SVI-Paste: Synthetic Dynamic Instance Copy-Paste

Anonymous CVPR submission

Paper ID 67

Abstract

*Data augmentation methods such as Copy-Paste have been studied as effective ways to expand training datasets while incurring minimal costs. While such methods have been extensively implemented for image level tasks, we found no scalable implementation of Copy-Paste built specifically for video tasks. In this paper, we leverage the recent growth in video fidelity of generative models to explore effective ways of incorporating synthetically generated objects into existing video datasets to artificially expand object instance pools. We first procure synthetic video sequences featuring objects that morph dynamically with time. Our carefully devised pipeline automatically segments then copy-pastes these dynamic instances across the frames of any target background video sequence. We name our video data augmentation pipeline Synthetic Dynamic Instance Copy-Paste, and test it on the complex task of Video Instance Segmentation which combines detection, segmentation and tracking of object instances across a video sequence. Extensive experiments on the popular Youtube-VIS 2021 dataset using two separate popular networks as baselines achieve strong gains of **+2.9 AP (6.5%)** and **+2.1 AP (4.9%)**. We make our code and models publicly available.*

1. Introduction

Analysing video data is one of the central tasks in the field of computer vision. With the proliferation of video data today, a fundamental challenge revolves around training networks [7, 15] that generalize and scale well in the face of large data diversity. It can be difficult to capture the immense variety and nuances of scenes in the real world through recorded image sequences. To tackle this, we have been relying on increasingly larger datasets [6, 62] to fulfill the needs of larger and deeper networks [31, 38]. However, each captured image usually requires human annotation, an endeavour that has become the central bottleneck in this pipeline as the number of recorded sequences grow. Video instance segmentation [47, 50, 52, 54, 58] (VIS) has emerged as a comprehensive video analysis task that encompasses recognition, segmentation and tracking of object instances across a video scene. To train a network for this task requires densely labelled image sequences where each object of interest is identified, labelled and its shape traced with a segmentation mask. The cost and time needed for segmentation labelling is often an order of magnitude higher than obtaining labels for other vision tasks such as classification where dense masks are not required. Furthermore, expanding any dataset manually requires finding suitable videos that match the complexity and scene structure of that dataset which can prove to be a difficult affair, especially for object categories that are intrinsically rare.

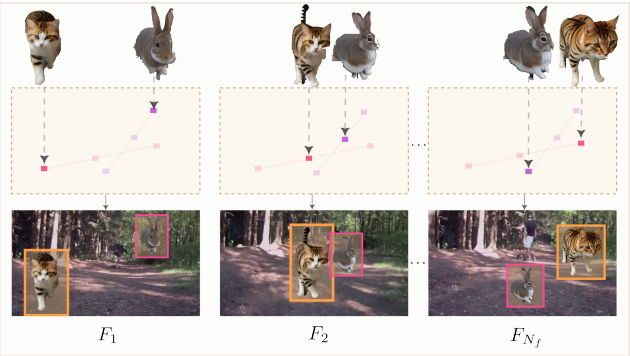

Figure 1. Our proposed data-augmentation framework generates synthetic object instances that are temporally dynamic and copy-pastes them using a linear trajectory onto each frame of a video sequence ($F_1, F_2, ..., F_{N_f}$). Our aim is to increase instance population of any existing video dataset.

Data augmentations [41] have been extensively studied as simple ways of artificially expanding a dataset. Copy-paste [8, 9, 12, 54] provides an object-aware augmentation pipeline where object instances are extracted from labelled datasets using segmentation masks and pasted onto existing background images or videos. These instances are drawn from the source dataset itself [9, 54] or from 3D models [34], neither of which provide us with a framework that is easily scalable since obtaining segmentation masks and 3D object models are both quite resource-intensive. Advancements in generative models have transformed computer vision in recent years. Methods [10, 60] have started to em-

ploy synthetic images generated from text-to-image generative models [39, 40] to improve the performance of the copy-paste augmentation. Off-the-shelf object segmenters [32, 37] are used to obtain new object instances from synthetic data, thereby eliminating the need for human labeling. Although synthetic data has shown to improve instance segmentation in the image domain, the application of synthetic data for copy-paste in the video domain has not yet been explored. To this end, we propose a novel data augmentation method for VIS, purpose-built for dense video tasks, and enable a natural expansion of existing datasets by introducing object instances that simulate the dynamism of objects in the real world.

The recent surge in the popularity and development of diffusion models for text-to-video (T2V) generation has resulted in the creation of networks [14, 16, 19] that can generate complex video scenes, based on text prompts, with remarkable realism. Inspired by these advancements in video fidelity spearheaded by diffusion-based generative models, we explore new ways of generating, segmenting, and then incorporating synthetically generated dynamic instances of specific object categories into existing video scenes. We do this to artificially inflate the object population, aiming to synthesise a greater variety of object features in complex video scenes. We devise an infinitely scalable data augmentation framework that requires no manual dense labeling for dynamic instance generation or segmentation.

Our pipeline, using a text-to-video generative model, produces synthetic video scenes that capture rich object semantics through diverse viewpoints and action-states. We extract segmentation masks for objects in each scene using an off-the-shelf self-supervised salient object segmenter. To ensure the validity of the generation and segmentation process, we use a zero-shot image recognition model trained on very large text-image multi-model datasets such as CLIP [38] to filter out any erroneously generated or segmented objects. Finally, we use images from YTVIS21 as background and randomly initialise the starting positions of generated objects. We explore different ways of copy-pasting dynamic objects in a video sequence and show that a linear trajectory with randomly sampled displacement gives best results. We illustrate this in Figure 1.

