

APES: Articulated Part Extraction from Sprite Sheets

Zhan Xu^{1,2} Matthew Fisher² Yang Zhou² Deepali Aneja² Rushikesh Dudhat¹ Li Yi³ Evangelos Kalogerakis¹

¹University of Massachusetts Amherst ²Adobe Research ³Tsinghua University

Abstract

Rigged puppets are one of the most prevalent representations to create 2D character animations. Creating these puppets requires partitioning characters into independently moving parts. In this work, we present a method to automatically identify such articulated parts from a small set of character poses shown in a sprite sheet, which is an illustration of the character that artists often draw before puppet creation. Our method is trained to infer articulated parts, e.g. head, torso and limbs, that can be reassembled to best reconstruct the given poses. Our results demonstrate significantly better performance than alternatives qualitatively and quantitatively. Our project page <https://zhan-xu.github.io/parts/> includes our code and data.

1. Introduction

Creating rich, animated characters has traditionally been accomplished by independently drawing each frame of the character. To accelerate this process, tools have been developed to allow precisely rigged 2D characters to be easily rendered in different poses by manipulating the rig. To create these rigs, artists often start by drawing several different poses and configurations of the complete character in a sprite sheet or turnaround sheet. They then manually segment out the common parts in these sheets and stitch them together to create the final character rig, which can then be articulated to reconstruct the original character drawings [24]. The obtained parts from different sprite sheets can also be used as assets and assembled freely to create new character rigs¹.

Significant expertise is required to create a well-rigged 2D character, and automatic rigging methods have several unique challenges. Animated characters can have a wide range of different limbs, accessories, and viewing angles, which prevents a single template from working across all characters. Furthermore, the amount of available examples for rigged, animated characters is relatively small when compared against real datasets that can be acquired by motion capture or other techniques. This limited data is particularly challenging to work with because characters are

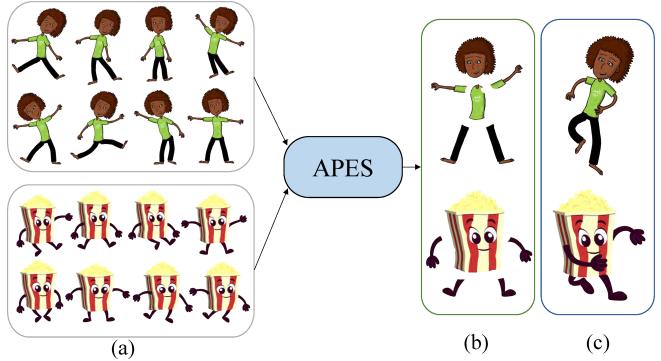


Figure 1. Given sprite sheets as input (a), APES produces articulated parts (b) that can best express poses in the sprite sheets. The obtained parts can further be warped to generate new poses (c), or manipulated freely to create new puppets.

often drawn and animated in different styles. Finally, poses shown in sprite sheets have both articulated variation and non-rigid deformation. Extracting articulated parts that express given poses requires effective analysis of the motion demonstrated in sprite sheets.

We propose a method to automatically construct a 2D character rig from a sprite sheet containing a few examples of the character in different poses. Our rig is represented as a set of deformable layers [52], each capturing an articulated part. We assume that all characters in the sprite sheet can be reconstructed by applying a different deformation to each puppet layer and then compositing the layers together. We start by learning a deep network that computes correspondences between all pairs of sprites. We then use these correspondences to compute possible segmentations of each sprite. Finally, we attempt to reconstruct the other sprites in the sprite sheet using the possible puppet segmentations, choosing the set with minimal overall reconstruction error.

We evaluate our method on several test sprite sheets. We show that our method can successfully produce articulated parts and significantly outperforms other representative appearance and motion-based co-part segmentation works [17, 39]. Our contributions are the following:

- A method for analyzing a sprite sheet and creating a corresponding articulated character that can be used as a puppet for character animation.
- A neural architecture to predict pixel motions and clus-

¹<https://pages.adobe.com/character/en/puppet-maker>

ter pixels into articulated moving parts without relying on a known character template.

- An optimization algorithm for selecting the character parts that can best reconstruct the given sprite poses.

2. Related Work

Rigid motion segmentation. Several approaches [40, 43–45, 48, 60] have been proposed to cluster pixels into groups following similar rigid motions. One line of work [43, 44] identifies rigid groups by discovering distinct motion patterns from 2D optical flow. These methods typically work well on smooth video sequences, but cannot generalize to images with large pose changes between each other. They also aim at object level segmentation, and often miss articulated parts within each object. Other works employ 3D geometric constraints and features to infer the underlying motion of pixels for clustering [40, 45, 48, 57, 60]. These methods also assume small motions, and require multi-view input to perform 3D geometric inference, thus are not applicable to artistic sprites.

Co-part segmentation. Several works focus on segmenting common foreground objects or parts from a set of images or video frames [8, 9, 17, 19, 26, 46, 47, 61]. They often utilize features from pretrained networks on ImageNet [36], thus are more suitable for natural images instead of non-photorealistic images, such as sprites. Most importantly, their segmentation relies more on appearance and semantic consistency rather than part motion. As a result, they miss articulated parts, even when trained on our datasets, as shown in our experiments in the case of SCOPS [17].

Other co-part segmentation works rely on motion to better extract articulated parts. Early methods use key-point tracking and various strategies for trajectory recovery and modeling [7, 10, 30, 49, 56]. However, they are often hand-tuned and prone to noisy tracking and large pose deformations. More recently, deep learning methods have shown promising results for motion-based co-segmentation [37, 39, 54]. However, they heavily rely on well-predicted optical flow. When input images have distinct and large pose changes, optical flow becomes unreliable. They are also more suitable for natural images of objects from a single category. When trained on sprite sheets with varying articulation structure, they produce unsatisfactory results, as shown in our experiments for the recent approach of [39].

3D mobility segmentation. Mobility-based segmentation for 3D point clouds has also been investigated in recent works [22, 23, 51, 53]. Yi *et al.* [58] predicted point cloud segmentation from a pair of instances under different object articulation states. Hayden et al. [13] proposed an unsupervised part model to infer parts in a 3D motion sequence. MultiBodySync [16] achieved consistent correspondence and segmentation from multiple articulation states of the

same object by spectral synchronization. All these approaches are designed for 3D point clouds or meshes. Although we are inspired by these approaches to handle large pose variations, our method incorporates several adaptations for processing 2D sprites, including a convolutional correspondence module for pixel correspondence, a neural voting strategy to handle efficient clustering of rigid motions in superpixel space, and an optimization strategy to find common parts leading to the best reconstruction of sprites.

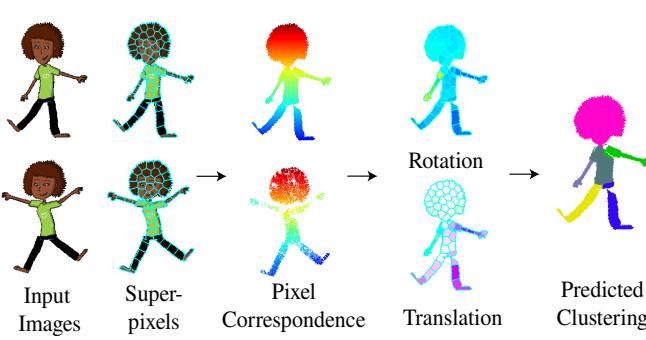
Puppet rigging and deformation. Prior works on puppet deformation [14, 32] assumes that the parts and their hierarchy are given i.e., the articulated parts have been specified by artists. Our approach is complementary to these methods, aiming to automate part extraction useful in their input. Recently, Xu *et al.* [55] proposed a neural network to infer a hierarchical rig for articulated characters. However, it relies only on the 3D geometry of the model, and does not take into account motion cues, as we do.

