IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 January 2024, accepted 15 January 2024, date of publication 18 January 2024, date of current version 15 February 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3355814*

RESEARCH ARTICLE

# CorrFractal: High-Resolution Correspondence Method Using Fractal Affinity on Self-Supervised Learning

**JIN-MO CHOI**[1], **BLAGOVEST I. VLADIMIROV**[2], **AND SANGJOON PARK**[2]

[1]Department of Computer Software, University of Science and Technology, Daejeon 34113, Republic of Korea
[2]Defence and Safety Convergence Research Division, Electronics and Telecommunications Research Institute, Daejeon 34129, Republic of Korea

Corresponding author: Sangjoon Park (sangjoon@etri.re.kr)

This work was supported by the Electronics and Telecommunications Research Institute (ETRI) Research and Development Support Program of Ministry of Science and ICT (MSIT)/Institute for Information & Communication Technology Planning & Evaluation (IITP), Republic of Korea, through the Project "Development of Beyond X-Verse Core Technology for Hyper-Realistic Interactions by Synchronizing the Real World and Virtual Space" under Project RS-2023-00216821.

**ABSTRACT** Existing supervised learning-based methods performed high-resolution visual correspondence using a decoder module. However, in self-supervised learning-based methods, it is difficult to use a decoder module that is easily influenced by labels. This paper will introduce a self-supervised learning-based visual correspondence method for high-resolution representation without decoder module. To this end, the paper proposed four modules. Each module has an output of the original resolution and distributes the role of the decoder module to perform high-resolution representation. The first module is the pattern boosted quantization module, which learns pattern information along with color information to create high-resolution pseudo labeling. The second module is the backbone module, which is created by applying aggregation to the backbone network to simultaneously handle semantic features and high-resolution features. The third module is the appearance module, which learns appearance information using the features of the high-resolution embedding space. The fourth module is the correspondence module, which gradually reconstructs a high-resolution visual correspondence using low-resolution input. It was confirmed using subtraction image that the proposed method improves the performance about representation of thin objects and object boundaries. Video segmentation performance was evaluated on the DAVIS-2017 val dataset using the $J\&F$ mean, yielding 65.4%.

**INDEX TERMS** Decoder module, high-resolution representation, pseudo labeling, self-supervised learning, visual correspondence.

## I. INTRODUCTION

Visual correspondence is the problem of predicting where a specific point of a reference image is located in a target image. It is a core method related to various fields, including image matching [1], [2], key point matching [3], [4], tracking [5], [6], [7], stereo vision problem [8], colorizing [9], and video segmentation [10], [11], [12], [13], [14], [15]. Different methods of solving the problems of long-term prediction, occlusion, and drift using self-supervised learning have recently been proposed [10], [11], [12], [13], [14], [15]. However, the high-resolution representation could not efficiently be handled due to the high computational complexity of the affinity matrix. In other words, it cannot distinguish clear boundaries between objects and cannot express objects composed of high frequencies (e.g., thin threads) [16], [17], [18], [19], [20], [21]. On the other hand, supervised learning-based methods efficiently solved high-resolution representation problems using a decoder module. The problem is that it is difficult to use a decoder module that is directly affected by labels in self-supervised methods. In addition, performing high-resolution

The associate editor coordinating the review of this manuscript and approving it for publication was Abdel-Hamid Soliman.

representation without using a decoder module significantly increases memory usage because it increases the computational complexity of the affinity matrix. To solve this problem, a method is needed that can use memory efficiently while maintaining high-resolution features. The proposed method will target the original resolution representation of the input image, unlike existing methods [12], [13] that use a resolution representation with stride 4 for high resolution representation.

Various modules have been presented to solve these issues. The first module is the pseudolabeling module that adds the Walsh–Hadamard pattern [22] to the Lab color space. It uses a mixture of pattern and color information to express an object's intrinsic characteristics. The second module is a network architecture aggregation for creating high-resolution features. The third module is the appearance module that learns the appearance using high-resolution features. The fourth module is the correspondence module that performs resolution reconstruction using the fractal structure of the affinity matrix.

The usage of pseudolabeling is an important and controversial issue in the topic of visual correspondence through self-supervised learning. Reference [11] performed pseudolabeling by applying k-means clustering on the ab channel of the Lab color space, consequently inspiring subsequent papers to utilize pseudolabeling and influencing the development of a more sophisticated pseudolabeling process. References [12] and [13] approached it as a regression problem by randomly selecting only one channel among the ab channels. Reference [14] solved the correspondence problem using cycle consistency, taking advantage of the fact that if the first frame is predicted in backward time, and then predicted again in forward time, it always has the same features as the first frame. The abovementioned methods have the advantage of being less influenced by hypotheses from human bias by avoiding to perform pseudolabeling. In contrast, existing methods that use cycle consistency have the disadvantage of limiting the affinity matrix size because they continuously predict the affinity matrix as long as the length on the time axis. Conversely, methods implementing pseudolabeling have relatively less restrictions on the affinity matrix size because usage of the time axis information is not necessary. Therefore, methods requiring pseudolabeling are advantageous when learning high-resolution features by being able to use a relatively large affinity matrix. However, the existing Lab color-based pseudolabeling method does not consider the object's pattern and does not give weight to high-frequency information, making it insufficient for expressing high-resolution features and clear boundary information. To better learn high-resolution features, the Walsh—Hadamard pattern is added along with the Lab color space and clustered with k-means.

Visual dense correspondence is a pixel-to-pixel matching method for objects existing in the reference and target images. This problem deals with object motion; thus, objects must be understood. To do this, methods for separately dealing with the appearance and motion features are studied [23], [24], [25]. This approach was similar to the concept of recognizing salient objects by integrating a dorsal pathway for motion perception and a ventral pathway for semantic perception in neuroscience [26]. Reference [27] constructed the appearance module and correspondence module by sharing embedding features from the backbone. In this approach, the correspondence module learns motion for visual correspondence, while the appearance module mitigates the degradation of appearance caused by the correspondence module by using a narrow search range for the affinity matrix. However, information loss occurred because the appearance module did not use the feature map of the spatial resolution corresponding to the spatial resolution of the quantized image. The proposed method enhances the performance of the appearance module by generating high-resolution feature maps through aggregation [28], [29], [30], [31], [32], [33] in the existing network architecture.

In visual correspondence, the optimal trade-off between the subsampling of the embedding space and the matching range must be found because the amount of memory used for the affinity matrix is proportional to the square of the feature sampling due to the affinity matrix being calculated as the inner product of the reference and target frames. Accordingly, most methods use the coarse subsampling of the reference and target features to prevent unrealistic memory usage. However, these methods impair the feature details. References [12] and [13] calculated the inner product only within the region of interest of the window size using restricted attention. Their approach had the advantage of maintaining the feature details because the affinity matrix was calculated at a relatively high resolution by narrowing the search area. In contrast, the motion beyond the search area had not been understood due to the narrow search space.

The proposed method uses coarse subsampling on restricted attention to simultaneously address the search area limitation and loss of feature details and obtain a wide search range. It uses the fractal structure of the affinity matrix on the correspondence module to restore the feature details.

Another important topic when dealing with visual correspondence is the drift problem that refers to the phenomenon of leaving mismatched traces according to the object trajectory. Reference [13] argued that schedule sampling and bidirectional learning of cycle consistency [10] alleviate the drift problem. However, cycle consistency does not only reduce the flexibility of the model selection by using excessive memory, but also encounters overfitting of the motion information by recursively using the prediction result of the backbone [12]. Reference [14] alleviated the drift problem by not recursively using backbone prediction. However, the high-resolution representation cannot be dealt with due to the high cycle consistency cost. Solving the drift problem requires the proposed method to learn about the relationship between two density correspondences. In the performed test, a more diverse density correspondence is used to alleviate the drift problem.