We highlight our main contributions as follows:

- We propose **S**ynthetic **D**ynamic **I**nstance Copy-**Paste** (SDI-Paste) as a novel synthetic data augmentation regime for the task of Video Instance Segmentation. Our method does not require any manual dense label annotation and is infinitely scaleable.
- We present a pipeline for on-demand crafting and segmenting of temporally dynamic objects in diverse scenes using only the category information of required objects. We then explore multiple ways of copy-pasting these instances onto existing video sequences and discover that a linear monotonic object trajectory with random jumps gives best results.
- Extensive experiments on the popular YTVIS21 dataset show the impressive performance of SDI-Paste. We equip two different online VIS networks with our pipeline and show a strong 6.5 % (2.9 AP) improvement. We also conduct multiple ablation studies for a thorough evaluation of the different parts of our framework. We release our synthetic dataset and code-base for future video data augmentation research.

2. Related Works

Video Instance Segmentation (VIS) [52] is a dense vision task that requires joint classification, segmentation and tracking of instances across video sequences. We mainly divide the VIS methods into two groups: offline and online. Offline (or per-clip) methods [2, 22, 29, 47] process the entire video clip simultaneously. This concurrent processing of numerous frames allows for deeper contextual understanding between them. However, these methods require significantly high processing power and memory during both training and inference. Prominent offline VIS methods either employ mask propagation [2, 29] or use transformer frameworks [22, 47]. Online (or per-frame) methods [17, 21, 50, 52–54] carry out instance segmentation using only a small number of frames within a local range, aiming to facilitate near real-time processing. At each step, instances are detected and assembled into short sequences from the available frames. For example, CTVIS [54] are built on top of image-level instance segmentation models [4, 5] by adding an additional pipeline that performs tracking. Memory mechanisms are utilized in IDOL [50] and CTVIS [54] to store instance identities as more frames become available, which improve tracking performance through consistent re-identification of object instances. While online VIS methods lag behind offline methods in terms of perfomance, they offer efficient training/inference cycle with lower memory usage. For this reason, we test SDI-Paste on online VIS. Specifically, we build and test our pipeline on CTVIS [54] and IDOL [50] as they are among the state-of-the-art online VIS networks.

Video Generation. Diffusion models [43] have recently seen rapid development in generation of high-resolution complex image scenes and are the most popular framework for Text-to-Image (T2I) synthesis [33, 35]. Stable Diffusion [40] performs sampling in latent feature space using an autoencoding framework to enable T2I generation. More recently, T2I methods have been explored for video generation conditioned on text prompts based on a diffusion framework [19]. Methods since have improved upon this

work using a pipeline where a pre-trained T2I model is used as backbone and motion/temporal modules are added and then trained using video data [3, 16, 18, 20]. Animatediff [14] uses a Stable Diffusion backbone and learns generalised motion priors to animate image scene. While it outputs video clips only 16 frames long, this is sufficient for online VIS methods as they usually take only a few input images. The generated animations can be chosen from a variety of image domain stylisations and is ready to be used out-of-the-box. All pre-trained models are made available and the code-base is well supported. For these reasons, we choose Animatediff to develop a novel data augmentation framework for VIS.

Image-based Data Augmentations are a low-cost, effective way to enlarge training datasets [25, 26, 42, 45]. Copy-paste [8, 9, 12] provides an object-aware data augmentation framework where objects are extracted from labelled datasets using segmentation masks and pasted onto existing background images. Ghiasi *et al.* [12] find that random pasting of instances on a background, without blending or context, is sufficient to achieve good improvements. For augmentation pipelines such as Copy-Paste, obtaining novel instances of the right category or from varying object viewpoints/representations can be challenging due to the large volume of instances needed. There have been works that use 3D rendering to insert objects into image scenes [23, 34, 44]. However, they require 3D model repositories that require human input. Recently, X-Paste [60] introduces object instances generated using Stable Diffusion [40] into the Copy-Paste framework for Instance Segmentation. Text-based image generative models [40] can produce an unlimited number of images which lends X-Paste a level of scalability that is difficult to match using existing datasets. Furthermore, the ability to generate any object provides an avenue of training networks to handle rare object classes.

Data augmentation strategies for Video tasks usually involve extensions of image-based methods [36]. While these improve performance, they fall short on leveraging the full scope of temporal dynamism that is unique to videos. DynaAugment [24] and Group RandAugment [1] extend common image-based augmentation methods for video classification. Other approaches involve mixing multiple video scenes akin to the copy-paste ethos. VideoMix [56] extends CutMix [55] by cutting and pasting frames from two different scenes. SV-Mix [46] and Learn2Augment [13] combine CutMix and Mixup [57] in a learnable framework. Zhao *et al.* [61] experiment with spatial and feature based augmentations for Video Object Tracking while Lee *et al.* adapt Copy-Paste for Video Inpainting [27]. However, we found no works exploring data augmentation for VIS.