3. Method

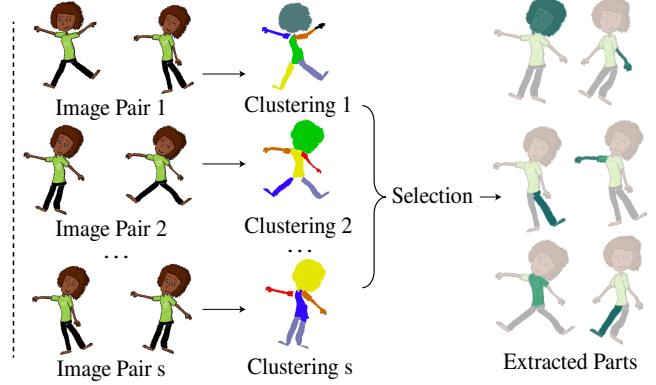
The goal of our method is to infer the articulated parts of a 2D cartoon character given only a few poses drawn by an artist under different articulations. The number of poses can vary for each character, e.g. 6 to 10 in our datasets. The input are P sprite RGB raster images \mathbf{I}_i and their accompanying foreground binary masks \mathbf{M}_i , $i = [1, \dots, P]$, where P is the total number of poses. The output is a set of articulated body parts that artists can subsequently animate based on standard part rigging methods and software [2, 6, 24] (see Fig. 1 for examples).

The pipeline of our method is shown in Fig. 2. First, given any pair of images (poses) from the inputs, the first module of our method, i.e *correspondence* module, (Sec. 3.1) infers the pixel correspondences which capture the candidate motions of pixels between images. These correspondences are then processed through the *clustering* module (Sec. 3.2), which attempts to find pixels with similar motion patterns and groups them into a set of candidate articulated parts as the output. This modular architecture has the advantage of *disentangling* motion from appearance and using only motion patterns for clustering. Finally, we gather candidate parts from all pairs and select a final set of parts to represent the target puppet (Sec. 3.3). The selected parts are required to have minimum overlap with each other and also reconstruct all poses with as-rigid-as-possible deformations.

The correspondence and clustering module have the form of neural networks that we both train jointly in a supervised manner (Sec. 4) based on publicly available puppet datasets (Sec. 5.1). We observe that networks can still generalize to real, artist-made cartoon characters and poses. The part selection solves a parameter-free optimiza-



(a) Correspondence and Clustering modules



(b) Whole pipeline

Figure 2. Pipeline overview. (a) Given any pair of images, the correspondence module (Sec.3.1) predicts candidate pixel correspondence between them. The clustering module (Sec.3.2) then groups superpixels with similar rigid transformation together. (b) After applying the correspondence and clustering modules to all pairs of images, we collected parts scattered across all poses. We select a subset through optimization that can best reconstruct the given poses while having minimal overlap between them.

tion problem which does not require training.

3.1. Correspondence module

Given a pair of images $\mathbf{I}_s, \mathbf{I}_t$ and corresponding binary foreground masks $\mathbf{M}_s, \mathbf{M}_t$ from the input set, the correspondence module predicts the candidate motion mapping of foreground pixels between two images. To achieve this, the module first concatenates each input image and mask, then transforms them into a feature map $\mathbf{F}_s \in \mathbb{R}^{H \times W \times 64}$ using a convnet. The network follows a U-Net architecture [35] and consists of ten convolutional layers in its encoder and another ten layers in its decoder. The convolutional layers in the encoder implement *gated convolution* [59], whose gating mechanism prevents background pixels indicated by the masks from influencing the foreground pixel correspondences. The feature vector of each pixel is normalized according to its L_2 norm such that it is unit length (i.e., it lies on the unit hypersphere [50]).

Next, given each foreground pixel $\mathbf{x} = (x, y)$ in the source image \mathbf{I}_s , its corresponding pixel $\mathbf{x}' = (x', y')$ in the target image \mathbf{I}_t is found as the pixel with the most similar feature vector in terms of cosine similarity:

$$\mathbf{x}' = \arg \max_{\mathbf{u} \in \mathbf{I}_t, \mathbf{M}_t(\mathbf{u})=1} (\mathbf{F}_s(\mathbf{x}) \cdot \mathbf{F}_t(\mathbf{u})) \quad (1)$$

We experimented with alternatives to extract correspondences such as RAFT [42] and COTR [18] trained on the same dataset as ours. Both resulted in worse results (see our experiments section for comparisons and discussion).

3.2. Clustering module

Given the pixel correspondences between the source and the target image $\mathbf{I}_s, \mathbf{I}_t$, our clustering module aims to discover character articulated parts by grouping pixels with similar motion transformations. Since it is not possible to estimate the transformation, i.e. 2D rotation and translation

from a single pixel, we instead gather votes for transformations from pairs of corresponding points $(\mathbf{x}_1, \mathbf{x}'_1)$ and $(\mathbf{x}_2, \mathbf{x}'_2)$, where $\mathbf{x}_1, \mathbf{x}_2$ are source pixels, and $\mathbf{x}'_1, \mathbf{x}'_2$ are their correspondences in the target image. Then we cluster these votes to discover the dominant rigid motion transformations and associated parts, similarly to Hough voting [4].

Voting pairs. Gathering votes from correspondences of all possible pixel pairs $\mathbf{x}_1, \mathbf{x}_2$ would be computationally expensive even for moderate image resolutions. In addition, distant pixels often belong to different parts, thus, their votes would tend to be irrelevant. To accelerate computations, we apply the superpixel segmentation method SLIC [1] to our input images and assume all pixels within a superpixel share the same motion transformation.

Rotation extraction. To extract the rotation from pairs of correspondences, one popular method is to use the orthogonal Procrustes analysis [15]. However, through our experiments, this approach turned out not to be robust – even slightly noisy correspondences can significantly distort the votes. Instead, we follow a convnet approach that learns to estimate the transformations from approximate correspondences. The input to our network is a map storing the voting pairs. Specifically, for each source pixel \mathbf{x}_1 , we store the 2D vector $\mathbf{x}_1 - \mathbf{x}_c$ representing its relative position with respect to its superpixel centroid \mathbf{x}_c , and also the corresponding 2D vector $\mathbf{x}'_1 - \mathbf{x}'_c$. This results in a $H \times W \times 4$ input voting map. Pixels without any correspondences are indicated by an additional binary mask.

The voting and mask maps are processed through a U-Net backbone and gated convolutions similar to the convnet of our correspondence module. The output is a $H \times W \times 64$ feature map representing motion features per pixel in the source image. We then apply average pooling spatially over each superpixel area to acquire motion features $\mathbb{R}^{K_s \times 64}$ for all K_s superpixels. Finally, an MLP layer is applied to map

the features to $\mathbb{R}^{K_s \times 2}$ space, representing (residual) sine and cosine of the rotation angles for K_s superpixels.

Translation extraction. Directly predicting both translation and rotation is possible, however we found it is more accurate to predict the rotation first, then update the motion features based on the rotation, and finally predict the translation (see also our ablation). In this manner, we discourage the network to express any small rotations merely as translations. The translation prediction network shares the same architecture of the rotation extraction network.

Clustering. Given extracted rotations and translations for superpixels, we proceed with characterizing their motion similarity, or in other words *affinity*. This affinity is computed based on motion residuals inspired by [16]. We apply the estimated rotation and translation of each superpixel to transform all other superpixels, and compute the position difference between the transformed superpixels and their corresponding superpixels. Specifically, given a super-pixel \mathbf{p}_i with extracted rotation matrix $\mathbf{R}_s[i]$ and translation $\mathbf{t}_s[i]$, the motion residual for the super-pixel \mathbf{p}_j is computed as:

$$\mathbf{D}_s(i, j) = \frac{\sum_{\mathbf{x} \in \mathbf{p}_j} (\mathbf{R}_s[i] \cdot \mathbf{x} + \mathbf{t}_s[i] - \mathbf{x}')}{|\mathbf{p}_j|} \quad (2)$$

where $|\mathbf{p}_j|$ is the number of pixels in the superpixel \mathbf{p}_j . The motion residual matrix $\mathbf{D}_s \in \mathbb{R}^{K_s \times K_s \times 2}$ is processed through more MLP layers to compute the superpixel affinity matrix $\mathbf{A}_s \in \mathbb{R}^{K_s \times K_s}$. More details on the architecture can be found in the supplementary.

