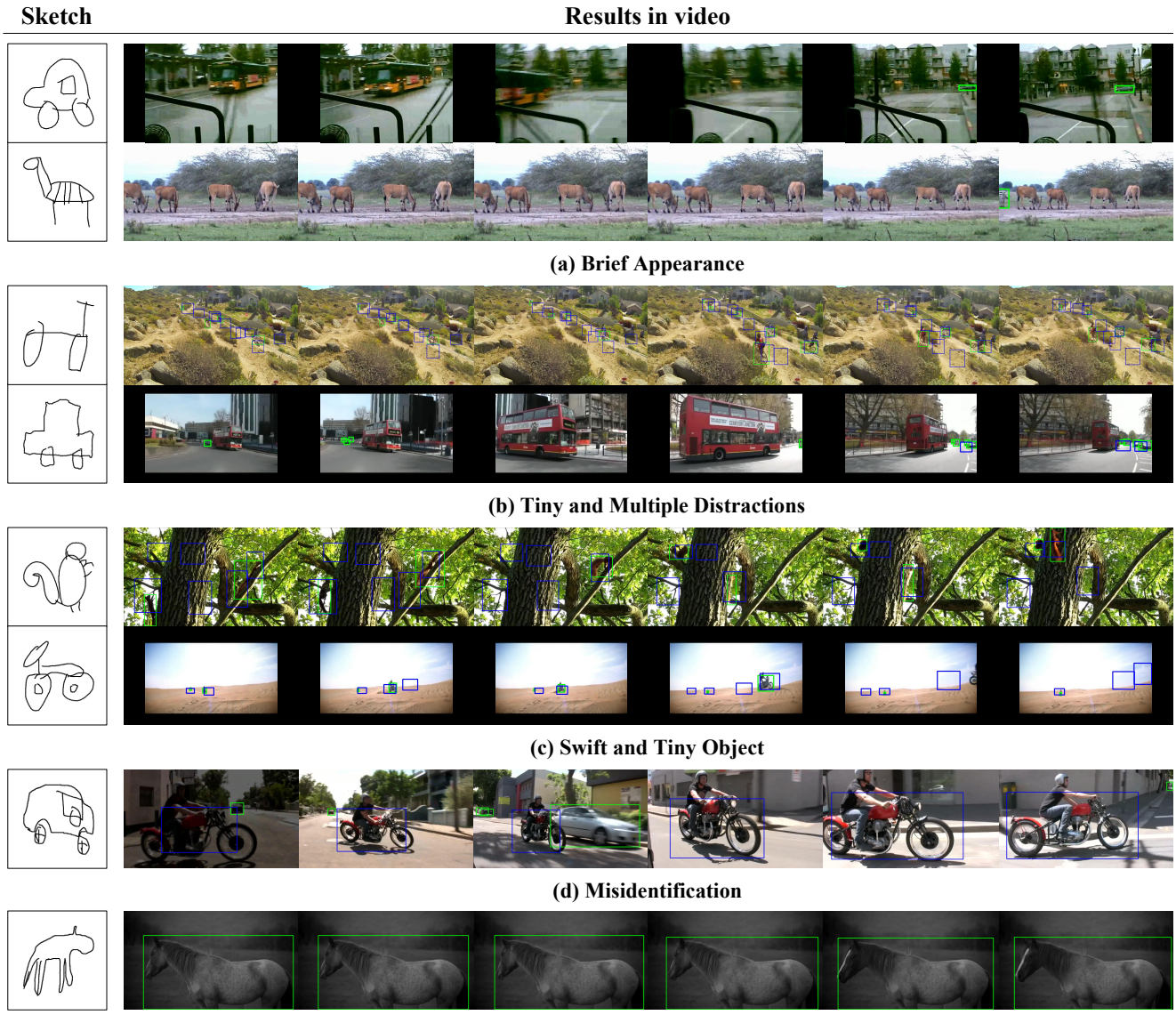


Figure 7. **Failure cases** of SVANet on QuickDraw dataset. Green and blue boxes represent ground truths and predictions, respectively.

looking for, rather than having to use specific keywords or search through a pre-existing database of images. This can make it easier for users to find the specific object they are looking for, especially if the image is not easily describable with keywords or if the image does not exist in a pre-existing database. Moreover, sketch query can transcend the language barrier, and can be less prone to ambiguity and errors as the user is able to provide a visual representation of the desired object. On the other hand, language query requires additional translation when the user's language changes (*e.g.*, English $\rightarrow$ French). Additionally, since sketch queries are basically embodiments of real-world objects, the model inherently learns over the visual similarity between the query sketch and the video objects. Therefore, the model can leverage such inductive bias of appearance matching for unseen categories. Sketch query offers a great degree of freedom and can overcome several limitations that other queries (*e.g.*, language or image) may include. There are several more advantages in using sketch query:

1. As the use of touch screen devices (*e.g.*, smartphones, tablets) has recently increased, acquisition of sketch data has become easier.