These methods augment data by transforming or combining existing labelled videos. For augmentation pipelines such as Copy-Paste, obtaining novel instances of the right category or from varying object viewpoints/representations can be challenging as many instances are needed. Some works use 3D rendering to insert objects into image scenes [23, 34, 44]. However, they employ 3D model repositories that require human input. This is costly for tasks like VIS that require many object types and unique instances.

In [59], Zhang *et al.* use a Generative Adversarial Network to produce a "dynamic" image by compressing the temporal information of foreground objects from videos into one single static image. The process of sampling to compress video information into a single static image is lossy - we posit valuable information is likely to be lost. Furthermore, this process is unsuitable for VIS as it requires dense segmentation masks for each frame of a video sequence whereas their "dynamic" sampling always treats an entire video sequence as a single still image. More recently, generative models have been utilised to simulate video road scenes for the task of autonomous driving [11, 28, 49]. While these approaches also use diffusion-based generative models for synthetic scene generation, their focus is on video scenes limited to street-view elements (such as cars, trees, building, etc) and built specifically for autonomous driving needs.

In this work, we propose a pipeline that leverages text-to-video models within a Copy-Paste framework that generates and integrates object instances of diverse categories. This pipeline is designed to be easily adapted for any video task involving segmentation or tracking of diverse object instances. We aim to inject synthetically-generated, temporally-dynamic instances to expand the pool of object instances in existing video datasets. We investigate this pipeline's effectiveness and compare with new baselines as, to the best of our knowledge, ours is the first work to explore data augmentation strategies specifically for VIS.

3. Methodology

In this work, we aim to investigate an effective data augmentation scheme for dense video tasks by incorporating synthetically generated data. A straightforward method to achieve this goal is to extract object instances from static images and paste the same instance repeatedly into each frame of a video sequence. However, static instances fail to represent the inherent dynamism of real world objects that is captured in video. Without considering the dynamic nature of video data, achieving satisfactory performance is challenging.

This work proposes a natural extension of the copy-paste framework, specifically designed for video, demonstrating that dynamic generative object features provide sufficient

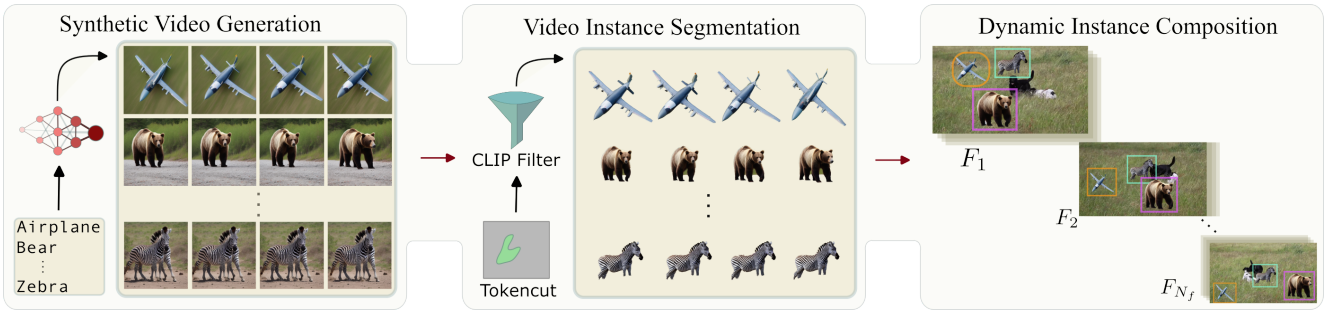

Figure 2. Illustration of our SDI-Paste pipeline. Firstly, Synthetic Video Generation uses text prompts to obtain diverse video scenes. Secondly, frames in each scene are segmented to acquire synthetic dynamic object instances. Finally, using a linear-random trajectory scheme, these dynamic instances are copy-pasted onto existing video sequences to compose the augmented dataset.

realism and supervision to improve performance. We employ a text-to-video generative diffusion model to create photo-realistic video frames. These frames are segmented and filtered using a zero-shot classification model to obtain object instance sequences. Finally, we insert these object instances as a dynamic sequence in moving locations throughout the successive frames of a video clip.

We name our framework **S**ynthetic **D**ynamic **I**nstance Copy-**Paste** (SDI-Paste) and show comprehensive testing for the challenging task of VIS. Our SDI-Paste pipeline includes three main steps: Synthetic Video Generation, Video Instance Segmentation, and Dynamic Instance Composition. Details follow below.

3.1. Synthetic Video Generation

To generate synthetic videos, we employ a text-to-video model, such as AnimateDiff [14], which incorporates motion dynamics into generated static image scenes. AnimateDiff animates images obtained through Stable Diffusion by using a separate motion modeling module trained to learn motion priors from large-scale video datasets. In this work, we aim to investigate augmentation using images featuring objects that change smoothly and dynamically over time, unlike Stable Diffusion, which exclusively produces static images [60].

We generate animated sequences for specific classes by providing the text-to-video model with a sentence prompt that describes the object in a dynamic scene. Through empirical experiments, we discovered that a simple yet effective way to produce a diverse set of objects and action-states was to include generic adjectives such as "moving" and "dynamic" to describe the object and to place it in a "changing" background. For example, to generate multiple and varied scenes for the object class "bear", the text prompt will be:

A close up video of one moving dynamic bear in changing background, moving camera, centred.

where **bear** is a variable depending on the object class. We use this sentence as the input to AnimateDiff. We find that this same sentence, when run repeatedly, results in a new video scene with visually unique objects and diverse action-states each time. We use the object categories from YouTube-VIS [52] to generate the necessary quantity of text inputs. When each of these text inputs is passed to AnimateDiff, it results in a short video clip comprising 16 frames.