Given the predicted affinity matrix \mathbf{A}_s , the grouping is achieved by using spectral clustering [29]. Here we follow the differential clustering approach [3, 16], which results in matrix $\mathbf{G}_s \in \mathbb{R}^{K_s \times C_s}$ representing a soft membership of superpixels to C_s clusters. We follow [16] to set the number of clusters based on the number of eigenvalues extracted from spectral clustering larger than a threshold. Here we set the threshold as 1% of the sum of the first 10 eigenvalues. By converting the soft membership to a hard one, the resulting clusters reveal articulated parts for the source pose based on its paired target pose.

3.3. Part Selection

By passing each pair of poses $\mathbf{I}_s, \mathbf{I}_t$ through our correspondence and segmentation modules, we obtain a set of parts for the source pose \mathbf{I}_s . Processing all pairs of poses yields a “soup” of candidate parts $\mathbf{Q} = \{\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_C\}$ scattered across all poses, where $C = \sum_s C_s$ is their total number. Obviously many of these parts are redundant e.g., the same arm extracted under different poses. Our part selection procedure selects a compact set of parts that (a) can reconstruct all poses with minimal error, and also (b) have minimum overlap with each other. To reconstruct poses, one possibility is to use rigid transformations of candidate

parts. Despite the fact that rigidity was used to approximately model the motion of parts in the previous section, not all the sprite sheet characters are fully rigidly deformed. There are often small non-rigid deformations within each part and around their boundaries. Thus we resort to as-rigid-as-possible (ARAP) deformation [41] for reconstructing poses using the selected parts more faithfully.

To satisfy the above criteria, we formulate a “set cover” optimization problem where the smallest sub-collection of “sets” (i.e., parts in our case) covers a universe $\mathbf{P} = \{\mathbf{p}_i\}$ of “elements” (i.e., all the superpixels across all poses). Specifically, by introducing a binary variable z_c indicating whether a part \mathbf{q}_c belongs to the optimal set (the “set cover”) or not, we formulate the following optimization problem:

$$\begin{aligned} & \min \sum_{\mathbf{q}_c \in \mathbf{Q}} z_c \\ \text{s.t. } & \sum_{\mathbf{q}_c : \mathbf{p}_i \in \mathbf{q}_c} z_c \geq 1 \text{ for all } \mathbf{p}_i \in \mathbf{P} \end{aligned} \quad (3)$$

We solve the above Integer Linear Programming (ILP) problem through relaxation. This yields a continuous linear programming problem solved using the interior point method [12]. We finally apply the randomized-rounding algorithm [34] to convert the continuous result to our desired binary predictions. The randomized-rounding can give us multiple possible solutions. We measure their quality by deforming the selected parts in each solution to best reconstruct all the given poses. The deformation is based on ARAP [41]. We choose the best solution with the minimal reconstruction error (see details in the supplementary).

4. Training

The correspondence and clustering modules are involved in our training procedure.

Correspondence module supervision. We train the correspondence module through a contrastive learning approach using supervision of pair-wise pixel correspondences. Specifically, given a pair of input images $\mathbf{I}_s, \mathbf{I}_t$, we minimized a correspondence loss [28, 31] that encourages the representation of ground-truth corresponding pixel pairs $(\mathbf{x}, \mathbf{x}')$ to be more similar than non-corresponding ones:

$$L_{\mathbf{x}, \mathbf{x}'}^{(corr)} = -\log \frac{\exp(\mathbf{F}_s(\mathbf{x}) \cdot \mathbf{F}_t(\mathbf{x}')/\tau)}{\sum_{\mathbf{u} \in \mathbf{U}_t} \exp(\mathbf{F}_s(\mathbf{x}) \cdot \mathbf{F}_t(\mathbf{u})/\tau)} \quad (4)$$

where \mathbf{U}_t is a predefined number of pixels we randomly sample from the foreground region of the image \mathbf{I}_t indicated by its mask \mathbf{M}_t . We set this number to 1024 in our experiments. The temperature τ is used to scale the cosine similarities. It is initially set as 0.07, and is learned simultaneously as we train the correspondence module [33]. The total correspondence loss L_c is averaged over all training corresponding pixel pairs.

During training, we alternatively replace the argmax of Eq. 1 with a soft version to preserve differentiability and enable backpropagation of losses from the clustering model. Specifically, we replace it with the weighted average of the top- κ closest target image foreground pixels to each source image pixel ($\kappa = 3$ in our implementation):

$$\mathbf{x}' = \frac{\sum_{\mathbf{u}_\kappa \in \mathcal{U}(\mathbf{x})} \exp(\mathbf{F}_s(\mathbf{x}) \cdot \mathbf{F}_t(\mathbf{u}_\kappa) / \tau) \cdot \mathbf{u}_\kappa}{\sum_{\mathbf{u}_\kappa \in \mathcal{U}(\mathbf{x})} \exp(\mathbf{F}_s(\mathbf{x}) \cdot \mathbf{F}_t(\mathbf{u}_\kappa) / \tau)} \quad (5)$$

where $\mathcal{U}(\mathbf{x})$ represent the top- κ most similar target pose pixels to \mathbf{x} using cosine similarity. The closest pixels are updated after each forward pass through our network.

Clustering module supervision. We train the clustering module with the binary cross-entropy (BCE) loss over the supervision of ground-truth affinity matrix $\mathbf{A}_s^{gt}(i, j)$.

$$L_s^{(aff)} = BCE(\mathbf{A}_s, \mathbf{A}_s^{gt}) \quad (6)$$

Similarly to [16], we introduce an additional loss on the motion residual matrix \mathbf{D}_s to encourage consistent rigid transformation predictions, i.e. $\mathbf{R}_s, \mathbf{t}_s$ in Eq. 2, across superpixels of the same part:

$$L_s^{(motion)} = \frac{\sum_{i,j} [\mathbf{A}_s^{gt}(i, j) = 1] \cdot \|\mathbf{D}_s(i, j)\|^2}{\sum_{i,j} [\mathbf{A}_s^{gt}(i, j) = 1]} \quad (7)$$

where $[.]$ is an indicator function.

Finally, we adopt the soft IoU loss [20] to push the clustering memberships of superpixels in matrix \mathbf{G}_s to be as similar as possible to the ground-truth ones \mathbf{G}_s^{gt} .

$$L_s^{(clust)} = \sum_{c=1}^{C_s^{(gt)}} \frac{\langle \mathbf{g}_c, \mathbf{g}_{\mathcal{H}(c)}^{gt} \rangle}{\|\mathbf{g}_c\|_1 + \|\mathbf{g}_{\mathcal{H}(c)}^{gt}\|_1 - \langle \mathbf{g}_c, \mathbf{g}_{\mathcal{H}(c)}^{gt} \rangle} \quad (8)$$

where \mathbf{g}_c and \mathbf{g}_c^{gt} represent the column of the \mathbf{G}_s and \mathbf{G}_s^{gt} respectively. C_s^{gt} is the total number of parts in the ground-truth. $\mathcal{H}(c)$ represents the matched column index c of predicted cluster to the ground-truth cluster based on Hungarian matching [21].

We note that the ILP solution does not participate in our network training implementation. End-to-end training would require methods for differentiating ILPs [11, 27], yet these would make training computationally too expensive.

Implementation Details. The correspondence and clustering modules are trained using the Adam optimizer using the sum of all the above losses. We refer readers to the supplemental for more details, and also to our project page for source code (the link is included in our abstract).

5. Experiments

In this section, we discuss our dataset and results. We also show qualitative and quantitative comparisons.

5.1. Datasets

To provide supervision to our neural modules, we make use of two publicly available datasets.