The study contributions are as follows:

1) Pattern-boosted quantization is proposed. Using this module, the representation power is increased by clustering the consistent pattern information of the quantized image.
2) The feature representation of the embedding space is improved by using aggregation on the backbone.
3) The feature representation is modified using the appearance module to better understand the high-resolution appearance.
4) A high-resolution reconstruction is performed by proposing the fractal structure of the affinity matrix on the correspondence module. The drift problem is alleviated by learning the relationship between multi-density correspondences.
5) The visual correspondence performance on various datasets is evaluated. Video segmentation performance was evaluated on the DAVIS-2017 val dataset [34] using the $J\&F$ mean, yielding 65.4%

The remainder of this paper is organized as follows: Section II introduces the preliminary work; Section III provides a detailed description of the proposed method; Section IV evaluates the method performance by conducting experiments on various datasets. The contributions of each module are considered through an ablation study; and Section V briefly summarizes.

## II. PRELIMINARY WORK

This section explains the deep learning-based semisupervised video object segmentation methods. Note that the terms traditionally used in the video object segmentation (VOS) field are presented to prevent terminological confusion. In unsupervised video object segmentation, segmentation is performed without providing any annotation information from the user in the test stage. Conversely, in semisupervised video object segmentation, the region of interest (mask) of the first frame is given by the test stage user, and the subsequent frames are segmented by referring to the region. The VOS fields are classified depending on how the region of interest is given during the test stage. Meanwhile, general machine-learning methods are divided into supervised and unsupervised learning depending on the presence or absence of annotation for learning in the training stage. To avoid confusion in the terms used in the VOS field, unsupervised video object segmentation is referred to as automated video object segmentation or zero-shot video segmentation, and semisupervised video object segmentation is referred to as semiautomated object segmentation or one-shot video segmentation. We use the automatic video object segmentation or semiautomatic video object segmentation terminology in the paper as suggested by [35].

Visual correspondence is the problem of predicting where the specific point of the reference image is in the target image. It must respond to the changes in an object's shape and appearance over time; hence, the object's inherent characteristics must be understood. Accordingly, visual correspondence methods have often been introduced as a family of semiautomatic VOS. Their performance is measured using evaluation methods in the field. Before introducing the proposed paper, we will first demonstrate the necessity of the proposed method by explaining the existing semiautomatic VOS using supervised and unsupervised learning.

### A. SEMIAUTOMATIC VOS USING SUPERVISED LEARNING

The semiautomatic VOS using supervised learning is studied in fields that require a precise performance, such as long-term memory and pixel-level representation using annotation information during the training time. As such, it is generally divided into online fine-tuning-, propagation-, and matching-based methods.

#### 1) ONLINE FINE-TUNING-BASED METHODS

Online fine-tuning-based methods use transfer learning [36] and comprises two steps: 1) the process of learning general features at training time; and 2) the process of learning the target appearance at inference time [35], [37]. Each step shares the same backbone network and is gradually fine-tuned to understand the target object [38]. The backbone network is generally initialized with pre-trained parameters learned from ImageNet and learns objectness by fine-tuning it with object segmentation datasets on the train [37]. Some studies learned motion information [39] or instance-level semantic information [40] to boost the performance. Learning in the training step is called "offline learning," while that in the inference step is called "online learning." Online fine-tuning-based methods relatively reduce the burden of model overfitting due to their structural simplicity [35]. However, they encounter the problem of a long inference running time [41]. Various online fine-tuning-based methods [38], [39], [40] alleviate this problem by omitting the decoder or refining the module and performing resolution reconstruction by aggregation. This approach is deemed to allow an accurate contour expression in the inference time with a relatively few trainable parameters and limited data [38].

#### 2) PROPAGATION-BASED METHODS

Propagation-based methods predict the object of interest (mask) of the current frame by propagating the mask of the previous frames [42]. This method assumes that the target object has a high relationship with the object of interest in the previous frames [37]. However, video objects do not only have continuous changes, they can also be vulnerable to the occlusion, rapid motion, multiple instance, and drift problems [43], [44], [45], [46], [47]. Propagation-based methods have been developed to reinforce the insufficient spatio-temporal context for the object of interest [35]. These methods are classified into short-term temporal propagation-based methods that utilize additional information along with a single previous mask and long-term temporal propagation-based methods that use multiple previous masks to understand the context information [37]. The additional context information used in the former includes optical flow [44] and reinforcement learning [45] and is often fine-tuned through estimation to alleviate the error accumulation [46], [47], [48]. Long-term temporal propagation-based methods, such as bilateral neural network, generative adversarial network, and RNN are also being employed [49], [50], [51], [52], [53]. Propagation-based methods are easy to use with a decoder module because they do not require a time-consuming online learning process.

### 3) MATCHING-BASED METHODS

Matching-based methods predict a target object from the object of interest (mask) of the reference frames by measuring the similarity of the embedding space for the reference and target frames. This approach is classified into implicit and explicit matching [37]. Implicit matching [46], [48], [54] is a method of implicitly inferring the similarity between frames through fully connected or convolutional layers. In comparison, explicit matching [55], [56], [57], [58], [59] is the procedure of explicitly calculating the similarity between pixels through an affinity matrix.

The embedding space of a matching-based method implicitly learns the object's appearance information; thus, it has robust characteristics against occlusion or appearance change [60]. Reference [55] efficiently constructed a memory bank from multiple reference frames, making it easy to learn context for a long period. It has recently showed an excellent performance, even on long-time video datasets [59]. Matching-based methods have a fast inference time because they do not require online learning [55]. They also have the advantage of easily using a decoder module (i.e., refine module) for the high-resolution representation. With the above background, these methods have recently been receiving the greatest attention in the semiautomatic VOS field [35].

## B. SEMIAUTOMATIC VOS USING UNSUPERVISED LEARNING

Semiautomatic VOS using supervised learning requires a dataset of a pixel-wise annotation. Creating this dataset is time-consuming and labor-intensive and requires expert participation. Attempts of solving this problem have required the investigation of unsupervised-based methods that do not use any annotations at training time. The achievement of self-supervised learning with unsupervised learning is particularly notable. We will focus herein on the introduction of self-supervised learning methods. The semiautomatic VOS using self-supervised learning is divided into pseudolabeling- and nonpseudo-labeling based methods. The former comprises methods that learn user-designed definitions to extract useful features from unlabeled datasets, whereas the latter do not learn user-designed features from unlabeled data and solve the problem using the dataset's own characteristics. For example, a cycle consistency method learns the relationships between objects on the time axis [10], [14], and a contrastive learning method narrows the distance of positive pairs and increases that of negative pairs [61].

### 1) PSEUDOLABELING-BASED METHODS

Like the explicit matching-based methods, this type of method estimates the target mask by calculating the relationship between the reference and target frames using the affinity matrix. Instead of using labeled data at the training time, they are learned through pseudolabeling, which provides labels generated from the features of the input itself. Reference [11] created the $q_{t-1}$ pseudolabeling by selecting the ab channels in the Lab color space and quantizing it with k-means clustering. This method calculated the affinity matrix by applying the matrix inner product between $f_{t-1}$, which is the embedding space of the reference frame, and $f_t$, which is that of the target frame, and then normalizing it with softmax. The affinity matrix for the global area is represented by $A^*$.

$$\hat{q}_t = A^*(f_{t-1}, f_t) \cdot q_{t-1} \tag{1}$$

$$Loss_c = \min_{\theta} L(q_t, \hat{q}_t) \tag{2}$$

Ref. [13] proposed restricted attention to solve the low-resolution embedding space problem. The affinity matrix using restricted attention calculates only the similarity within a limited area from the sample of interest in the reference frame. This is based on the following characteristic: the objects between the adjacent frames smoothly move in the space and time axis. Similar to [12], [55] augmented the memory from multiple reference frames to improve one's understanding of objects over long-term periods, hence the robust characteristics against the drift and occlusion problems.

### 2) NONPSEUDO-LABELING-BASED METHODS

These methods perform visual correspondence using only the features between the reference and current frames. Widely used methods include cycle consistency and contrastive learning. Cycle consistency takes advantage of the fact that when an object of interest is predicted backward pass on the time axis, and then predicted forward pass, it returns to the initial area of interest. Reference [10] implemented cycle consistency for image patches. Reference [14] constructed cycle consistency using the affinity matrix to be multiplied through a backward pass, and then multiplied again through a forward pass, similar to the identity matrix. In contrastive learning, the distance for the positive pairs is narrowed down, while that for the negative pairs is increased. Reference [61] achieved a good performance by setting the relationship between the adjacent frames on the time axis to positive pairs. The advantage of the cycle consistency method is that it contains information about the time axis order. It can also exclude the human bias caused by pseudolabeling. A matrix multiplication operation between affinity matrices should be performed depending on the cycle consistency configuration, which limits the embedding space. Contrastive learning relies on semantic features to learn the relationships between adjacent frames as positive pairs. Hence, nonpseudo-labeling-based methods have a relatively low-resolution representation compared to pseudolabeling-based ones.