We show some examples of synthetic frames in Figure 3 where dynamic objects naturally morph over time. We found that the generative model sometimes introduce small feature aberrations, such as an extra ear on the rabbit (row 1) or additional feet on the fox (row 3). However, despite these deformed features that become visible upon close inspection, these objects are still immediately visually recognisable. We demonstrate that incorporating these aberrations into the augmentation process leads to a remarkable improvement in performance. We hypothesize that these aberrations provide an extra challenge to the network as it learns to not only identify the correct classifications but also learns to track as objects morph and deform due to changing features, viewpoints and actions. Furthermore, in some instances, these aberrant features can be seen to simulate the sudden appearance/disappearance of object features (such as a limb) that might manifest in a real scene.

3.2. Video Instance Segmentation

The next step is video instance segmentation where we acquire segmentation masks for objects in all generated video frames. Since each generated frame has a single salient object against a generic background, we can use any off-the-shelf salient object segmentor to extract foreground instance masks. In our work, we use TokenCut [48], a graph-based algorithm that leverages features obtained from a self-supervised transformer to detect and segment salient objects in images and videos.

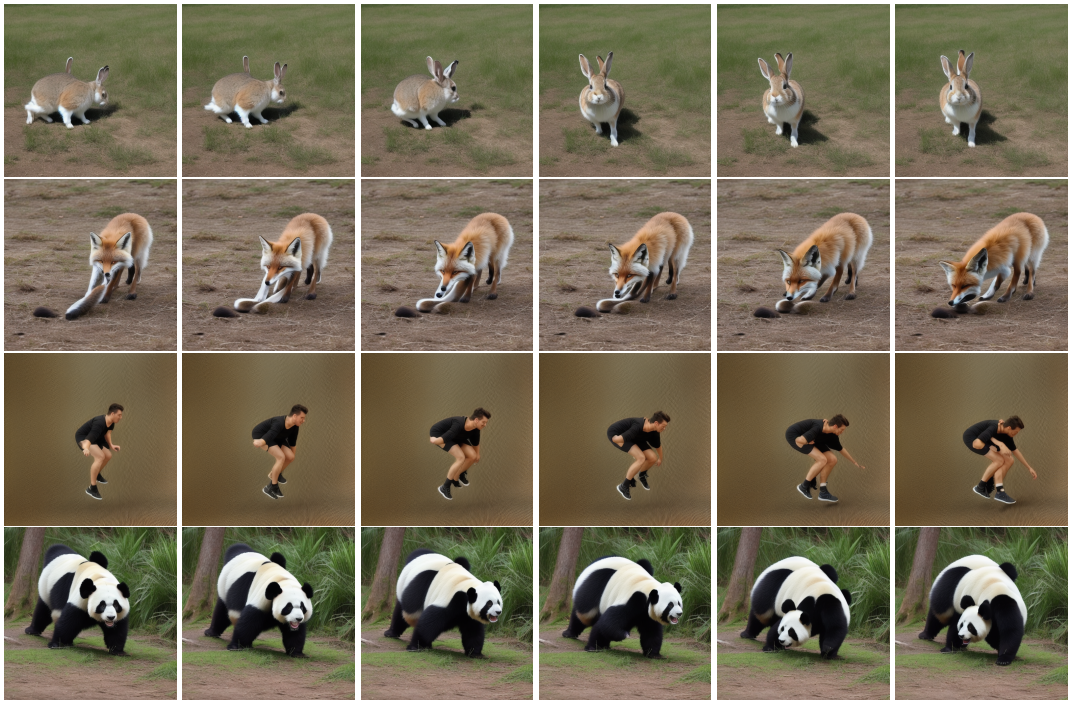

Figure 3. Examples of dynamic video frames generated with AnimateDiff [14]. We can observe single salient foreground objects undergoing seamless shape and viewpoint transitions as a result of their actions. Object features are mostly preserved barring some aberrations such as extra ears or feet.

To ensure the segmentation masks successfully extract the foreground object, we filter them using CLIP [38, 60]. In our setting, CLIP is employed as an assessor that matches the input text prompt and the resulting image content to obtain a relevance score for each frame. This score is used to judge the semantic relevance of the generated image and filter out instances with failed generations or erroneous segmentations. Additionally, masks occupying less than 5% or more than 95% of the total image area are removed.

3.3. Dynamic Instance Composition

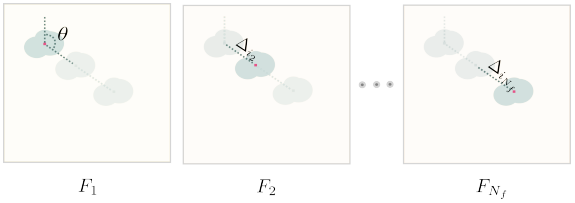

Figure 4. Figure showing linear instance placement trajectory. The direction θ is constant for all frames but the displacement Δ varies frame by frame.

We employ a class-balanced strategy [60] to sample instances from the segmented masks of the generated objects, and then randomly paste them onto a sequence of background frames. We use videos directly from the target dataset as the background and paste the generated instances on top of existing objects in each frame. Objects that are fully occluded after composition are removed.