OkaySamurai dataset. First, we use the publicly available puppets from the OkaySamurai website². The dataset consists of 57 artist-created and rigged characters, with varying numbers of articulated parts and spanning different categories such as full or half body humanoids, dolls, robots, often having accessories such as clothes and hand-held objects. The advantage of this dataset is that the rigged characters are already segmented into parts, which can be used to train our neural modules and allow numerical evaluation. We split the data such that 30 puppets are used for training, 7 for hold-out validation, and 20 for testing. For each training and validation puppet, we generate 200 random poses and sample 100 pose pairs to train the correspondence and clustering modules. The different poses are created by specifying random angles in a range $[-0.3\pi, 0.3\pi]$ to their skeletal joints. We also apply small, additional non-rigid deformations on each body part to improve the pose diversity (see supplementary for details).

Creative Flow+ dataset. Despite augmentation for poses, the number of training puppets in the OkaySamurai remains limited. Since our correspondence module is appearance-sensitive, we can pre-train it separately on larger datasets with ground-truth correspondences. One such example is the recent Creative Flow+ dataset [38]. The dataset contains 2D artistic, cartoon-like renderings of animation sequences along with ground truth pixel-wise correspondences. The dataset does not contain segmentation of articulated parts, yet, it is still a useful source to pretrain our correspondence module. The animation sequences are generated from various 3D meshes. We removed the ones having no articulated pose structure, e.g., the ones generated from ShapeNet models. Since we are also interested in training on pose pairs with large motion variations, we sample poses at least 30 frames in between. In total, we pick 8058 pairs of Creative-Flow+ cartoon renderings for training, 1165 pairs for validation, and 1078 pairs for evaluating our correspondences against alternatives.

SPRITES dataset. We use one more dataset to evaluate how well our method generalizes to other data not involved in our training. We obtained 10 sprite sheets manually created by artists³. We refer to this dataset as “SPRITES”. For each sprite sheet, we gathered 6-10 poses of the character, all artist-drawn. The characters of this dataset are not rigged, nor segmented into parts, thus we use this dataset for qualitative evaluation.

Training strategy. We first pre-train our correspondence module using the InfoNCE loss of Eq. 4 on the Creative-

²<https://www.okaysamurai.com/puppets/>

³we obtained permission to publish them

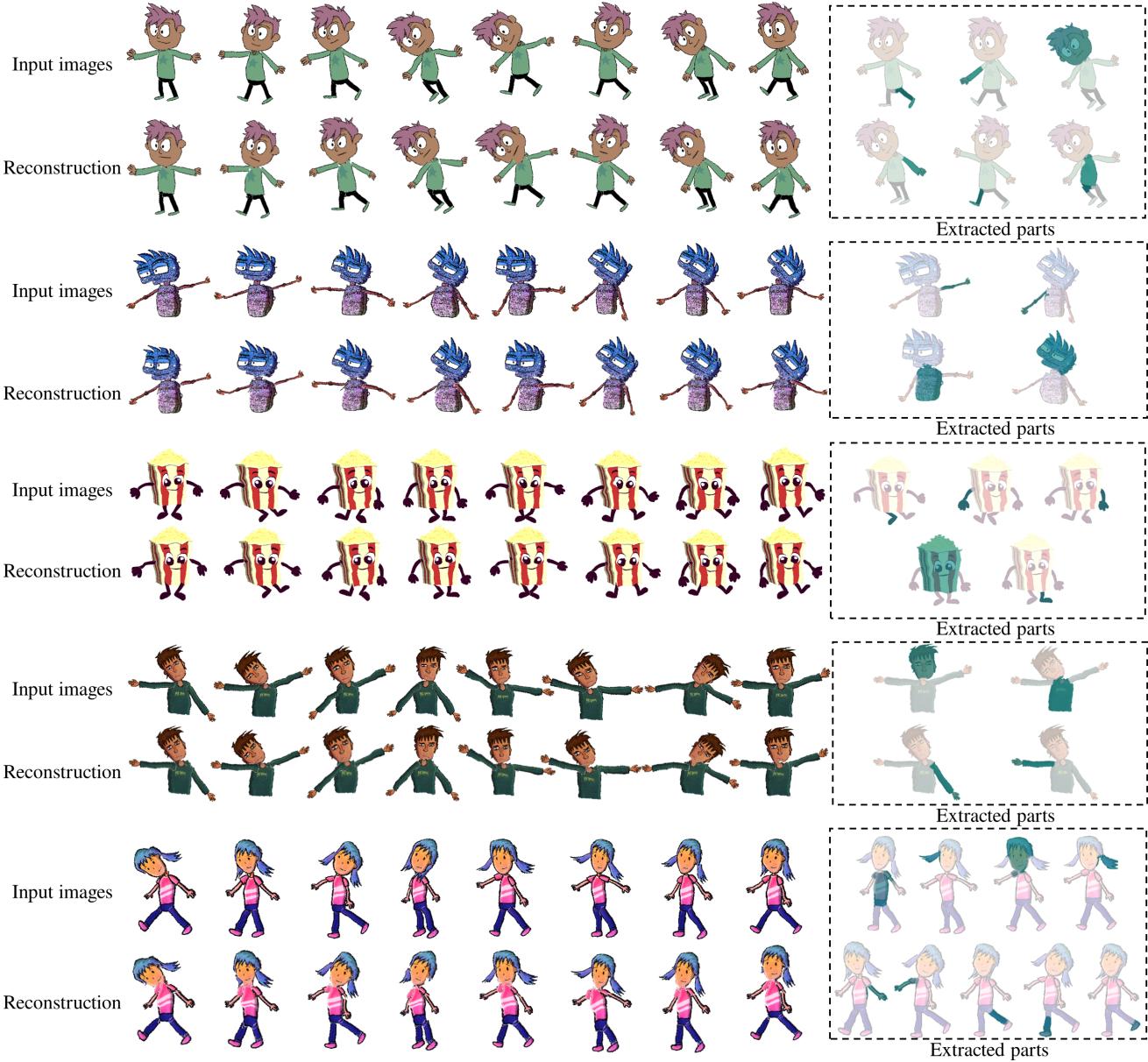


Figure 3. *Left:* The top row of each OkaySamurai test puppet shows the input poses of the sprite sheet. The bottom row shows our reconstructed poses. *Right (box):* The predicted articulated parts from APES.

Flow+ dataset. Starting from the pre-trained correspondence module, we then train both neural modules on the training split of the OkaySamurai dataset. This strategy offered the best performance. We also found helpful to apply color jittering augmentation to each training pair.

Evaluation metrics. The test split from CreativeFlow+ dataset can be used for evaluating correspondence accuracy. We use the end-point-error (**EPE**) as our evaluation protocol which measures the average distance between the predicted and the ground-truth corresponding pixels.

The test split of OkaySamurai is used to evaluate part extraction. For each testing puppet, we generate 10 different

poses to get 200 test poses. We process all possible pairs (45 pairs per puppet) through our trained correspondence, clustering modules and part selection procedure to output selected articulated parts for each puppet. We also deform the selected parts by ARAP to best reconstruct each input pose (see Sec. 3.3). For evaluating the output parts, we first perform Hungarian matching between ground-truth and reconstructed parts based on Intersection over Union (IoU), with $1 - IoU$ is used as cost. The resulting average **part IoU** is used as our main evaluation metric. As additional evaluation metrics, we also use the difference between the reconstructed and ground-truth poses by **MSE**, **PNSR** and

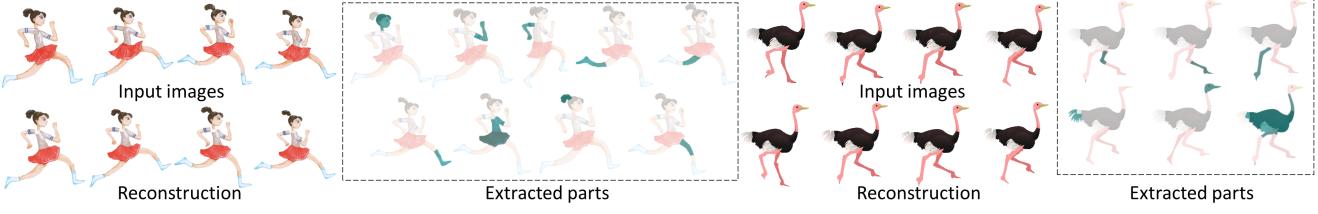


Figure 4. Part extraction from sprite sheets created by artists (“SPRITES” dataset). In the box, we show the predicted articulated parts.