## III. PROPOSED METHOD

CorrFractal is a visual correspondence method that is based on self-supervised learning for high-resolution representation. It comprises the methods shown in Fig. 1. It is a pattern-boosted quantization method for creating a quantized image, including the object's general characteristics. The backbone module provides high-resolution features to the appearance and correspondence modules through aggregation. The appearance module learns the appearance information of the high-resolution features. The correspondence module learns the object motion by calculating the similarity for multiple resolutions and simultaneously performs a high-resolution reconstruction.

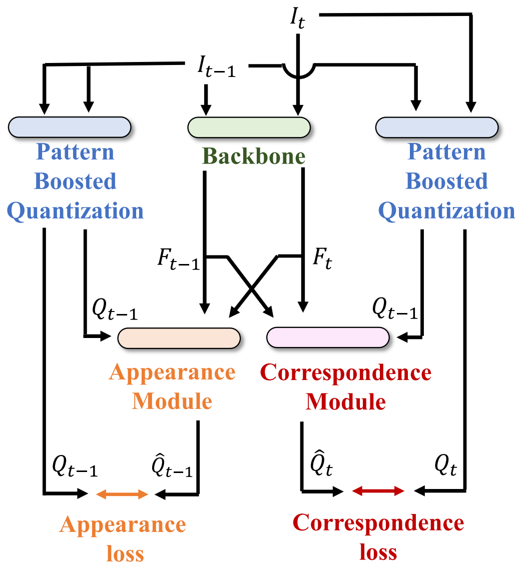


**FIGURE 1.** Flowchart of the proposed method. The proposed method consists of four modules. Each module outputs the original resolution equal to the resolution of the input frame. In the training stage, both appearance loss and correspondence loss are used to learn. However, in the inference stage, only the estimate value of the correspondence module is used.

Each module comprising CorrFractal is configured to output the original resolution. For example, the embedding spaces $F_{t-1} = \phi(I_{t-1};\theta)$ and $F_t = \phi(I_t;\theta)$ are calculated using the reference $I_{t-1}$ and current $I_t$ frames, respectively. For all notations in this paper, we will use the uppercase to represent the original resolution and the lowercase to denote the lower one.

### A. PATTERN-BOOSTED QUANTIZATION

The existing pseudolabeling-based method is superior over the existing non-pseudolabeling-based one in terms of the high-resolution representation; hence, the proposed method adopts the pseudolabeling-based approach. The existing pseudolabeling-based methods use color information as an important clue for understanding objects [11], [12], [13] because it is one of the important clues expressing the object's inherent characteristics. A natural environment, however, depicts cases of similar color information for the background and the object [11]. In particular, the number of sampling is limited when quantization is performed. This makes it more difficult to distinguish between the object and the background. According to neuroscience, the ventral pathway plays an important role in understanding objects. The important clues used by the ventral pathway also include color, contour, and pattern information [62], [63], [64], [65]. Based on neuroscience research results, pattern and color information are selected as clues for implementing pseudolabeling

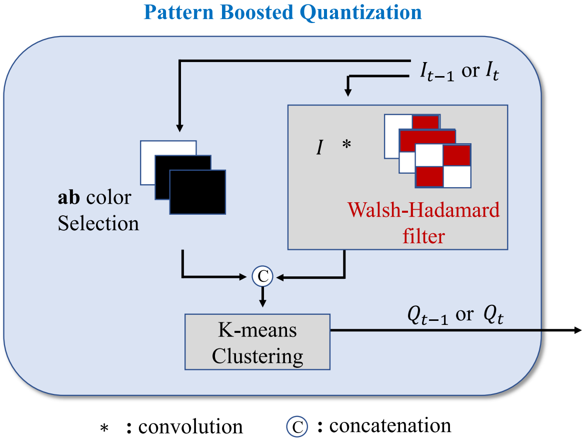


**FIGURE 2.** Diagram of the pattern boosted quantization. The resolution of pseudo labeling is the same as the resolution of the input frame. The Walsh–Hadamard filter is useful for boosting pattern features and boundary features.

Fig. 2 illustrates a flowchart of the pattern-boosted quantization. The color information was extracted by selecting only the ab channels in the Lab color space, as previously used [11], [13]. Meanwhile, the pattern information was extracted by creating it through the application of convolution using an RGB frame and a Walsh—Hadamard filter. Creating various mixed patterns from the color and pattern information required the random selection of channels, their concatenation, and the application of the k-means clustering to make $Q_{t-1}$ and $Q_t$, which are pattern-boosted quantization for the reference and target frames. Through the Walsh—Hadamard filter, pseudolabeling learns not only similar patterns inside the object, but also the object's contour information using high-frequency patterns

### B. NETWORK ARCHITECTURE

The semiautomatic VOS based on self-supervised learning has a relatively lower-resolution representation than the semi-automatic VOS based on supervised learning. Reference [12] presented feature representation based on stride 4. Other

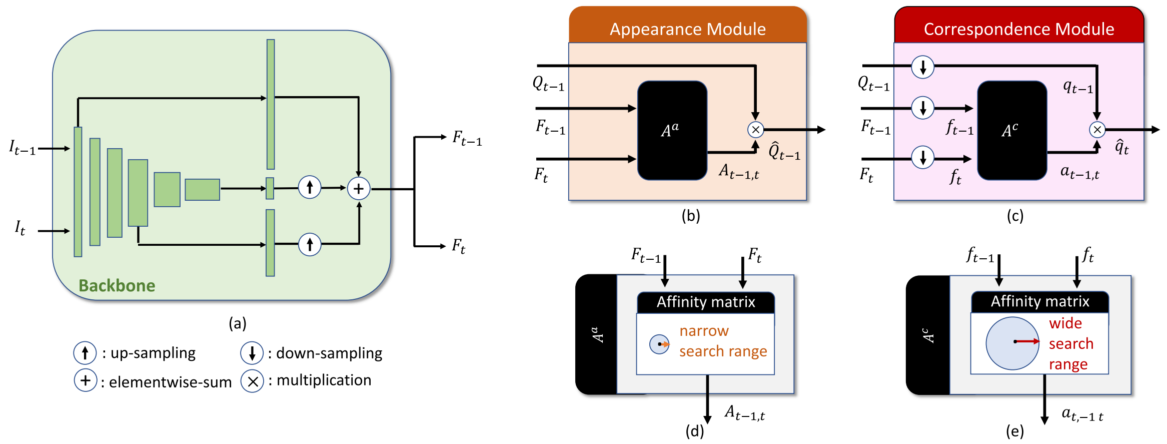


**FIGURE 3.** Overview of the network architecture and appearance and correspondence modules. (a) Backbone module uses aggregation to create high resolution features. (b) Appearance module directly calculates high-resolution features using a narrow search range, as depicted in (d). (c) Correspondence module cannot directly calculate high-resolution features because it uses a wide search range as depicted in (e).

methods have a lower-resolution feature representation. A decoder module (i.e., refine module) used in supervised learning is an easy approach for the high-resolution feature representation. However, self-supervised learning methods do not provide exemplary labeling; therefore, decoder modules directly affected by labeled data cannot be used.

In generative adversarial networks [66], methods are used to gradually increase the resolution from low-resolution images to generate high-resolution ones [67], [68]. These methods aim to stabilize high-resolution learning with the help of low-resolution images. Online fine-tuning-based methods use aggregation to solve problems with long inference times. These methods claim to express accurate contours with few trainable parameters and limited data.