We introduce a dynamic instance copy-paste strategy for video tasks that can be employed to a variable number of frames within a background sequence. We assume that there are N_f consecutive frames in a background video sequence: $\{F_1, F_2, ..., F_{N_f}\}$. First, we randomly select the number of novel objects N_i we expect to introduce into this background sequence: $N_i \in [1, N_{max}], N_{max} = 20$. Then we sample N_i animated instance categories. Each sampled category consists of N_f instances $(i_1, i_2, ..., i_{N_f})$ with one instance for each frame of the background video sequence.

Starting with the first frame in the background sequence, we randomly sample the starting xy-coordinates (x_0, y_0) for the first instance of each sampled category:

$$x_0 \sim \mathcal{U}[0, W], y_0 \sim \mathcal{U}[0, H] \quad (1)$$

where W and H give the width and height of the background image respectively. For each subsequent frame, a trajectory system is imposed for the positioning of the remaining instances. Empirically, we achieved the best results following a linear trajectory for pasting instances across background frames. For each object placement, we fix a constant direction throughout the background sequence but

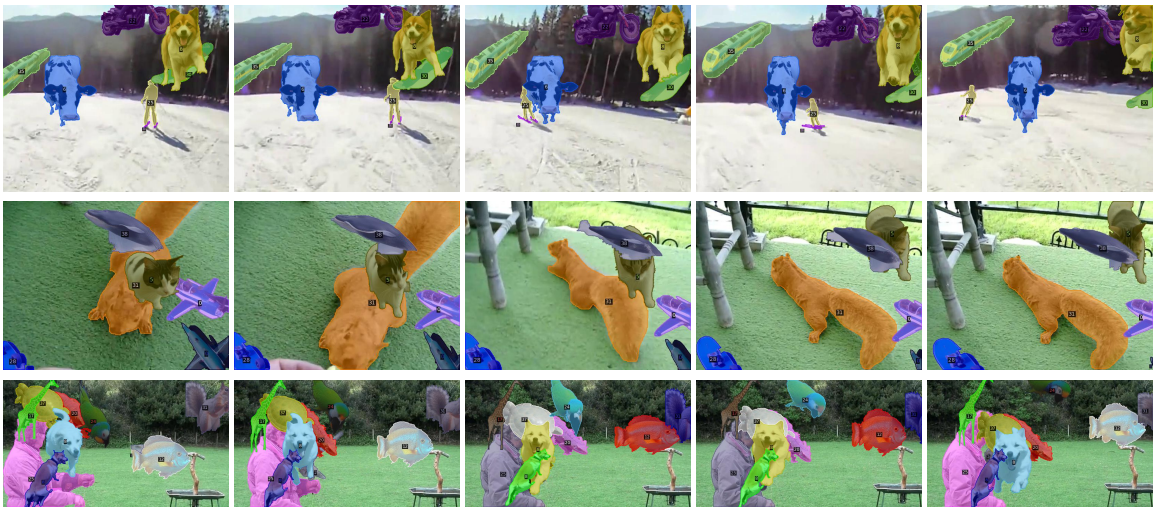

Figure 5. Example of dynamic video frames obtained after instance composition. Dynamic Instances are copy-pasted onto a background image with its existing objects to enlarge the instance pool for each sequence.

allow for variable displacement between frames. Specifically, for each object instance $i_j, j \in [1, N_f]$, we randomly sample an angle over a uniform distribution:

$$\theta_{i_j} \sim [0, 360]^\circ \tag{2}$$

We follow the direction θ_{i_j} when pasting instances in all the following background frames. However, we allow the displacement of instance positions between frames to change randomly. We show this process in Figure 4. For each instance i_j, we randomly sample a displacement:

$$\Delta_{i_j} \sim \mathcal{U}[0, \Delta_{max}] \tag{3}$$

where Δ_{max} is the maximum allowable displacement and is a hyper-parameter we set during training. Based on θ_{i_j} and Δ_{i_j}, we obtain the pixel displacement along x- and y-axis using standard trigonometry:

$$(\delta x_{i_j}, \delta y_{i_j}) = (\Delta_{i_j} * \cos\theta_{i_j}, \Delta_{i_j} * \sin\theta_{i_j}) \tag{4}$$

Then, we apply this to the initial instance position to obtain the xy-coordinates for each subsequent frame:

$$(x_{i_j}, y_{i_j}) = (x_{i_{j-1}} + \delta x_{i_j}, y_{i_{j-1}} + \delta y_{i_j}) \tag{5}$$

We follow X-Paste [60] in determining object scale as we paste instances onto the background frames. For each instance, we sample a scale S_i from a Gaussian distribution $N(\mu_C, \sigma_C^2)$ and paste it with scale $S_i^2 HW$ on a background frame where H, W denote the image height and width. For each category in the dataset, we calculate the mean μ_C and standard variance σ_C^2 of object scales ($\sqrt{O_M/(HW)}$) within that category. Here, O_M is the object mask area and HW gives the total image area. In Figure 5, we show some sample video frames after instance composition.

4. Experiments

4.1. Implementation

Datasets. We train, test and evaluate our method on YTVIS21 [52], a popular VIS dataset. YTVIS21 is a smaller subset of the YouTube-VOS (Video Object Segmentation) dataset [51] from which 40 common object categories were retained. It consists of 2,900 videos each 3 to 6 seconds long with 4,883 unique objects. During training, we use YTVIS21 as the background images on which we paste generated objects.