LPIPS [62]. High reconstruction error indicates implausible parts used in deformation.

5.2. Articulated Parts Selection from Sprite Sheets

Figure 3 shows our articulated part extraction results from our method for characteristic sprite sheets from the OkaySamurai dataset. We also include reconstruction results based on the deformation procedure described in Section 3.3 on the second row of each example. Our method successfully recovers articulated parts in most cases, although boundaries of parts are not always accurate (e.g., see shoulders and hips in the last example). Figure 4 shows results from the SPRITES dataset. Our method is able to detect intuitive articulated parts in these artist-drawn poses, although regions near part boundaries (e.g., legs, tail of bird) are slightly grouped off.

Our supplementary material includes additional qualitative results from the CreativeFlow+ dataset. In addition, the supplementary video shows applications of our method to automatic puppet creation and automatic synthesis of animation skeletons based on our identified parts.

5.3. Comparisons

Articulated part extraction. Our method (APES) is the first to deal with articulated part extraction from sprite sheets. There are no prior methods that have been applied to this problem. Yet, one important question is whether methods that have been developed for part co-segmentation in photorealistic images can be applied to our problem.

One may argue that appearance cues might be enough to detect the common parts across different poses of a character. To test this hypothesis, we perform comparisons with **SCOPS** [17], a state-of-the-art co-part segmentation method. The method is self-supervised, and the self-supervision is applied to real-world images. We train SCOPS on the same training sources with our method (CreativeFlow+ and OkaySamurai), and we also add supervisory signal using our clustering loss. We call this super-

vised variant as **SCOPS-s**. We note that SCOPS does not make use of optical flow or external correspondences, thus APES still uses more supervision than SCOPS-s. Nevertheless, we consider useful to show this comparison, since SCOPS is a characteristic example of a method that does not consider motion cues. We also note that SCOPS uses a semantic consistency loss that makes segmentations more consistent across objects of the same category. We tested SCOPS-s with and without this loss; we refer to these variants as **SCOPS-s (sc)** and **SCOPS-s (nosc)**. We exhaustively tested the loss weights to find the best configuration, and select the best number of output parts (12 parts). We note that the output segmentation map from SCOPS includes a background region – we ignore it in our evaluation. The resulting part regions can be evaluated with the same metrics, averaged over all poses of the test puppets of OkaySamurai. We finally note that SCOPS cannot perform reconstruction, thus, we report only segmentation performance in the OkaySamurai test dataset.

Table 1 presents the average part IoU for the OkaySamurai test set. Note that we also include the performance of the original self-supervised SCOPS approach just for reference. All SCOPS variants have low performance e.g., APES’ IoU in OkaySamurai is 71%, twice as high compared to the best SCOPS variant (35%). The results indicate that appearance-based co-part segmentation is not effective at extracting articulated parts in sprite sheets.

An alternative co-segmentation approach is the one by Siarohin *et al* [39], which relies on motion cues. We denote it as **MoCoSeg**. Like SCOPS, it is self-supervised and trained on videos. Similarly, to make a fair comparison, we re-train this method on the same training sources as ours, and also add supervisory signal using both our clustering loss and our correspondence loss on its flow output. We call this supervised variant as **MoCoSeg-s**. We tuned their loss weights and output number of parts to achieve best performance in the validation split.

Table 1 shows the performance of MoCoSeg-s and MoCoSeg, both of which are low. We suspect that the motion inferred by MoCoSeg is correlated to their part segmentation, which is more suitable for objects with consistent articulation structure. This indicates that such motion-based segmentation methods are not appropriate for our setting.

Figure 5 shows characteristic outputs from the above

Method	IoU
SCOPS [17]	27.4%
SCOPS-s (sc)	33.1%
SCOPS-s (nosc)	35.8%
MoCoSeg [39]	26.0%
MoCoSeg-s	32.3%
APES	71.0%

Table 1. Results in the OkaySamurai test set.

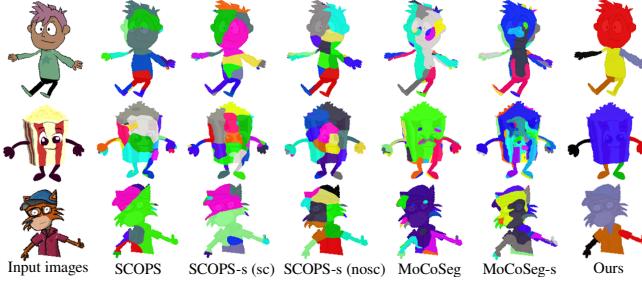


Figure 5. Identified parts from different methods on characteristic poses from the OkaySamurai dataset. Note that different color indicates different parts (colors have no semantic correspondence). APES identifies articulated parts much more successfully.

Method	RAFT [42]	COTR [18]	APES
EPE (end-point-error)	28.07	31.93	22.90

Table 2. Quantitative results on the Creative Flow test split based on the EPE metric. Our method achieves the lowest EPE.

SCOPS and MoCoSeg variants and APES. Our method can infer articulated parts from the input poses much more accurately compared to competing methods, aligning better with the underlying articulated motion.

Correspondences. We also evaluate our correspondence module against alternatives. First we compare our method with an optical flow method **RAFT** [42]. For RAFT, we remove their online-generated masks to allow longer-range optical flows and incorporate foreground masks in the correlation operation so that only foreground pixels have positive correlation. In addition, we shifted the predicted corresponding pixels to their nearest foreground pixels. We note that we fine-tuned RAFT on the same training datasets as ours (CreativeFlow+ and OkaySamurai training splits). This worked better compared to training it from scratch on our datasets, or using its pre-trained model without fine-tuning. Table 2 shows quantitative results on our test split of CreativeFlow+. Our correspondence module produces much more accurate correspondences compared to RAFT.

We also compare to a pixel correspondence method based on transformers, called **COTR** [18]. We fine-tuned COTR on the same training splits as ours, and shifted the predicted corresponding pixels to their nearest foreground pixels. Still, COTR’s results are inferior to APES. For visualization of correspondence results from our method and others, please see our supplementary material.

5.4. Ablation Study

We perform a set of ablation experiments on the OkaySamurai dataset since it includes ground-truth articulated parts for evaluation. We compare with the following variants of our method: **RT_simult**: we predict rotation and translation of each superpixel simultaneously, instead of sequentially as in our original method. **No Eq. 7**: we train the

Method	IoU	MSE	PNSR	LPIPS
RT_simult	69.5%	741.58	20.19	0.10
No Eq. 7	70.1%	749.58	20.14	0.10
No Eq. 6	59.4%	838.11	19.47	0.11
RAFT corr.	63.5%	788.39	19.78	0.11
COTR corr.	60.2%	862.28	19.40	0.12
APES	71.0%	733.12	20.20	0.10

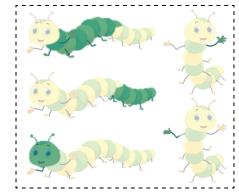
Table 3. Ablation study of our variants (OkaySamurai dataset).

segmentation module with Eq. 6 and Eq. 8 only, without supervision for the motion residual matrix. **No Eq. 6**: we train the segmentation module with Eq. 7 and Eq. 8 only, without supervision on the affinity matrix. **RAFT corr**: Instead of our UNet-based correspondence module, we use RAFT to predict correspondences used in the following steps. **COTR corr**: we use COTR to produce correspondences instead.

We report all our evaluation metrics in Table 3, including reconstruction metrics, since all the above variants employ the same part selection and reconstruction stage. We observe inferior results from all reduced variants.