The proposed method requires semantic features because it must express the motion between the reference and target frames. High-resolution features should also be maintained. Aggregation is a good approach for solving this problem. Fig. 3(a) shows the network architecture of the proposed method. The aggregation method depicted in the figure comprises a very simple method. Our own experiments for the proposed method found no significant performance difference in the complex forms of aggregation [29]; thus, it was constructed in as simple a form as possible. Elementwise sum was used to efficiently create high-resolution features from the necessary channels extracted from each stage [28]. $F_{t-1}$ and $F_t$ generated from the backbone were employed to extract the motion matching for objects and the information for the resolution reconstruction through a multiresolution subsampling in the subsequent module.

### C. APPEARANCE MODULE

The appearance module was developed to alleviate the motion overfitting by learning high-resolution appearance features. These features can be learned by calculating the similarity for a low-resolution embedding space like the matching-based methods (explicit matching), through a decoder module, or by the direct calculation of the similarity from a high-resolution embedding space. The proposed method chooses to directly calculate the similarity from the high-resolution embedding space. The problem here is that the computational complexity of the affinity matrix is calculated as $(f_w \times f_h)^\intercal \cdot (f_w \times f_h)$, where $f_w$ is the embedding space width, and $f_h$ is the embedding space height. This means that the memory consumption becomes more depleted as the resolution increases. To solve this problem, [13] suggested restricted attention that calculates only the similarity of the local region for the sample of interest.

The appearance module uses restricted attention to calculate $\hat{Q}_{t-1}$, which is the visual correspondence for the reference frame. Fig. 3(b) depicts the appearance module. Like a general affinity matrix, the affinity matrix $A^a$ uses $F_t$, which is the embedding space of the current frame, and $F_{t-1}$, which is the embedding space of the reference frame, as the input values. However, unlike the general affinity matrix, $A^a$ calculates the visual correspondence for the reference frame; hence, the window size, which is the local search range, should be very small for it to be more influenced by the reference features. In the proposed method, the window size is set to 3 × 3 to acquire the effect of regularization. Consequently, the appearance module obtains a very narrow search area, lowering the memory burden of calculating the original resolution features. The appearance module is used through $Q_{t-1}$, a pseudolabeling composed of the original resolution. The high-resolution appearance features can then be learned. The proposed method is used to improve the correspondence module performance by learning the appearance module as a sibling structure.

### D. CORRESPONDENCE MODULE

The correspondence module plays the most important role of predicting where the region of interest (mask) of the

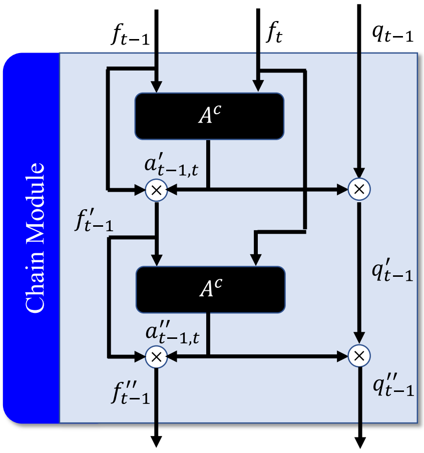


**FIGURE 4. Diagram of the chain module. The chain module has a chain structure by recursively receiving input values for the features and pseudo labeling of the reference frame. The features of the target frame act as an attractor.**

reference frame is located in the current frame. Fig. 3(c) briefly depicts the correspondence module. Unlike the affinity matrix of the appearance module, the affinity matrix, $A^c$, of the correspondence module cannot directly deal with the high-resolution feature map, $F_{t-1}$, $F_t$ and pseudo labeling, $Q_{t-1}$, because it basically requires a relatively wide search area by performing a motion-related feature matching. Existing methods calculate the affinity matrix on the low-resolution feature map to solve the search area problem. However, the low resolution-based method infers inaccurate boundary results by losing feature details, which consequently results in the accumulation of correspondence errors caused by incorrect pixel matching. The correspondence module solves this problem by first calculating the low-resolution predicted value $\hat{q}_t$ using low-resolution input and gradually reconstructing the high resolution. Fractal affinity and multi-density correspondence methods are used accordingly.

### 1) FRACTAL AFFINITY

Fractal affinity gradually calculates the high-resolution similarity by connecting matrices in a chain form. Achieving this requires the shape of each affinity input to be maintained and an attractor to exist. The relevant equation is as follows:

$$(f^n_{t-1}, q^n_{t-1}) = Chain(f^{n-1}_{t-1}, q^{n-1}_{t-1}, F^*_t) \tag{3}$$

Fractal affinity receives $f^{n-1}_{t-1}$ and $q^{n-1}_{t-1}$, $F^*_t$ as the input and outputs $f^n_{t-1}$ and $q^n_{t-1}$; $f^{n-1}_{t-1}$ is the reference frame of the embedding space for the $n-1th$ chain; $q^{n-1}_{t-1}$ is a pseudo labeling of the $n-1th$ chain; $F^*_t$ is the current frame of the embedding space resized according to different input sizes. Additionally; $f^n_{t-1}$ is the reference frame of the embedding space for the next chain; and $q^n_{t-1}$ is the pseudolabeling of the next chain.

The resolution is gradually increased by performing an upsampling after calculating each chain. At this time, the current frame, $F^*_t$, is directly received from the backbone and resized according to the reference frame embedding space. The last chain module deals with the original resolution of a pseudolabeling; thus, $F^*_t$ has the $F_t$ value, and $\hat{Q}_t$ infers the original resolution. The window size on restricted attention for a lower resolution has a constant value, while that on restricted attention for the original resolution has a $1\times 1$ value. The proposed method solves the memory problem in this manner using a small window at a high resolution and a relatively large window at a low resolution.

The fractal affinity structure widens the search area of the window size when performed at the same resolution. This assumes that the embedding space values are linearly distributed. The proposed method defines this structure as a chain module.

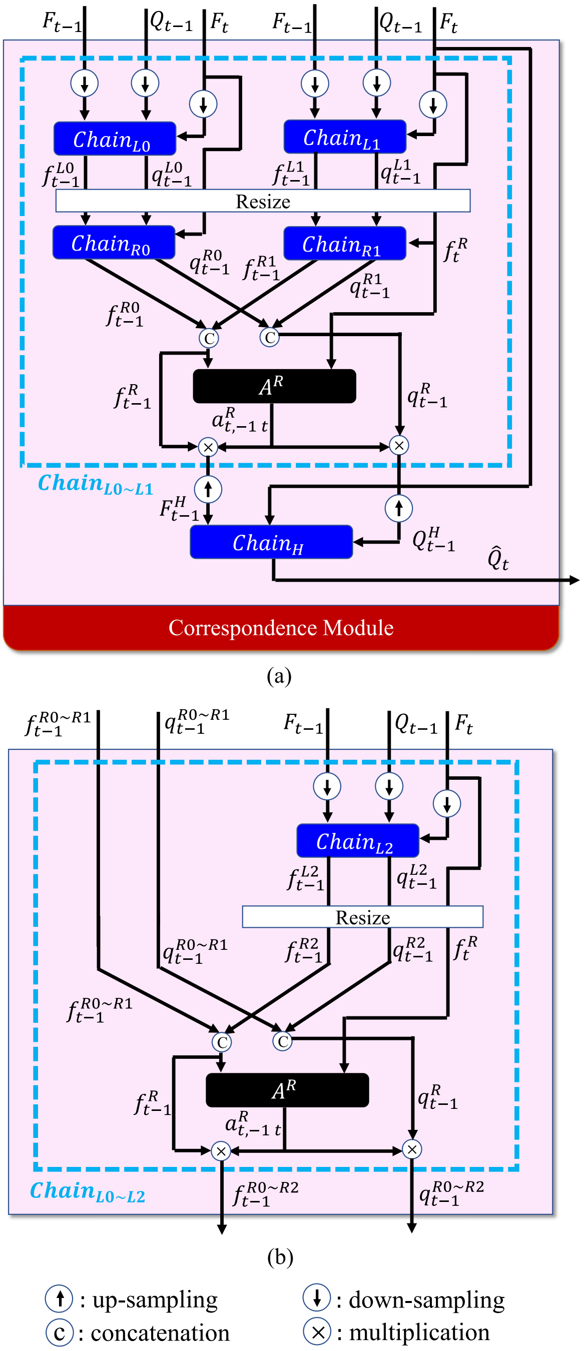
$$q'_{t-1} = A^c(f_{t-1}, f_t) \cdot q_{t-1} \tag{4}$$
$$f'_{t-1} = A^c(f_{t-1}, f_t) \cdot f_{t-1} \tag{5}$$
$$q''_{t-1} = A^c(f'_{t-1}, f_t) \cdot q'_{t-1} \tag{6}$$
$$f''_{t-1} = A^c(f'_{t-1}, f_t) \cdot f'_{t-1} \tag{7}$$

This configuration can reduce the memory consumption of the affinity matrix proportional to the square of the window's range, but implies the possibility of finding incorrect matches if it fails to search for similar features in the initial chain. Therefore, arbitrarily increasing the number of chains can actually reduce the performance. The proposed method suggests the appropriate number of chains through the experiments.