Baseline Frameworks. We design SDI-Paste as a plug-and-play dataset module that can be added to any online VIS training regime and reap immediate performance gains. To demonstrate, we test SDI-Paste on two popular online VIS frameworks: CTVIS [54] and IDOL [50]. For CTVIS, we test on a ResNet-50 [15] backbone which is pre-trained on COCO [30]. The baseline is trained for 32000 iterations. To incorporate the SDI-Paste pipeline, we divide the training regimen into two parts: first we pre-train with SDI-Paste enabled for 16000 iterations. We use this as pre-trained, disable SDI-Paste, and finetune for 16000 iterations on the base dataset only. We follow standard training settings as listed on [54]. Similarly, for IDOL, we test on pre-trained ResNet-50 [15] backbone and follow the same training regimen to obtain the baseline and a version of the model trained with SDI-Paste. While CTVIS takes in 10 sequences as input at each training step, IDOL requires only 2. SDI-Paste can support up to 16 frames of input as we are limited by the video throughput of our generative network.

Method	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
CTVIS (Baseline)	44.3	65.3	48.8	40.5	55.2
CTVIS (SDI-Paste)	**47.2** (+6.5%)	**68.2**	**51.8**	**41.3**	**56.3**
IDOL (Baseline)	42.7	64.3	46.3	39.6	53.9
IDOL (SDI-Paste)	**44.8** (+4.9%)	**67.7**	**49.5**	**39.8**	**54.2**

Table 1. Results comparing the performance of two online VIS networks with and without SDI-Paste.

While both of these methods offer models with larger backbones that compete with state-of-the-art in performance, they come with large compute resource overheads. Thus, we choose to test our pipeline on lighter versions of the models to demonstrate the efficacy of SDI-Paste as a pipeline that improves on any baseline regardless of the base network. We report standard metrics: AP, AP_{50}, AP_{75}, AR_1 and AR_{10}.

SDI-Paste Settings. SDI-Paste is a modular framework comprising of a text-to-video generator (AnimateDiff) and an image foreground segmentor (TokenCut). For AnimateDiff [14], we use the 40 object categories in YTVIS21 to produce animated video instances and obtain 470 video sequences for each category. Each sequence is 16 frames long resulting in 300800 generated image frames. We choose RealisticVision image stylisation for AnimateDiff as it produces the most natural looking videos. The remaining settings are set as recommended by the authors.

Likewise, for TokenCut [48], we use the recommended setting to segment each frame in a video sequence individually. Each segmented object is input into a CLIP model [38] and filtered. We determined qualitatively that a clip score threshold of 0.21 filters poorly segmented/generated objects. We also remove objects that occupy less than 5% and more than 95% of the total image area since such objects are likely to have been mis-segmented. During dynamic instance composition, we set maximum number N_{max} of novel objects introduced to each video sequence to be 20.

4.2. Results

Main Results. In Table 1, we compare our SDI-Paste trained models against baseline for CTVIS and IDOL and see solid improvements of 6.5% and 4.9% respectively. We achieve this not by changing network structure or altering training strategies but purely through injection of synthetic dynamic instances onto the base training dataset. The discrepancy in improvement between CTVIS and IDOL can be attributed to the difference in number of images input to the network during each training step: CTVIS uses 10 sequential images whereas IDOL uses only two. This results in the network seeing more synthetic instances in CTVIS relative to IDOL. We cannot match the performance figures of our two baselines in comparison to their original papers [54] and [50] as we trained them from scratch in constrained training and dataset settings; our main aim was to demonstrate the efficacy of SDI-Paste regardless of the base network or its performance capability.

Comparison with other methods. In Table 2, we make comparisons against other related data augmentation methods. We use CTVIS as our base VIS framework and compare SDI-Paste with a Copy-Paste [12] baseline where we copy-paste instances from across YTVIS21 (i.e. same instance copy-pasted across all the frames in a sequence). To test the effectiveness of dynamic object instances compared to static ones, we also compare against X-Paste [60] on YTVIS21. We use the code-base and recommended settings from [60] to generate and segment object instances and obtain comparable number of images to our SDI-Paste setting. We adapt X-Paste for a video task by pasting each static object instance repeatedly onto each frames in a video sequence. We observe that Copy-Paste improves the baseline CTVIS by 2.7%. X-Paste, with its synthetic static instances, posts an improvement of 5.6% over baseline whereas our SDI-Paste pipeline outperforms them both with an improvement of 6.5% over baseline (0.9% over X-Paste and 3.7% improvement over Copy-Paste). While these results show the effectiveness of the copy-paste framework in enabling a solid boost in model performance, we see that the VIS task is better served by SDI-Paste where dynamic object instances are injected onto the base dataset. We posit this is due to the dynamic instances capturing more diverse object features from varying viewpoints and shape deformations when compared to static instances as in X-Paste. Please note that we show the best results from a pool of maximum number of experiments that was possible within our computation budget.