6. Conclusion

We presented APES, a method that extracts articulated parts from a sparse set of character poses of a sprite sheet. As far as we know, APES is the first method capable of automatically extracting deformable puppets from unsegmented character poses. We believe that methods able to parse character artwork and generate rigs have the potential to significantly automate the character animation workflow.



Limitations. While we can handle a wide range of character styles and part configurations, there are some limitations. Parts with non-rigid or subtle motion cannot be extracted well. For example, given the caterpillar poses shown on the right, our method extracts the head, arms, and body chunks, yet it does not segment the thin legs and the individual abdomen segments, since these do not seem to have distinct rotations with respect to the rest of the body. As discussed in our experiments, the boundary of parts is not always accurate. Our method selects parts such that they minimally overlap during part selection. As a result, small articulated parts might be missed and replaced by larger ones. Handling strongly overlapping parts more explicitly, layer order changes between sprites (such as a character turning around), and large occlusions would make our method applicable to a wider range of sprite sheet cases.

Acknowledgements. Our research was partially funded by NSF (EAGER-1942069) and Adobe.

References

- [1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE TPAMI*, 34(11), 2012. 3
- [2] Adobe. *Character Animator, version*. 2021. <https://www.adobe.com/products/character-animator.html>. 2
- [3] Federica Arrigoni and Tomas Pajdla. Motion segmentation via synchronization. In *ICCV Workshops*, 2019. 4
- [4] Dana H Ballard. Generalizing the hough transform to detect arbitrary shapes. *PR*, 13(2), 1981. 3
- [5] P.J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE TPAMI*, 14(2), 1992. 11
- [6] Péter Borosán, Ming Jin, Doug DeCarlo, Yotam Gingold, and Andrew Nealen. Rigmesh: automatic rigging for part-based shape modeling and deformation. *ACM TOG*, 31(6), 2012. 2
- [7] Hyung Jin Chang and Yiannis Demiris. Highly articulated kinematic structure estimation combining motion and skeleton information. *IEEE TPAMI*, 40(9), 2017. 2
- [8] Subhabrata Choudhury, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Unsupervised part discovery from contrastive reconstruction. In *NeurIPS*, 2021. 2
- [9] Edo Collins, Radhakrishna Achanta, and Sabine Susstrunk. Deep feature factorization for concept discovery. In *ECCV*, 2018. 2
- [10] Luca Del Pero, Susanna Ricco, Rahul Sukthankar, and Vittorio Ferrari. Discovering the physical parts of an articulated object class from multiple videos. In *CVPR*, 2016. 2
- [11] Aaron Ferber, Bryan Wilder, Bistra Dilkina, and Milind Tambe. Mipaal: Mixed integer program as a layer. In *AAAI*, 2020. 5
- [12] Robert M Freund. Primal-dual interior-point methods for linear programming based on newton’s method. *Massachusetts Institute of Technology*, 2004. 4
- [13] David S Hayden, Jason Pacheco, and John W Fisher. Non-parametric object and parts modeling with lie group dynamics. In *CVPR*, 2020. 2
- [14] Tobias Hinz, Matthew Fisher, Oliver Wang, Eli Shechtman, and Stefan Wermter. Chartergan: Few-shot key-point character animation and reposing. *arXiv preprint arXiv:2102.03141*, 2021. 2
- [15] Berthold KP Horn. Closed-form solution of absolute orientation using unit quaternions. *Josa a*, 4(4), 1987. 3
- [16] Jiahui Huang, He Wang, Tolga Birdal, Minhyuk Sung, Federica Arrigoni, Shi-Min Hu, and Leonidas Guibas. Multi-bodysync: Multi-body segmentation and motion estimation via 3d scan synchronization. In *CVPR*, 2021. 2, 4, 5
- [17] Wei-Chih Hung, Varun Jampani, Sifei Liu, Pavlo Molchanov, Ming-Hsuan Yang, and Jan Kautz. Scops: Self-supervised co-part segmentation. In *CVPR*, 2019. 1, 2, 7
- [18] Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. Cotr: Correspondence transformer for matching across images. In *ICCV*, 2021. 3, 8, 12
- [19] Armand Joulin, Francis Bach, and Jean Ponce. Multi-class cosegmentation. In *CVPR*, 2012. 2
- [20] Philipp Krähenbühl and Vladlen Koltun. Parameter learning and convergent inference for dense random fields. In *Int. Conf. Machine Learning*, 2013. 5
- [21] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2), 1955. 5
- [22] Hao Li, Guowei Wan, Honghua Li, Andrei Sharf, Kai Xu, and Baoquan Chen. Mobility fitting using 4d ransac. *Computer Graphics Forum*, 35(5), 2016. 2
- [23] Ting Li, Vinutha Kallem, Dheeraj Singaraju, and René Vidal. Projective factorization of multiple rigid-body motions. In *CVPR*, 2007. 2
- [24] Songrun Liu, Alec Jacobson, and Yotam Gingold. Skinning cubic b  zier splines and catmull-clark subdivision surfaces. *ACM TOG*, 33(6), 2014. 1, 2
- [25] Erika Lu, Forrester Cole, Tali Dekel, Weidi Xie, Andrew Zisserman, David Salesin, William T Freeman, and Michael Rubinstein. Layered neural rendering for retiming people in video. In *SIGGRAPH Asia*, 2020. 11
- [26] Xiankai Lu, Wenguan Wang, Jianbing Shen, David Crandall, and Jiebo Luo. Zero-shot video object segmentation with co-attention siamese networks. *IEEE TPAMI*, 2020. 2
- [27] Jayanta Mandi and Tias Guns. Interior point solving for lp-based prediction+optimisation. In *NeurIPS*, 2020. 5
- [28] Natalia Neverova, David Novotny, Vasil Khalidov, Marc Szafraniec, Patrick Labatut, and Andrea Vedaldi. Continuous surface embeddings. In *NeurIPS*, 2020. 4
- [29] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *NeurIPS*, 2002. 4
- [30] Peter Ochs and Thomas Brox. Object segmentation in video: A hierarchical variational approach for turning point trajectories into dense regions. In *ICCV*, 2011. 2
- [31] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4
- [32] Omid Poursaeed, Vladimir Kim, Eli Shechtman, Jun Saito, and Serge Belongie. Neural puppet: Generative layered cartoon characters. In *WACV*, 2020. 2
- [33] 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 *Int. Conf. Machine Learning*, 2021. 4
- [34] Prabhakar Raghavan and Clark D Tompson. Randomized rounding: a technique for provably good algorithms and algorithmic proofs. *Combinatorica*, 7(4), 1987. 4
- [35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241, 2015. 3
- [36] 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*, 115(3), 2015. 2