### 2) MULTIDENSITY CORRESPONDENCE

We have dealt with a method for the high-resolution reconstruction from low-resolution features through fractal affinity. However, fractal affinity basically performs feature matching based on the affinity matrix; thus, a reconstruction cannot be performed on areas where the features have been severely damaged by down-sampling from the backbone. In other cases, expressing the object motion becomes difficult when only the local similarity is matched considering the high-resolution features. An incorrect matching may be induced if the object of interest is outside the window radius. The proposed method alleviates this problem by using multiresolution features as the correspondence module input.

Fig. 5(a) depicts a diagram of the multi-density correspondence of the training process. $Chain_L$, $Chain_R$, and $Chain_H$ are composed of the fractal affinity. $F_t$ acts as an attractor. $Chain_L$ and $Chain_R$ have the same window sizes and gradually increase the resolution. $Chain_H$ has a window size of $1 \times 1$ and reconstructs the original resolution. During training, the proposed method calculates the similarity for a multiresolution split into two branches from the backbone as input for the correspondence module. $Chain_{L0}$ receives input that is downsampled to a low resolution, while $Chain_{L1}$ receives input that is downsampled to the middle resolution. At this time, $Chain_{L0}$ and $Chain_{L1}$ select a subsampling rate using the probability distribution to learn the relationship with various resolutions. The proposed method experimentally



(a)

(b)

**FIGURE 5. Diagram of the correspondence module. (a) Learning relationships for the training stage. (b) Learning relationships for the inference stage.**

selects two Gaussian distributions with means of $\mu_1$ and $\mu_2$ and a standard deviation of $\sigma$ for sub-sampling.

The resulting values for $Chain_{L0}$ and $Chain_{L1}$, $(f_{t-1}^{L0}, q_{t-1}^{L0})$ and $(f_{t-1}^{L1}, q_{t-1}^{L1})$ are resized to match the resolution. $Chain_{R0}$ and $Chain_{R1}$ are applied. $f_{t-1}^{R0}$ and $f_{t-1}^{R1}$ are then concatenated to make $f_{t-1}^{R0,R1}$. $q_{t-1}^{R0}$ and $q_{t-1}^{R1}$ are concatenated to make $q_{t-1}^{R0,R1}$. For simplicity, the proposed method represents $Chain_R$ as a single subsampling chain. However, if the image handled by $Chain_L$ is very small, a gradual upsampling must be performed using more chains. The denominator value of subsampling is usually smaller than the window search range when compared to the spatial resolution of the previous chain's image. $(f_{t-1}^R, q_{t-1}^R, f_t^R)$ calculates the affinity matrix through $A_R$, which has a 3 × 3 window size and applies the softmax to the attention mechanism from each resolution. In this way, we learn the relationship for the fractal affinity between $L0$ and $L1$ which have different spatial resolutions. We denote this structure as $Chain_{L0\sim L1}$.

$$\begin{aligned} Chain_{L0\sim L1} &= A^R(f_{t-1}^{R0,R1}, f_t^R) \cdot q_{t-1}^{R0,R1} \\ &= (f_{t-1}^{R0\sim R1}, q_{t-1}^{R0\sim R1}) \end{aligned} \tag{8}$$

A generalized visual correspondence for the spatial resolution can be learned by applying random sampling with different Gaussian distributions to $Chain_{L0}$ and $Chain_{L1}$. However, a problem exists when selecting a specific subsampling in the test. The simplest method selects the expected value of the Gaussian distribution as the representative subsampling for $Chain_{L0}$ and $Chain_{L1}$. This method means that consistent subsampling is performed, regardless of the object characteristics. Consequently, this allows the objects and the backgrounds to be contained within a single sample, which is one of the causes of the drift problem.

The proposed method solves this problem by simultaneously calculating the relationships with various spatial resolutions in the test. For example, if $Chain_{L0}$ selects a static sampling rate, $Chain_{L1}$ compensates for the mismatched part of $Chain_{L0}$ with an attention mechanism. If $Chain_{L0\sim L1}$ is assumed to be a single $Chain_L$ that possesses the characteristics of both $Chain_{L0}$ and $Chain_{L1}$, the relationship with $Chain_{L2}$ can also be calculated. In this case, the proposed method is expressed as $Chain_{L0\sim L2}$ (Fig. 5(b)). Additional chain modules are used to express the abundant relationships between the spatial resolutions in the test.

$$Chain_{L0\sim L2} = A^R(f_{t-1}^{R0\sim R1,R2}, f_t^R) \cdot q_{t-1}^{R0\sim R1,R2} \tag{9}$$

The loss function of the proposed method is defined as follows: $Loss_a$ is the appearance loss, while $Loss_c$ is the correspondence loss. $\alpha_1$ and $\alpha_2$ are each defined as 0.5, The final loss is determined by $Loss_a$ and $Loss_c$.

$$Loss_a = \min_{\theta} L(Q_{t-1}, \hat{Q}_{t-1}) \tag{10}$$

$$Loss_c = \min_{\theta} L(Q_t, \hat{Q}_t) \tag{11}$$

$$Loss = \alpha_1 Loss_a + \alpha_2 Loss_c \tag{12}$$

In the test, only $\hat{Q}_t$, which is the predicted correspondence value, is used following the method of [27]. Although the appearance module is not involved in the inference, it improves the visual correspondence performance by learning helpful information as a sibling structure, as in [69].

### E. IMPLEMENTATION PLATFORM

The proposed method uses ResNet-18 [70] as the backbone. Three branches of aggregation are noted (Fig. 3(a)). Each branch comprises the output of the first convolution layer and the second and fourth stages. The last stage has 1/4 spatial resolution of the input frame. Pooling layers exist in the second and fourth stages. The embedding space is created through an elementwise-sum of the three aggregation branches by upsampling. The first convolution layer has the same spatial resolution as the input image; thus, the embedding space also has the same spatial resolution. The window size for the affinity matrix of the appearance module is $3 \times 3$, while that for the affinity matrix of the correspondence module is $11 \times 11$.

#### 1) TRAINING

The proposed method performs scratch learning using the YouTube-VOS [71] and OxUvA datasets [72]. The model comprises a base model that uses the $I_{t-1}$ frame as the reference frame and a full model that employs the reference frame as $I_0, I_5, I_{t-1}, I_{t-3}$ and $I_{t-5}$. The OxUvA dataset is used to learn the base model, while the YouTube-VOS dataset is utilized to learn the full model. The image frame is resized to $256 \times 256 \times 3$. The proposed model is trained for five epochs using the Adam optimizer. The initial learning rate is $2e^{-4}$ and reduced by half at two, three, and four epochs. $Chain_L$ subsamples two values selected by the Gaussian distribution as the denominator and learns the relationship between the two spatial resolutions as $Chain_{L0\sim L1}$.

#### 2) INFERENCE

The proposed method considers the relationship between six resolutions as at test time for multi-density correspondence. Therefore, a subsampling of 1/2, 1/3, 1/4, 1/6, 1/12, and 1/18 is performed for each $Chain_L$ and estimated as $Chain_{L0\sim L5}$ to express the relationship between various spatial resolutions.

### F. DATASET DETAILS

We have utilized various datasets for diverse purposes. For training, we employed the OxUvA and YouTube-VOS datasets. Additionally, for evaluating different objectives, we utilized the DAVIS-2017 dataset, VIP dataset [73], and JHMDB dataset [74]. In the subsection, we will provide detailed explanations for each dataset, along with descriptions of their purposes and evaluation methods.