4.3. Ablation Study

Ablating trajectory system. During Dynamic Instance Composition, we investigate three different methods of copy-pasting instances onto a sequence of images: Linear, Bezier and Linear-random. For the linear system, we paste instances across the video frames with a straight-line trajectory as directed by the angle θ_{i_j} with constant displacement of objects between the frames. The linear-random system adopts the same straight-line trajectory but allows for randomly sampled displacement of objects between frames (as discussed in 3.3). For the Bezier system, we trace the path

Method	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
Baseline	44.3	65.3	48.8	40.5	55.2
CopyPaste [12]	45.5	65.9	50.3	40.4	54.6
X-Paste [60]	46.8	68.1	50.5	40.5	55.9
SDI-Paste (ours)	**47.2**	**68.2**	**51.8**	**41.3**	**56.3**

Table 2. Results comparing existing data augmentation methods with SDI-Paste using CTVIS with ResNet-50 as baseline.

of the object using a Bezier curve with a random length. Comparing these three systems, we find that the linear-random trajectory consistently gives the best result while the linear system is only slightly behind. The Bezier trajectory gives the worst performance. This might be due to SDI-Paste being introduced as a pre-training step where simple linear trajectories are easier for the network to track. Possibly, an only-Bezier-curve trajectory does not account for the majority of object movement in YTVIS21. A further experiment could be a system that combines diverse ways of moving objects improves results. We leave this for future works.

Trajectory Method	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
Bezier curve	45.2	67.8	50.5	40.5	54.2
Linear	46.4	**70.0**	50.1	39.1	53.0
Linear-random	**47.2**	68.2	**51.8**	**41.3**	**56.3**

Table 3. Results comparing different trajectory systems adopted during Dynamic Instance Composition.

Ablating segmentation method. In X-Paste [60], Zhao *et al.* employ a CLIP-guided selection of segmentation masks obtained from four different supervised salient object segmentation networks. We test this strategy against TokenCut [48] as the only segmenter and find that it consistently outperforms the X-Paste pipeline.

Segmentation Pipeline	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
X-Paste CLIP-guided [60]	46.2	**69.9**	50.1	39.9	54.5
TokenCut [48]	**47.2**	68.2	**51.8**	**41.3**	**56.3**

Table 4. Comparison of two segmentation methods: X-Paste CLIP-guided strategy [60] and TokenCut [48].

Ablating effect of more instances. To verify the effect of increasing the number of synthetic instances generated on model performance, we create two datasets where, for each category, we generate either 150 or 470 dynamic instance sequences (each sequence consists of 16 image frames). These amount to 96000 and 300800 image frames generated respectively. We see in Table 5 that increasing the

Number of sequences	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
150	45.6	67.4	50.5	40.8	55.9
470	**47.2**	**68.2**	**51.8**	**41.3**	**56.3**

Table 5. Results comparing the effect of increasing the number of generated dynamic instances on CTVIS. Each sequence consists of 16 frames.

number of dynamic instances improves model performance by 3.5%. As expected, with a larger number of unique instances available for training, the network shows remarkable improvement on the same training regimen. Given our compute resources, using AnimateDiff to produce object instances was a significantly heavy task which limited our ability to test with larger generated datasets. We expect the model performance to benefit further from an even larger instance pool to draw from. We make the generated dataset, as well as code to reproduce it at any size, available and leave this task for future research.

5. Conclusion

In this paper, we introduce SDI-Paste as a novel synthetic data augmentation pipeline for VIS. SDI-Paste combines a generative text-to-video model and a self-supervised object segmentor in a carefully designed pipeline that yields dynamic object instances. We copy and paste these instances across a base dataset to achieve strong improvement over baseline and other existing augmentation strategies. The individual modules of our pipeline can be swapped for better and newer modules as T2V generators and object segmentors improve over time. The essence of our framework is infinitely scaleable and adaptable which we hope will make SDI-Paste a beneficial data augmentation regimen for other dense video tasks.

References

[1] Fengmin An, Bingbing Zhang, Zhenwei Wang, Wei Dong, and Jianxin Zhang. Group randaugment: Video augmentation for action recognition. In *2022 5th International Conference on Data Science and Information Technology (DSIT)*, pages 1–5. IEEE, 2022. 3

[2] Gedas Bertasius and Lorenzo Torresani. Classifying, segmenting, and tracking object instances in video with mask propagation, 2021. 2

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 3

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020. 2

[5] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation, 2022. 2

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[7] 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. 1

[8] Nikita Dvornik, Julien Mairal, and Cordelia Schmid. Modeling visual context is key to augmenting object detection datasets. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 364–380, 2018. 1, 3

[9] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1301–1310, 2017. 1, 3

[10] Chengxiang Fan, Muzhi Zhu, Hao Chen, Yang Liu, Weijia Wu, Huaqi Zhang, and Chunhua Shen. Divergen: Improving instance segmentation by learning wider data distribution with more diverse generative data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3986–3995, 2024. 1

[11] Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. *arXiv preprint arXiv:2310.02601*, 2023. 3

[12] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2918–2928, 2021. 1, 3, 7, 8

[13] Shreyank N Gowda, Marcus Rohrbach, Frank Keller, and Laura Sevilla-Lara. Learn2augment: Learning to compose videos for data augmentation in action recognition. In *European conference on computer vision*, pages 242–259. Springer, 2022. 3

[14] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2, 3, 4, 5, 7

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 6

[16] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022. 2, 3

[17] Miran Heo, Sukjun Hwang, Jeongseok Hyun, Hanjung Kim, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. A generalized framework for video instance segmentation, 2023. 2