2. Sketch is not bound by the user's age, race, and nation, and even those with language difficulties can communicate their thoughts;

3. Our model is effective even with a low-quality sketch (*e.g.*, QuickDraw), therefore users are not required to

draw well;

F.2. Limitations

While we believe that focusing on category-level localization in the SVOL problem can take advantage of the abstract nature of sketches, it is also important to consider the limitations of this setting. In this setting, the system may lose nuanced understanding of sketches that could be useful for precisely identifying objects of the same category with different details. For example, if we want to localize a car of a specific make and model, the system may not be able to do so accurately as it is not explicitly taught to differentiate objects within the same category during training. In order to improve the versatility of the SVOL system, future research may investigate on incorporating fine-grained data sources to differentiate objects within the same category.

We also recognize that transfer performance for unseen categories is still far from enough, yet this shows that SVOL is a challenging problem and suggests that better solutions should be found. We hope our findings and analysis will encourage further research in this direction.

F.3. Future Work

We hope future work will develop approaches for the following.

SVOL in large-scale video collection. On an online video platform, users often need to quickly and efficiently find the location of a specific object of interest amid large-scale video collections. In order to be practical in such situations, the SVOL system should be able to retrieve relevant videos, and accurately localize the target objects within the set of retrieved videos. This is similar to the setting for video corpus moment retrieval [84]. Such a system could greatly enhance the user experience by allowing them to quickly locate the desired object within the video corpus, making the process of finding relevant information faster and more efficient. Although it is beyond the scope of our current work, we believe it to be promising area for future research.

Domain adaptation methods. The significant difference in the appearance and structure between sketches and natural videos poses a challenge for the SVOL system to accurately match them. To alleviate this issue, various domain adaptation techniques [25, 26, 51] can be employed. These techniques aim to align the feature representations of the sketches and natural videos, thus reducing the domain gap. By utilizing these techniques, we anticipate further improvements in the performance of the SVOL system.

Fine-grained SVOL. In this work, we define the SVOL task to be agnostic to shape and pose within the same class, allowing us to localize objects in a video by sketching only key features that are unique and distinctive to that object, such as the ears, eyes, and tails of a cat. This setting has the advantage of being able to identify and locate the object in the video, even if the object's shape and pose change over time. However, this setting also has its limitations, as it may miss important details that could be useful for differentiating objects within the same class. Therefore, it may be worth exploring a more fine-grained approach to the SVOL problem, by focusing on detailed instance-level information such as shape, pattern, and pose, which can be used to distinguish objects of the same category [83]. This shape- and pose-specific approach, however, also comes with its own challenges. For instance, it may be difficult to match a still sketch to a moving object in a video, as the shape or pose of objects continues to change over time. Thus, to make this approach work effectively, it is essential to have a suitable data pairing that takes into account the dynamic nature of the video. Additionally, further research could explore ways to effectively balance between leveraging the abstract nature of sketches and preserving enough fine-grained details for precise object identification.

On-the-fly SVOL. As opposed to our SVOL setting, which requires a *complete* sketch to be drawn before localization can begin, the "on-the-fly" setting allows for localization to start as soon as the user begins drawing [4]. This approach utilizes each stroke that is drawn in real-time to match it to objects in the video. This allows for sketch-object matching with an *incomplete* sketch (*i.e.*, just a few strokes), which can greatly reduce the time and effort required to draw an accurate sketch. Furthermore, the system can provide immediate feedback based on the ongoing localization results as the user continues to draw, allowing for a more efficient and user-friendly experience. It can help the user to understand how well their sketch is matching with the objects in the video and make adjustments accordingly. This can also make the task of drawing accurate sketches more manageable for users with less experience or skill.