#### 1) OXUVA DATASET

OxUvA dataset is provided for evaluating single-object tracking algorithms, consisting of 366 sequences spanning 14 hours of video. The dataset involves target objects periodically disappearing and is characterized as a long-term, large-scale tracking dataset with an average duration of more than 2 minutes. The proposed method utilizes this dataset for training on the base model.

#### 2) YOUTUBE-VOS DATASET

YouTube-VOS dataset is provided for semiautomatic video object segmentation. The training set provides 3471 videos with dense object annotation, 65 categories, and 5945 unique object instances. The validation set and test set are provided by categorizing them into seen categories and unseen categories. The proposed method utilizes this dataset for training on the full model.

#### 3) DAVIS-2017 DATASET

DAVIS-2017 dataset is provided for automatic video object segmentation and semiautomatic video object segmentation. The train set and test set are provided in both 480p resolution and 1080p resolution. The DAVIS-2017 val dataset provides instance segmentation for 30 videos.The proposed method utilized this dataset to evaluate the performance of semiautomatic video object segmentation during the inference stage on the base model and full model. The evaluation method used $J$(mean) for measuring region similarity and $F$(Mean) for calculating object boundary similarity

#### 4) VIP DATASET

VIP dataset provides information on 19 parts for 50 videos to capture semantic part details. For performance evaluation, the dataset employs mIOU (mean Intersection over Union), which calculates the intersection area of predicted results and ground truth (GT) areas. The proposed method utilizes this dataset to assess the performance of part segmentation on full model.

#### 5) JHMDB DATASET

JHMDB dataset serves as an action recognition dataset, providing information on 15 keypoints related to humans across 268 videos. J-HMDB utilizes PCK@$\tau$ (Probability of Correct Key-point) to measure whether the ground truth (GT) exists within a bounding box of size $\tau$ proportional to human size, for performance evaluation. The proposed method employs this dataset for keypoint matching on full model.

## IV. EXPERIMENTS

We performed experiments on key point matching and part and object segmentations to evaluate the label propagation. If the label of the first frame is given at the test time, we conducted the experiments to examine how the high-resolution representation affects the performance when subsequent frames are propagated. An ablation study was also performed to confirm the module performance.

### A. OBJECT SEGMENTATION

We experiment on the DAVIS-2017 dataset at 480p resolution to evaluate the VOS performance. Table 1 shows the experiment results on the Youtube-VOS. A full model applying a memory bank was used in this experiment. Instead of fine-tuning the memory by dividing it into short- and long-term memories as in [12], we learned all terms at once. As shown in Table 1, the proposed method did not show the best performance under self-supervised learning. However, we focused on solving the high-resolution representation

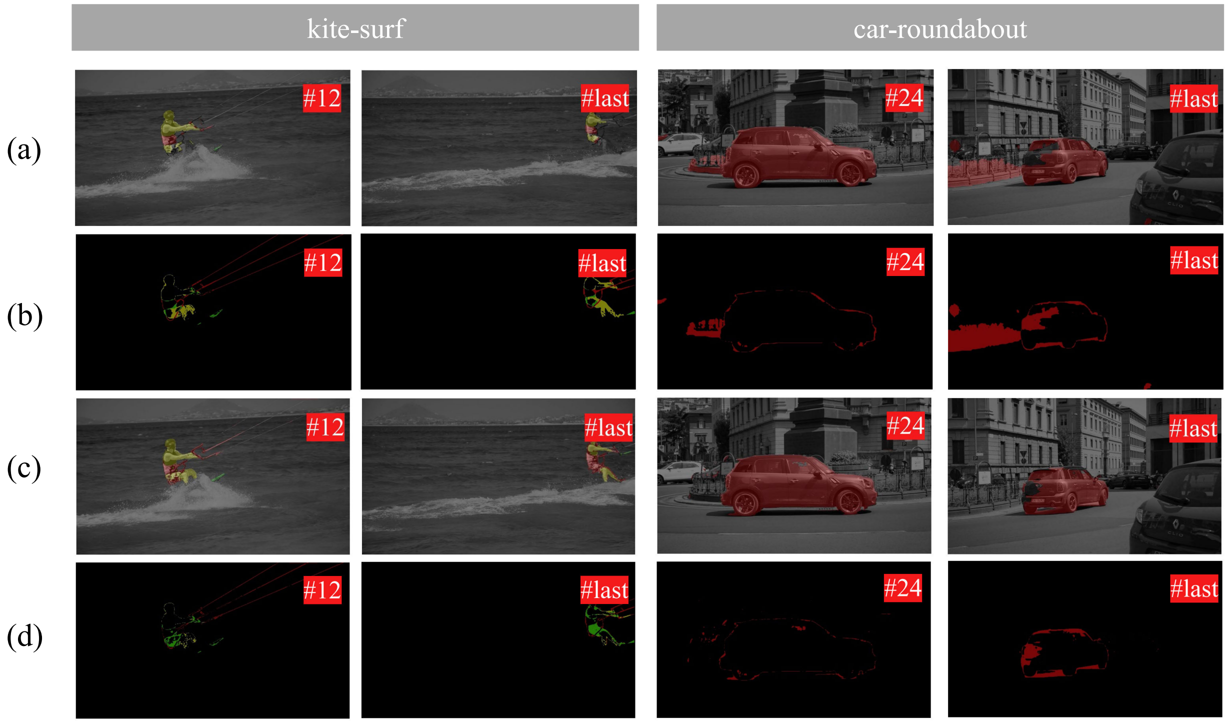


**FIGURE 6.** Comparison of the qualitative results on DAVIS 2017 dataset.

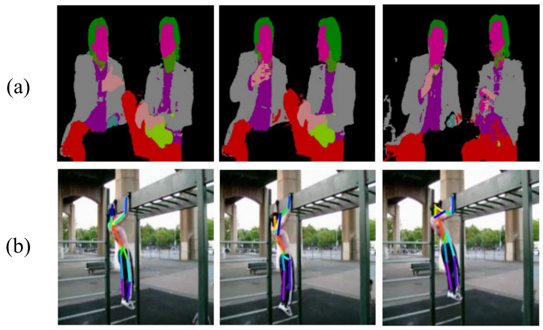


**FIGURE 7.** Qualitative results of the part segmentation and the key point matching.

in the self-supervised learning methods because it was a problem that needed to be solved. Fig. 6 compares the results with those obtained by [12], who solved the spatial resolution well among the existing methods applying self-supervised learning. The proposed method created subtraction images by performing an xor bit operation on the target mask and ground truth to evaluate the qualitative results for the high-resolution representation. Fig. 6(a) depicts images predicting the target mask presented in [12], while (b) illustrates the subtraction images depicted in the same reference. Fig. 6(c) shows images predicting the target mask of the proposed method, while (d) displays its subtraction images. The proposed method reduced the errors in the early video clips by precisely matching the object boundaries. Accordingly, the error accumulation caused by the boundaries was minimized. It had a representation, even for thin objects (e.g., rope). The proposed method alleviated the drift problem by learning the relationships for the multi-spatial resolution

### B. PART SEGMENTATION AND KEY POINT MATCHING

The proposed method used the VIP dataset to evaluate the part segmentation and the JHMDB dataset to evaluate the key point matching. Fig. 7(a) and Table 2 present the part segmentation results. The proposed method described the object's boundary relatively well, but the matching inside the object was relatively weak. Fig. 7(b) and Table 2 show the key point matching results. Similar to part segmentation, the performance tended to deteriorate due to the inability to estimate the object's internal area on the heat map.

### C. ABLATION STUDY

An ablation study was performed on a base model using only a single reference image on the OxUvA dataset. Table 3 shows the $J\&F$ mean results according to the module settings. $A$ is the appearance module. $P$ is the pattern-boosted quantization module. $B$ is the backbone module adopting aggregation. $C$ is the proposed correspondence module. The performance of the proposed method gradually improved with the addition of each module.