[18] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 3

[19] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 2

[20] Yaosi Hu, Chong Luo, and Zhenzhong Chen. Make it move: controllable image-to-video generation with text descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18219–18228, 2022. 3

[21] De-An Huang, Zhiding Yu, and Anima Anandkumar. Minvis: A minimal video instance segmentation framework without video-based training, 2022. 2

[22] Sukjun Hwang, Miran Heo, Seoung Wug Oh, and Seon Joo Kim. Video instance segmentation using inter-frame communication transformers, 2021. 2

[23] Kevin Karsch, Varsha Hedau, David Forsyth, and Derek Hoiem. Rendering synthetic objects into legacy photographs. *ACM Transactions on graphics (TOG)*, 30(6):1–12, 2011. 3

[24] Taeoh Kim, Jinhyung Kim, Minho Shim, Sangdoo Yun, Myunggu Kang, Dongyoon Wee, and Sangyoun Lee. Exploring temporally dynamic data augmentation for video recognition. *arXiv preprint arXiv:2206.15015*, 2022. 3

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 3

[26] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 3

[27] Sungho Lee, Seoung Wug Oh, DaeYeun Won, and Seon Joo Kim. Copy-and-paste networks for deep video inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4413–4421, 2019. 3

[28] Xuan Li, Yutong Wang, Lan Yan, Kunfeng Wang, Fang Deng, and Fei-Yue Wang. Paralleleye-cs: A new dataset of synthetic images for testing the visual intelligence of intelligent vehicles. *IEEE Transactions on Vehicular Technology*, 68(10):9619–9631, 2019. 3

[29] Huaijia Lin, Ruizheng Wu, Shu Liu, Jiangbo Lu, and Jiaya Jia. Video instance segmentation with a propose-reduce paradigm, 2021. 2

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context. 2014. 6

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1

[32] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7086–7096, 2022. 2

[33] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 2

[34] Yair Movshovitz-Attias, Takeo Kanade, and Yaser Sheikh. How useful is photo-realistic rendering for visual learning? In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 202–217. Springer, 2016. 1, 3

[35] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2

[36] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6964–6974, 2021. 3

[37] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern recognition*, 106:107404, 2020. 2

[38] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 5, 7

[39] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3

[41] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019. 1

[42] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[43] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 2

[44] Hao Su, Charles R Qi, Yangyan Li, and Leonidas J Guibas. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. In *Proceedings of the IEEE international conference on computer vision*, pages 2686–2694, 2015. 3

[45] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 3

[46] Yi Tan, Zhaofan Qiu, Yanbin Hao, Ting Yao, Xiangnan He, and Tao Mei. Selective volume mixup for video action recognition. *arXiv preprint arXiv:2309.09534*, 2023. 3

[47] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8741–8750, 2021. 1, 2

[48] Yangtao Wang, Xi Shen, Yuan Yuan, Yuming Du, Maomao Li, Shell Xu Hu, James L Crowley, and Dominique Vaufreydaz. Tokencut: Segmenting objects in images and videos with self-supervised transformer and normalized cut. 2022. 4, 7, 8

[49] Yuxi Wei, Zi Wang, Yifan Lu, Chenxin Xu, Changxing Liu, Hao Zhao, Siheng Chen, and Yanfeng Wang. Editable scene simulation for autonomous driving via collaborative llm-agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15077–15087, 2024. 3

[50] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation, 2022. 1, 2, 6, 7

[51] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. 2018. 6

[52] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5188–5197, 2019. 1, 2, 4, 6

[53] Shusheng Yang, Yuxin Fang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Crossover learning for fast online video instance segmentation, 2021.

[54] Kaining Ying, Qing Zhong, Weian Mao, Zhenhua Wang, Hao Chen, Lin Yuanbo Wu, Yifan Liu, Chengxiang Fan, Yunzhi Zhuge, and Chunhua Shen. Ctvis: Consistent training for online video instance segmentation. 2023. 1, 2, 6, 7

[55] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 3

[56] Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, and Jinhyung Kim. Videomix: Rethinking data augmentation for video classification. *arXiv preprint arXiv:2012.03457*, 2020. 3

[57] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 3

[58] Tao Zhang, Xingye Tian, Yu Wu, Shunping Ji, Xuebo Wang, Yuan Zhang, and Pengfei Wan. Dvis: Decoupled video instance segmentation framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1282–1291, 2023. 1

[59] Yumeng Zhang, Gaoguo Jia, Li Chen, Mingrui Zhang, and Junhai Yong. Self-paced video data augmentation with dynamic images generated by generative adversarial networks. *arXiv preprint arXiv:1909.12929*, 2019. 3

[60] Hanqing Zhao, Dianmo Sheng, Jianmin Bao, Dongdong Chen, Dong Chen, Fang Wen, Lu Yuan, Ce Liu, Wenbo Zhou, Qi Chu, et al. X-paste: Revisit copy-paste at scale with clip and stablediffusion. *arXiv preprint arXiv:2212.03863*, 2022. 1, 3, 4, 5, 6, 7, 8

[61] Jie Zhao, Johan Edstedt, Michael Felsberg, Dong Wang, and Huchuan Lu. Leveraging the power of data augmentation for transformer-based tracking. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6469–6478, 2024. 3

[62] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 1