- [37] Sara Sabour, Andrea Tagliasacchi, Soroosh Yazdani, Geoffrey Hinton, and David J Fleet. Unsupervised part representation by flow capsules. In *Int. Conf. Machine Learning*, 2021. 2
- [38] Maria Shugrina, Ziheng Liang, Amlan Kar, Jiaman Li, An-gad Singh, Karan Singh, and Sanja Fidler. Creative flow+ dataset. In *CVPR*, 2019. 5
- [39] Aliaksandr Siarohin, Subhankar Roy, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Motion-supervised co-part segmentation. In *ICPR*, 2021. 1, 2, 7
- [40] Yafei Song, Xiaowu Chen, Jia Li, and Qinping Zhao. Embedding 3d geometric features for rigid object part segmen-tation. In *ICCV*, 2017. 2
- [41] Olga Sorkine and Marc Alexa. As-rigid-as-possible surface modeling. In *SGP*, volume 4, 2007. 4, 11
- [42] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 3, 8, 12
- [43] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning motion patterns in videos. In *CVPR*, 2017. 2
- [44] Pavel Tokmakov, Cordelia Schmid, and Karteek Alahari. Learning to segment moving objects. *IJCV*, 127(3), 2019. 2
- [45] Roberto Tron and René Vidal. A benchmark for the comparison of 3-d motion segmentation algorithms. In *CVPR*, 2007. 2
- [46] Yi-Hsuan Tsai, Guangyu Zhong, and Ming-Hsuan Yang. Semantic co-segmentation in videos. In *ECCV*, 2016. 2
- [47] Sara Vicente, Carsten Rother, and Vladimir Kolmogorov. Object cosegmentation. In *CVPR*, 2011. 2
- [48] René Vidal, Yi Ma, Stefano Soatto, and Shankar Sastry. Two-view multibody structure from motion. *IJCV*, 68(1), 2006. 2
- [49] J.Y.A. Wang and E.H. Adelson. Layered representation for motion analysis. In *CVPR*, 1993. 2
- [50] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Int. Conf. Machine Learning*, 2020. 3
- [51] Xiaogang Wang, Bin Zhou, Yahao Shi, Xiaowu Chen, Qin-ping Zhao, and Kai Xu. Shape2motion: Joint analysis of motion parts and attributes from 3d shapes. In *CVPR*, 2019. 2
- [52] Nora S Willett, Wilmot Li, Jovan Popovic, Floraine Berthouzoz, and Adam Finkelstein. Secondary motion for performed 2d animation. In *UIST*, 2017. 1
- [53] Xun Xu, Loong-Fah Cheong, and Zhuwen Li. 3d rigid motion segmentation with mixed and unknown number of mod-els. *IEEE TPAMI*, 43(1), 2019. 2
- [54] Zhenjia Xu, Zhijian Liu, Chen Sun, Kevin Murphy, William T Freeman, Joshua B Tenenbaum, and Jiajun Wu. Unsupervised discovery of parts, structure, and dynamics. In *ICLR*, 2019. 2
- [55] Zhan Xu, Yang Zhou, Evangelos Kalogerakis, Chris Lan-dreth, and Karan Singh. Rignet: Neural rigging for articu-lated characters. *ACM TOG*, 39, 2020. 2
- [56] Jingyu Yan and Marc Pollefeys. A factorization-based ap-proach for articulated nonrigid shape, motion and kinematic chain recovery from video. *IEEE TPAMI*, 30(5), 2008. 2
- [57] Gengshan Yang and Deva Ramanan. Learning to segment rigid motions from two frames. In *CVPR*, 2021. 2
- [58] Li Yi, Haibin Huang, Difan Liu, Evangelos Kalogerakis, Hao Su, and Leonidas Guibas. Deep part induction from articu-lated object pairs. *ACM TOG*, 37(6), 2018. 2
- [59] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *ICCV*, 2019. 3, 12
- [60] Chang Yuan, Gerard Medioni, Jinman Kang, and Isaac Cohen. Detecting motion regions in the presence of a strong parallax from a moving camera by multiview geometric con-straints. *IEEE TPAMI*, 29(9), 2007. 2
- [61] Chi Zhang, Guankai Li, Guosheng Lin, Qingyao Wu, and Rui Yao. Cyclesegnet: Object co-segmentation with cycle refinement and region correspondence. *IEEE TIP*, 2021. 2
- [62] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 7

Appendix

A. Applications

In our supplementary video <https://youtu.be/YQtbRXFKNZE>, we show two applications based on extracted articulated parts from APES: part-based animation and puppet creation via part swapping. We refer readers to the video for a demonstration of these applications.

For part-based animation, we create a set of control points or joints based on the extracted parts. We follow a simple heuristic to obtain them. First, we compute the centroid point of the whole character, and designate the part closest to the centroid as the “central” part (this coincides with the torso in Fig. 6). We create a joint at the center of this central part. The rest of the parts are designated as “limbs”. We create joints (“pin joints”) at the center of the intersection between the limbs and the central part (e.g., hips, shoulders in Fig. 6). Finally, we extract the medial axis of each limb, and create a joint at the medial axis point furthest to the pin joint for each limb (these points correspond to fingers and toes in Fig. 6). The resulting joints form a simple control structure (e.g., an animation skeleton) that can be loaded into standard animation software. We use the Adobe’s Character Animator in the supplementary video. By manipulating the joint positions and angles, users can animate the characters based on the joints and parts extracted by APES.

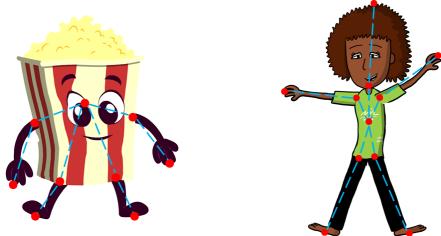


Figure 6. Control points (red dots) and skeletons created based on the extracted parts from APES.

B. Non-rigid Deformation for Reconstruction

As explained in Section 3.3, our part selection procedure uses a random rounding algorithm to solve the integer linear programming problem. The random rounding algorithm gives us multiple possible solutions. We pick the best solution in terms of reconstruction error after deforming each candidate set of parts to reconstruct the input poses. We describe here the deformation procedure.

Let $\mathbf{Q} = \{\mathbf{q}_c\}_{c=1}^C$ be a set of selected parts (where C is their total number), and the image \mathbf{I}_t be the target pose to be reconstructed. A straight-forward method for recon-

struction is fitting the optimal translation and rotation transformation for each part based on the predicted correspondences using the Procrustes orthogonal analysis used in ICP methods [5]. Such method is computationally efficient, but is sensitive to the noise in the predicted correspondences, and cannot capture non-rigid deformations that may exist in the input poses. Instead, we follow an as-rigid-as-possible deformation procedure (ARAP [41]). The ARAP takes as input a control mesh, and target positions for one or more of its vertices. To create a control mesh for each part $\mathbf{q}_c \in \mathbf{Q}$, we first uniformly sample a set of vertices \mathcal{V}_c from its oriented bounding box. Then we create the Delaunay triangulation of the vertices. Each control mesh also incorporates the texture from the original appearance of its associated part. We treat the target positions $\{\mathcal{V}'_c\}_{c=1}^C$ of the mesh vertices as unknowns in an optimization problem that attempts to deform the control meshes of all parts such that their resulting appearance is as close as possible to the target pose \mathbf{I}_t , while at the same time the part deformations are as-rigid-as-possible. Specifically, we solve the following problem:

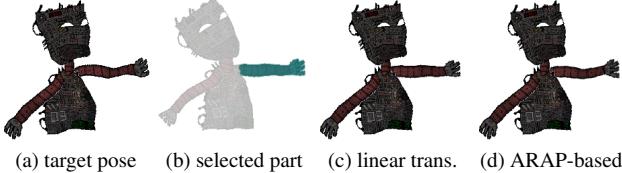
$$L(\mathcal{V}'_c) = \left\| \sum_{c=1}^C \Phi(\mathcal{V}'_c) - \mathbf{I}_t \right\|_2^2 + \lambda_r \sum_{c=1}^C \mathcal{R}(\mathcal{V}_c, \mathcal{V}'_c) \quad (9)$$

where λ_r is set to 0.05 in our experiments. The first term measures the reconstruction error between the deformed parts (caused by the shifted vertices) and the target pose. Φ renders each part with the shifted control vertices \mathcal{V}'_c based on barycentric coordinates and bilinear interpolation [25], which is differentiable. The second term is a regularization term that preserves the local shape during deformation. Specifically, it penalizes deviation from local rigid deformation for each control vertex neighborhood [41]:

$$\mathcal{R}(\mathcal{V}_c, \mathcal{V}'_c) = \sum_{\mathbf{v}_i \in \mathcal{V}_c} \sum_{\mathbf{v}_j \in \mathcal{N}(\mathbf{v}_i)} \|(\mathbf{v}'_j - \mathbf{v}'_i) - \mathbf{R}_i(\mathbf{v}_j - \mathbf{v}_i)\|^2 \quad (10)$$

where $\mathcal{N}(\mathbf{v}_i)$ is the neighborhood of the control vertex \mathbf{v}_i and \mathbf{R}_i is the best fit rotation of its neighborhood to the deformed configuration. The best fit rotations are computed via orthogonal Procrustes analysis. The optimization problem is solved iteratively. At each iteration, we alternate between solving for the deformed control vertices, and optimal rotations, as proposed in [41]. Fig. 7 shows an example of our reconstruction approach.