Fig. 8 compares the quantized image for the input frame using only the ab channels and the image with patterns added to the ab channels. Fig. 8(a) depicts the input frame; (b) is the pseudolabeling using only the ab color channels; and (c) is the pseudolabeling using the ab

**TABLE 1.** Quantitative results for the video object segmentation on the DAVIS 2017 dataset.

| Method | Supervised | Backbone | Dataset | $J\&F$ (Mean) | $J$ (Mean) | $F$ (Mean) |
|---|---|---|---|---|---|---|
| Video Colorization [11] | ✗ | ResNet-18 | Kinetics [75] | 34.0 | 34.6 | 32.7 |
| CycleTime [10] | ✗ | ResNet-50 | VLOG [76] | 48.7 | 46.4 | 50.0 |
| CorrFlow [13] | ✗ | ResNet-18 | OxUvA | 50.3 | 48.4 | 52.2 |
| MAST [12] | ✗ | ResNet-18 | Youtube-VOS | 65.5 | 63.3 | 67.6 |
| CRW [14] | ✗ | ResNet-18 | Kinetics | 68.3 | 65.5 | 71.0 |
| MAMP [77] | ✗ | ResNet-18 | Youtube-VOS | 69.7 | 68.3 | 71.2 |
| CLSC [78] | ✗ | ResNet-18 | Kinetics | 70.5 | 67.4 | 73.6 |
| LIIR [79] | ✗ | ResNet-18 | Youtube-VOS | 72.1 | 69.7 | 74.5 |
| LFGF [80] | ✗ | ResNet-18 | FlyingThings [81]+Youtube-VOS | 72.4 | 70.5 | 74.4 |
| **Ours** | ✗ | ResNet-18 | Youtube-VOS | 65.4 | 63.4 | 67.4 |
| OSVOS [38] | ✓ | VGG-16 | DAVIS$_{17}$ | 60.3 | 56.6 | 63.9 |
| STM [55] | ✓ | ResNet-50 | DAVIS$_{17}$+Youtube-VOS | 81.8 | 79.2 | 84.3 |

**TABLE 2.** Quantitative results for the part segmentation and the key point matching on the VIP and JHMDB dataset.

|  | VIP | JHMDB |  |
|---|---|---|---|
| Methods | mIOU | PCK@1 | PCK@2 |
| CycleTime [10] | 28.9 | 57.3 | 78.1 |
| CRW [14] | 38.6 | 59.3 | 80.3 |
| VFS [61] | 43.2 | 60.9 | 80.7 |
| SFC [82] | 38.4 | 61.9 | 83.0 |
| CLSC [78] | 40.8 | 61.7 | 82.6 |
| LIIR [79] | 41.2 | 60.7 | 81.5 |
| **Ours** | 36.9 | 59.0 | 79.8 |

**TABLE 3.** Quantitative results for the *J&F* (mean) according to the configuration of each module.

| Modules | $J\&F$ (Mean) | $J$ (Mean) | $F$ (Mean) |
|---|---|---|---|
| Base | 49.5 | 48.4 | 50.5 |
| Base+A | 51.3 | 51.0 | 51.6 |
| Base+A+P | 52.7 | 51.0 | 54.4 |
| Base+A+P+B | 54.5 | 52.2 | 56.8 |
| Base+A+P+B+C | 57.2 | 55.4 | 59.0 |

**TABLE 4.** Comparison of the quantitative results of the *J&F* (mean) according to the $\mu_1$ and $\mu_2$ settings of Gaussian distribution for the multi-density correspondence.

| $\mu_1$,$\mu_2$ | $J\&F$ (Mean) | $J$ (Mean) | $F$ (Mean) |
|---|---|---|---|
| 4, 6 | 54.7 | 52.9 | 56.6 |
| 4, 8 | 56.0 | 54.1 | 57.9 |
| 4, 10 | 53.4 | 51.6 | 55.1 |
| 4, 12 | 55.2 | 53.6 | 56.8 |

**TABLE 5.** Comparison of the quantitative results of the *J&F* (mean) according to the $\sigma$ settings of the Gaussian distribution for the multi-density correspondence.

| $\sigma$ | $J\&F$ (Mean) | $J$ (Mean) | $F$ (Mean) |
|---|---|---|---|
| 0.01 | 56.1 | 53.9 | 58.2 |
| 1.0 | 56.6 | 54.8 | 58.4 |
| 1.5 | 57.2 | 55.4 | 59.0 |
| 2.0 | 56.8 | 55.4 | 58.1 |
| 10 | 56.6 | 54.6 | 58.6 |

**TABLE 6.** Comparison of the quantitative results according to the number of affinity matrices contained in a single chain module.

| # of affinity matrices | $J\&F$ (Mean) | $J$ (Mean) | $F$ (Mean) |
|---|---|---|---|
| 1 | 56.0 | 53.2 | 58.8 |
| 2 | 57.2 | 55.4 | 59.0 |
| 3 | 54.8 | 52.7 | 56.9 |
| 4 | 54.7 | 52.7 | 56.7 |

color channels and the Walsh—Hadamard filter. For an equal comparison, a 16-channel quantization was used for both methods. The experiments showed that the method with added patterns generally represented the objects better. The pattern-boosted quantization module exhibited a high tendency of grouping similar textures within objects into a single clustering. By contrast, the method that used color only resulted in a relatively larger over-segmentation. In other words, in a situation where clustering channels are limited, the region that must be segmented cannot be properly segmented because it is already consuming resources due to incorrect segmentation. The pattern-boosted quantization module provided additional clues, even when the background and the object were similar by expressing the object's boundary

We conducted two experiments to obtain the hyperparameters for the subsampling in the multi-density correspondence. The first experiment focused on the $J\&F$ mean depending on

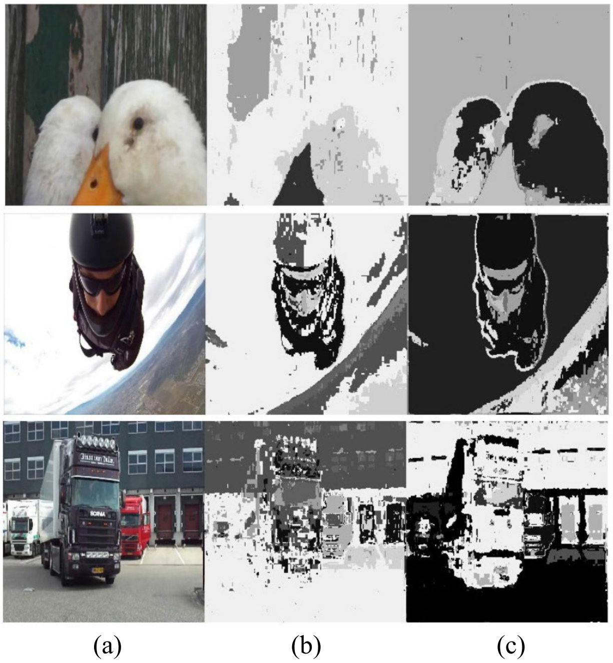


**FIGURE 8.** Comparison of pseudolabeling results.

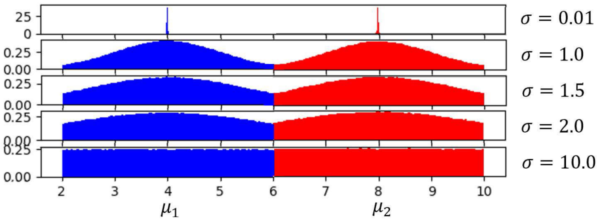


**FIGURE 9.** Comparison of the Gaussian distributions according to standard deviation for multi-density correspondence.

the denominator values, $\mu_1$ and $\mu_2$, for the static subsampling (Table 4). The second experiment aimed to calculate the $J\&F$ mean according to the standard deviation of the Gaussian distribution with $\mu_1$ and $\mu_2$ (Table 5). In the first experiment, the highest $J\&F$ mean was obtained when $\mu_1$ was 4, and $\mu_2$ was 8. Thus, in Table 5, $\mu_1$ was set to 4 and $\mu_2$ was set to 8. Each Gaussian distribution was truncated (Fig.9) to avoid the distribution being concentrated between $\mu_1$ and $\mu_2$.