F.4. Broader Impacts

Our SVANet makes predictions based on learned statistics of the collected dataset, which may reflect biases present the data, including ones with negative societal impacts. The predictions may not be accurate, thus users exercise caution and should not rely solely on them in real-world applications and it is recommended to use it in conjunction with other forms of analysis and decision-making. Further consideration is warranted regarding this issue.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing Moments in Video With Natural Language. In *ICCV*, pages 5803–5812, 2017. 2

[2] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded Up Robust Features. In *ECCV*, pages 404–417, 2006. 2

[3] 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*, pages 2293–2302, 2022. 1

[4] 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*, pages 9779–9788, 2020. 9

[5] Federico Boniardi, Abhinav Valada, Wolfram Burgard, and Gian Diego Tipaldi. Autonomous Indoor Robot Navigation Using a Sketch Interface for Drawing Maps and Routes. In *ICRA*, pages 2896–2901. IEEE, 2016. 1

[6] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-To-End Referring Video Object Segmentation With Multimodal Transformers. In *CVPR*, pages 4985–4995, 2022. 2

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language Models Are Few-Shot Learners. In *NeurIPS*, pages 1877–1901, 2020. 2

[8] 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. 2

[9] Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. WebQA: Multihop and Multimodal QA. In *CVPR*, pages 16495–16504, 2022. 3

[10] Wengling Chen and James Hays. Sketchygan: Towards Diverse and Realistic Sketch to Image Synthesis. In *CVPR*, pages 9416–9425, 2018. 1

[11] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer Tracking. In *CVPR*, pages 8126–8135, 2021. 2

[12] Yihong Chen, Yue Cao, Han Hu, and Liwei Wang. End-to-End Video Object Detection With Spatial-Temporal Transformers. In *CVPR*, pages 10337–10346, 2020. 1

[13] Zedu Chen, Bineng Zhong, Guorong Li, Shengping Zhang, and Rongrong Ji. Siamese Box Adaptive Network for Visual Tracking. In *CVPR*, pages 6668–6677, 2020. 2

[14] Navneet Dalal and Bill Triggs. Histograms of Oriented Gradients for Human Detection. In *CVPR*, pages 886–893, 2005. 2

[15] Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, and Mingkui Tan. Visual Grounding via Accumulated Attention. In *CVPR*, pages 7746–7755, 2018. 2

[16] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. TransVG: End-to-End Visual Grounding With Transformers. In *ICCV*, pages 1769–1779, 2021. 2

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[18] Sounak Dey, Pau Riba, Anjan Dutta, Josep Llados, and Yi-Zhe Song. Doodle To Search: Practical Zero-Shot Sketch-Based Image Retrieval. In *CVPR*, pages 2179–2188, 2019. 1, 4

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[20] Mathias Eitz, James Hays, and Marc Alexa. How Do Humans Sketch Objects? *ACM TOG*, pages 1–10, 2012. 1, 3, 4, 6

[21] Hehe Fan and Yi Yang. Person Tube Retrieval via Language Description. In *AAAI*, pages 10754–10761, 2020. 2

[22] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-Shot Object Detection With Attention-RPN and Multi-Relation Detector. In *CVPR*, pages 4013–4022, 2020. 1

[23] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect To Track and Track To Detect. In *ICCV*, pages 3038–3046, 2017. 1

[24] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. *arXiv preprint arXiv:1606.01847*, 2016. 2

[25] Yaroslav Ganin and Victor Lempitsky. Unsupervised Domain Adaptation by Backpropagation. In *ICML*, pages 1180–1189, 2015. 9

[26] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-Adversarial Training of Neural Networks. *JMLR*, pages 2096–2030, 2016. 9

[27] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: Temporal Activity Localization via Language Query. In *ICCV*, pages 5267–5275, 2017. 2

[28] Wei Han, Pooya Khorrami, Tom Le Paine, Prajit Ramachandran, Mohammad Babaeizadeh, Honghui Shi, Jianan Li, Shuicheng Yan, and Thomas S Huang. Seq-NMS for Video Object Detection. *arXiv preprint arXiv:1602.08465*, 2016. 1

[29] Ting-I Hsieh, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. One-Shot Object Detection With Co-Attention and Co-Excitation. In *NeurIPS*, 2019. 1

[30] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural Language Object Retrieval. In *CVPR*, pages 4555–4564, 2016. 2

[31] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-Image Translation With Conditional Adversarial Networks. In *CVPR*, pages 1125–1134, 2017. 1

[32] Ying Jiang, Congyi Zhang, Hongbo Fu, Alberto Cannavò, Fabrizio Lamberti, Henry YK Lau, and Wenping Wang. HandPainter-3D Sketching in VR With Hand-Based Physical Proxy. In *CHI*, pages 1–13, 2021. 1

[33] Jonas Jongejan, Henry Rowley, Takashi Kawashima, Jongmin Kim, and Nick Fox-Gieg. The Quick, Draw!-AI Experiment. *Mount View, CA, accessed Feb*, page 4, 2016. 1, 3, 4, 6