Non-rigid deformation for augmentation. As discussed in Section 5.1, we apply small, non-rigid deformations on each training body part to improve the pose diversity during training for the OkaySamurai dataset. To do so, for each part we uniformly sample control vertices in its oriented bounding box, and randomly shift the points by offsets sampled from Gaussian distribution $\mathcal{N}(0, \sigma^2)$ where σ is set as 2% of the maximum image dimension. We form the control



(a) target pose (b) selected part (c) linear trans. (d) ARAP-based

Figure 7. Example of reconstruction via non-rigid deformation. By using the best-fit rigid transformation, the deformation (c) of the left arm cannot reconstruct well the target pose (a). Using the ARAP-based reconstruction (d), the arm is aligned better to the target appearance.

meshes for parts, and deform them to reach the shifted control vertices using an ARAP deformation similar to Eq. 9.; here the reconstruction error measures difference between deformed and original control vertex positions.

C. Architecture Details

We provide here additional details of our network architecture.

Correspondence Module. Table 4 lists the layers used in the correspondence module along with the size of their output map. We call the architecture of our module as “Gated UNet” since the convolutional layers in the encoder implement Gated Convolution [59].

Clustering Module. Table 5 shows the architecture of the clustering module.

D. Implementation Details

The correspondence module is first trained alone. We set the learning rate to 10^{-3} and decrease it to 10^{-4} after 5 epochs. Then we train both the correspondence and clustering modules using the soft nearest neighbor in Equation 5 of the main text. We set the learning rate to 10^{-6} for the correspondence module and 10^{-4} for the segmentation module with a batch size of 8 for this stage. We use the Adam optimizer. The height H and width W of the images and voting maps are always 256. The number of clusters C_s during training is set to 12.

E. Correspondence Comparison

Fig. 8 shows a qualitative comparison example of predicted correspondences between our method, RAFT [42] and COTR [18]. To help RAFT and COTR better use the foreground masks, we map the predicted target positions to their nearest neighbors in the foreground. RAFT and COTR cannot produce correspondences reliably e.g., for hand and head regions, as shown below.

Table 4. **Correspondence module architecture (Gated UNet).** $Conv3x3$ is convolutional layer with kernel size 3. LN is LeakyReLU with negative slope as 0.2. BN is BatchNorm. Dot between LR and Sigmoid is element-wise product. $Upsample$ uses bilinear interpolation for upsampling. We note that there are skip connections between corresponding layers from encoder and decoder, following the original U-Net architecture.

	Layers	Output
Input	Concat(image, mask)	$256 \times 256 \times 4$
Encoder	$BN(LR(Conv3x3) \cdot Sigmoid(Conv3x3))$	$256 \times 256 \times 32$
	$BN(LR(Conv3x3) \cdot Sigmoid(Conv3x3))$	$256 \times 256 \times 32$
	MaxPooling(2)	$128 \times 128 \times 32$
	$BN(LR(Conv3x3) \cdot Sigmoid(Conv3x3))$	$128 \times 128 \times 64$
	$BN(LR(Conv3x3) \cdot Sigmoid(Conv3x3))$	$128 \times 128 \times 64$
	MaxPooling(2)	$64 \times 64 \times 64$
	$BN(LR(Conv3x3) \cdot Sigmoid(Conv3x3))$	$64 \times 64 \times 128$
	$BN(LR(Conv3x3) \cdot Sigmoid(Conv3x3))$	$64 \times 64 \times 128$
	MaxPooling(2)	$32 \times 32 \times 128$
	$BN(LR(Conv3x3) \cdot Sigmoid(Conv3x3))$	$32 \times 32 \times 256$
Decoder	$BN(LR(Conv3x3) \cdot Sigmoid(Conv3x3))$	$32 \times 32 \times 256$
	MaxPooling(2)	$16 \times 16 \times 256$
	$BN(LR(Conv3x3) \cdot Sigmoid(Conv3x3))$	$16 \times 16 \times 256$
	$BN(LR(Conv3x3) \cdot Sigmoid(Conv3x3))$	$16 \times 16 \times 256$
	Upsample(2)	$32 \times 32 \times 256$
	$BN(ReLU(Conv3x3))$	$32 \times 32 \times 128$
	$BN(ReLU(Conv3x3))$	$32 \times 32 \times 128$
	Upsample(2)	$64 \times 64 \times 128$
	$BN(ReLU(Conv3x3))$	$64 \times 64 \times 64$
	$BN(ReLU(Conv3x3))$	$64 \times 64 \times 64$
Decoder	Upsample(2)	$128 \times 128 \times 64$
	$BN(ReLU(Conv3x3))$	$128 \times 128 \times 32$
	$BN(ReLU(Conv3x3))$	$128 \times 128 \times 32$
	Upsample(2)	$256 \times 256 \times 32$
	$BN(ReLU(Conv3x3))$	$256 \times 256 \times 32$
	$BN(ReLU(Conv3x3))$	$256 \times 256 \times 32$
	Upsample(2)	$256 \times 256 \times 32$
	$BN(ReLU(Conv3x3))$	$256 \times 256 \times 32$
	$BN(ReLU(Conv3x3))$	$256 \times 256 \times 64$

Table 5. **Clustering module architecture.** The symbol “ $pred.R$ ” means predicted rotations, and “ $pred.T$ ” means predicted translations. To update the voting map, we apply the predicted rotations to source pixels.

	Layers	Output
Input	Concat(voting map, mask)	$256 \times 256 \times 5$
Gated UNet	similar to Table 4 with intermediate channel numbers as 16, 32, 64, 128, 256	
	N/A	$K_s \times 16$
Average pooling per superpixel	16 → 64 → 2	$pred.R: K_s \times 2$
	Updated input	
Gated UNet	Concat(updated voting map, mask)	
	similar to Table 4 with intermediate channel numbers as 16, 32, 64, 128, 256	
Average pooling per superpixel	N/A	$K_s \times 16$
	16 → 64 → 2	$pred.T: K_s \times 2$

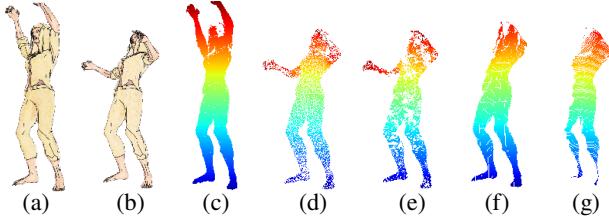


Figure 8. Visualization of the predicted correspondences from our method, RAFT and COTR. From left to right: (a) source image (b) target image (c) color-coded source pixels (d) ground-truth target pixels (e) prediction of our method (f) prediction of RAFT (g) prediction of COTR. Corresponding pixels have same color in the visualization of correspondence maps. Our method matches pixels more accurately compared to other methods.

F. More Qualitative Results

We show here additional qualitative results from the CreativeFlow+ dataset. We note that this dataset does not include segmentations of articulated parts. Their provided segmentation maps are based on mesh components, which often do not match articulation. Thus, as an additional test set, we manually segmented 12 characters into rigid parts from their test split, and used them as reference to measure the performance of our trained model on them. We achieve an IoU of 67%, which is slightly lower than the IoU we achieved for the OkaySamurai dataset (71%). Two examples are shown in Fig. 9.

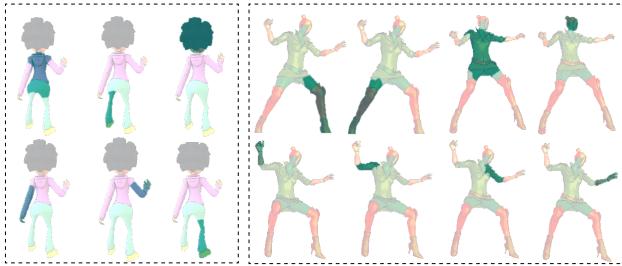


Figure 9. Part extraction results in the CreativeFlow+ dataset.

We include the results for the test sprite sheets in OkaySamurai dataset and SPRITE in our code repository.