[34] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J Kim. Hotr: End-to-End Human-Object Interaction Detection With Transformers. In *CVPR*, pages 74–83, 2021. 2

[35] Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler. What Are You Talking About? Text-to-Image Coreference. In *CVPR*, pages 3558–3565, 2014. 2

[36] Kin Chung Kwan and Hongbo Fu. Mobi3dsketch: 3D Sketching in Mobile AR. In *CHI*, pages 1–11, 2019. 1

[37] Sumin Lee, Sangmin Woo, Yeonju Park, Muhammad Adi Nugroho, and Changick Kim. Modality Mixer for Multimodal Action Recognition. In *WACV*, pages 3298–3307, 2023. 2

[38] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. TVQA+: Spatio-Temporal Grounding for Video Question Answering. *arXiv preprint arXiv:1904.11574*, 2019. 2

[39] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of Siamese Visual Tracking With Very Deep Networks. In *CVPR*, pages 4282–4291, 2019. 2

[40] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High Performance Visual Tracking With Siamese Region Proposal Network. In *CVPR*, pages 8971–8980, 2018. 2

[41] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. Neural Speech Synthesis With Transformer Network. In *AAAI*, volume 33, pages 6706–6713, 2019. 3

[42] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. OSCAR: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *ECCV*, pages 121–137. Springer, 2020. 3

[43] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 1

[44] Fang Liu, Xiaoming Deng, Changqing Zou, Yu-Kun Lai, Keqi Chen, Ran Zuo, Cuixia Ma, Yong-Jin Liu, and Hongan Wang. Scenesketcher-V2: Fine-Grained Scene-Level Sketch-Based Image Retrieval Using Adaptive GCNs. *TIP*, 31:3737–3751, 2022. 1

[45] Jialu Liu. Image Retrieval Based on Bag-of-Words Model. *arXiv preprint arXiv:1304.5168*, 2013. 2

[46] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single Shot Multibox Detector. In *ECCV*, pages 21–37, 2016. 1

[47] David G Lowe. Distinctive Image Features From Scale-Invariant Keypoints. *IJCV*, pages 91–110, 2004. 2

[48] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *NeurIPS*, 32, 2019. 3

[49] Zhaoliang Lun, Matheus Gadelha, Evangelos Kalogerakis, Subhransu Maji, and Rui Wang. 3D Shape Reconstruction From Sketches via Multi-View Convolutional Networks. In *3DV*, pages 67–77. IEEE, 2017. 1

[50] Anton Osokin, Denis Sumin, and Vasily Lomakin. OS2D: One-Stage One-Shot Object Detection by Matching Anchor Features. In *ECCV*, pages 635–652, 2020. 1

[51] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment Matching for Multi-Source Domain Adaptation. In *ICCV*, pages 1406–1415, 2019. 9

[52] Tiziano Portenier, Qiyang Hu, Attila Szabo, Siavash Arjomand Bigdeli, Paolo Favaro, and Matthias Zwicker. Faceshop: Deep Sketch-Based Face Image Editing. *arXiv preprint arXiv:1804.08972*, 2018. 1

[53] 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*, pages 8748–8763, 2021. 3

[54] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training. 2018. 2

[55] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language Models Are Unsupervised Multitask Learners. *OpenAI blog*, page 9, 2019. 2

[56] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-Shot Text-To-Image Generation. In *ICML*, pages 8821–8831. PMLR, 2021. 3

[57] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *CVPR*, pages 779–788, 2016. 1

[58] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection With Region Proposal Networks. In *NeurIPS*, 2015. 1, 2

[59] Pau Riba, Sounak Dey, Ali Furkan Biten, and Josep Llados. Localizing Infinity-Shaped Fishes: Sketch-Guided Object Localization in the Wild. *arXiv preprint arXiv:2109.11874*, 2021. 1, 2, 6

[60] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet Large Scale Visual Recognition Challenge. *IJCV*, pages 211–252, 2015. 1, 3, 4, 5

[61] Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Stylemeup: Towards Style-Agnostic Sketch-Based Image Retrieval. In *CVPR*, pages 8504–8513, 2021. 1, 4

[62] Daisuke Sakamoto, Koichiro Honda, Masahiko Inami, and Takeo Igarashi. Sketch and Run: A Stroke-Based Interface for Home Robots. In *CHI*, pages 197–200, 2009. 1

[63] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The Sketchy Database: Learning To Retrieve Badly Drawn Bunnies. *ACM TOG*, pages 1–12, 2016. 1, 3, 4, 6

[64] Jifei Song, Qian Yu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Deep Spatial-Semantic Attention for Fine-Grained Sketch-Based Image Retrieval. In *ICCV*, pages 5551–5560, 2017. 1

[65] Rui Su, Qian Yu, and Dong Xu. Stvgbert: A Visual-Linguistic Transformer Based Framework for Spatio-Temporal Video Grounding. In *ICCV*, pages 1533–1542, 2021. 2

[66] Haifeng Sun, Jiaqing Xu, Jingyu Wang, Qi Qi, Ce Ge, and Jianxin Liao. DLI-Net: Dual Local Interaction Network for Fine-Grained Sketch-Based Image Retrieval. *TCSVT*, 32(10):7177–7189, 2022. 1

[67] Aditay Tripathi, Rajath R Dani, Anand Mishra, and Anirban Chakraborty. Sketch-Guided Object Localization in Natural Images. In *ECCV*, pages 532–547, 2020. 2

[68] Laurens Van der Maaten and Geoffrey Hinton. Visualizing Data Using T-SNE. *JMLR*, 2008. 6

[69] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All You Need. In *NeurIPS*, pages 5998–6008, 2017. 2, 3

[70] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching Networks for One Shot Learning. In *NeurIPS*, 2016. 1

[71] Fang Wang, Le Kang, and Yi Li. Sketch-Based 3D Shape Retrieval Using Convolutional Neural Networks. In *CVPR*, pages 1875–1883, 2015. 1

[72] Han Wang, Jun Tang, Xiaodong Liu, Shanyan Guan, Rong Xie, and Li Song. Ptseformer: Progressive Temporal-Spatial Enhanced Transformer Towards Video Object Detection. In *ECCV*, pages 732–747. Springer, 2022. 1

[73] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-DeepLab: End-to-End Panoptic Segmentation With Mask Transformers. In *CVPR*, pages 5463–5474, 2021. 2

[74] Sheng-Yu Wang, David Bau, and Jun-Yan Zhu. Sketch Your Own Gan. In *ICCV*, pages 14050–14060, 2021. 1

[75] Sangmin Woo, Sumin Lee, Yeonju Park, Muhammad Adi Nugroho, and Changick Kim. Towards Good Practices for Missing Modality Robust Action Recognition. *arXiv preprint arXiv:2211.13916*, 2022. 2

[76] Sangmin Woo, Jinyoung Park, Inyong Koo, Sumin Lee, Minki Jeong, and Changick Kim. Explore and Match: End-to-End Video Grounding With Transformer. *arXiv preprint arXiv:2201.10168*, 2022. 2, 3

[77] Haiping Wu, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Sequence Level Semantics Aggregation for Video Object Detection. In *ICCV*, pages 9217–9225, 2019. 1

[78] Peng Xu, Timothy M Hospedales, Qiyue Yin, Yi-Zhe Song, Tao Xiang, and Liang Wang. Deep Learning for Free-Hand Sketch: A Survey. *TPAMI*, 2022. 1

[79] Peng Xu, Kun Liu, Tao Xiang, Timothy M Hospedales, Zhanyu Ma, Jun Guo, and Yi-Zhe Song. Fine-Grained Instance-Level Sketch-Based Video Retrieval. *TCSVT*, pages 1995–2007, 2020. 1

[80] Shuai Yang, Zhangyang Wang, Jiaying Liu, and Zongming Guo. Deep Plastic Surgery: Robust and Controllable Image Editing With Human-Drawn Sketches. In *ECCV*, pages 601–617. Springer, 2020. 1

[81] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. LAVT: Language-Aware Vision Transformer for Referring Image Segmentation. In *CVPR*, pages 18155–18165, 2022. 2

[82] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch Me That Shoe. In *CVPR*, pages 799–807, 2016. 1, 4

[83] Qian Yu, Jifei Song, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Fine-Grained Instance-Level Sketch-Based Image Retrieval. *IJCV*, 129(2):484–500, 2021. 1, 9

[84] Hao Zhang, Aixin Sun, Wei Jing, Guoshun Nan, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. Video Corpus Moment Retrieval With Contrastive Learning. In *SIGIR*, pages 685–695, 2021. 9

[85] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where Does It Exist: Spatio-Temporal Video Grounding for Multi-Form Sentences. In *CVPR*, pages 10668–10677, 2020. 2

[86] Qianyu Zhou, Xiangtai Li, Lu He, Yibo Yang, Guangliang Cheng, Yunhai Tong, Lizhuang Ma, and Dacheng Tao. Transvod: End-To-End Video Object Detection With Spatial-Temporal Transformers. *TPAMI*, 2022. 1