In our experiment, it was difficult to find a general regularity for $\mu_1$ and $\mu_2$ in Table 4, which we analyzed the problem related to motion matching and representation. When low-resolution subsampling was performed by $\mu_2$, the object motion was expressed well, but the object details were damaged. When relative high-resolution subsampling was performed by $\mu_2$, estimating the object motion can easily be a failure, but the object detail was expressed well. As a result, static subsampling found it difficult to adapt in complex environments. Table 5 shows that the value of the $J\&F$ mean is higher when the standard deviation value is above a certain level than when it is very small. This means that learning the relationships between various resolutions helps improve the performance.

Table 6 measures the $J\&F$ mean depending on how many affinity matrices were composed within the chain module. The experiments showed that the best performance was achieved when there were two affinity matrices.

## V. CONCLUSION

The proposed method in this work represents a visual correspondence method for the high-resolution representation on self-supervised learning without decoder module. For this purpose, the proposed method presented four modules. The pattern-boosted quantization module created efficient pseudo labeling by learning the pattern information along with the color information. The pseudo labeling made by the pattern-boosted quantization not only allocated similar patterns inside an object to a single area, but also sensitively responded to the contour and provided clues to distinguish the background. The backbone module used aggregation to make high-resolution features while preserving the semantic features. The appearance module learned high resolution-based appearance using the embedding space generated from the backbone network. This module utilized a small search range on the affinity matrix for memory efficiency. The correspondence module performed motion matching while gradually increasing the spatial resolution. Although the embedding space of the reference frame employed low-resolution information, the embedding space of the current frame received high-resolution features directly from the backbone network and played the role of an attractor. The correspondence module also alleviated the drift problem by learning the relationships between various resolutions. Each module cooperated with each other to achieve a high-resolution representation. As a result, the proposed method demonstrated a good performance for the high-resolution representation by performing visual correspondence on object boundaries and thin objects.

This paper was developed for high-resolution representation in self-supervised learning without a decoder module. However, the following issues have been identified: Firstly, overfitting occurs at the boundaries of objects. Secondly, there is a high memory consumption of multiple density correspondence during the training stage, resulting in slow learning. Thirdly, there is a very high memory consumption of multiple density correspondence during the inference stage, leading to slow execution. To address the first issue, we are currently researching setting the resolution of all chain modules to be the same during the training stage, aiming to alleviate overfitting to high-resolution features. To tackle the second and third issues, we are exploring additional mapping methods.

## REFERENCES

[1] M. Cho, J. Lee, and K. M. Lee, "Feature correspondence and deformable object matching via agglomerative correspondence clustering," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 1280–1287.

[2] M. Bansal and K. Daniilidis, "Joint spectral correspondence for disparate image matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2802–2809.

[3] Y. Xu, B. Wang, W. Liu, and X. Bai, "Skeleton graph matching based on critical points using path similarity," in *Proc. ACCV*, 2009, pp. 456–465.

[4] X. Bai and L. J. Latecki, "Path similarity skeleton graph matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 7, pp. 1282–1292, Jul. 2008.

[5] W. Hu, M. Hu, X. Zhou, T. Tan, J. Lou, and S. Maybank, "Principal axis-based correspondence between multiple cameras for people tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 663–671, Apr. 2006.

[6] I. J. Cox and M. L. Miller, "On finding ranked assignments with application to multitarget tracking and motion correspondence," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 31, no. 1, pp. 486–489, Jan. 1995.

[7] F. Papenmeier, H. S. Meyerhoff, G. Jahn, and M. Huff, "Tracking by location and features: Object correspondence across spatiotemporal discontinuities during multiple object tracking," *J. Experim. Psychol., Hum. Perception Perform.*, vol. 40, no. 1, pp. 159–171, Feb. 2014.

[8] D. N. Bhat and S. K. Nayar, "Ordinal measures for image correspondence," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 4, pp. 415–423, Apr. 1998.

[9] J. Lee, E. Kim, Y. Lee, D. Kim, J. Chang, and J. Choo, "Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5800–5809.

[10] X. Wang, A. Jabri, and A. A. Efros, "Learning correspondence from the cycle-consistency of time," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2561–2571.

[11] C. Vondrick, A. Shrivastava, S. Guadarrama, and K. Murphy, "Tracking emerges by colorizing videos," in *Proc. ECCV*, 2018, pp. 391–408.

[12] Z. Lai, E. Lu, and W. Xie, "MAST: A memory-augmented self-supervised tracker," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6478–6487.

[13] Z. Lai and W. Xie, "Self-supervised learning for video correspondence flow," in *Proc. BMVC*, 2019.

[14] A. Jabri, A. Owens, and A. A. Efros, "Space-time correspondence as a contrastive random walk," 2020, *arXiv:2006.14613*.

[15] X. Li, S. Liu, S. De Mello, X. Wang, J. Kautz, and M.-H. Yang, "Joint-task self-supervised learning for temporal correspondence," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019, pp. 317–327.

[16] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "ICNet for real-time semantic segmentation on high-resolution images," 2017, *arXiv:1704.08545*.

[17] C. Jin, R. Tanno, T. Mertzanidou, E. Panagiotaki, and D. C. Alexander, "Learning to downsample for segmentation of ultra-high resolution images," 2021, *arXiv:2109.11071*.

[18] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5686–5696.

[19] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," 2019, *arXiv:1908.07919*.

[20] G. Zhang, Y. Ge, Z. Dong, H. Wang, Y. Zheng, and S. Chen, "Deep high-resolution representation learning for cross-resolution person re-identification," *IEEE Trans. Image Process.*, vol. 30, pp. 8913–8925, 2021.

[21] H. Yu, N. Xu, Z. Huang, Y. Zhou, and H. Shi, "High-resolution deep image matting," 2020, *arXiv:2009.06613*.

[22] B. J. Fino and V. R. Algazi, "Unified matrix treatment of the fast Walsh–Hadamard transform," *IEEE Trans. Comput.*, vols. C–25, no. 11, pp. 1142–1146, Nov. 1976.

[23] T. Zhou, S. Wang, Y. Zhou, Y. Yao, J. Li, and L. Shao, "Motion-attentive transition for zero-shot video object segmentation," in *Proc. AAAI*, 2020, pp. 13066–13073.

[24] W. Li, J. Mu, and G. Liu, "Multiple object tracking with motion and appearance cues," 2017, *arXiv:1909.00318*.

[25] S. Hyun, J. Kim, and J.-P. Heo, "Self-supervised video GANs: Learning for appearance consistency and motion coherency," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10821–10830.

[26] L. L. Cloutman, "Interaction between dorsal and ventral processing streams: Where, when and how?" *Brain Lang.*, vol. 127, no. 2, pp. 251–263, Nov. 2013.

[27] J.-m. Choi, J. Son, and S. Park, "Learning video correspondence using appearance module for target tracking," in *Proc. IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Jan. 2021, pp. 287–290.

[28] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," 2016, *arXiv:1612.03144*.

[29] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," 2017, *arXiv:1707.06484*.

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," 2015, *arXiv:1505.04597*.

[31] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2016, *arXiv:1608.06993*.

[32] L. Zhu, R. Deng, M. Maire, Z. Deng, G. Mori, and P. Tan, "Sparsely aggregated convolutional networks," 2018, *arXiv:1801.05895*.

[33] G. Larsson, M. Maire, and G. Shakhnarovich, "FractalNet: Ultra-deep neural networks without residuals," 2016, *arXiv:1605.07648*.

[34] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool, "The 2017 Davis challenge on video object segmentation," 2017, *arXiv:1704.00675*.

[35] T. Zhou, F. Porikli, D. Crandall, L. Van Gool, and W. Wang, "A survey on deep learning technique for video segmentation," 2021, *arXiv:2107.01153*.

[36] K. He, R. Girshick, and P. Dollár, "Rethinking ImageNet pre-training," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4918–4927.

[37] M. Gao, F. Zheng, J. J. Q. Yu, C. Shan, G. Ding, and J. Han, "Deep learning for video object segmentation: A review," *Artif. Intell. Rev.*, vol. 56, no. 1, pp. 457–531, Jan. 2023.

[38] S. Caelles, "One-shot video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jan. 2017, pp. 221–230.

[39] H. Xiao, J. Feng, G. Lin, Y. Liu, and M. Zhang, "MoNet: Deep motion exploitation for video object segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1140–1148.

[40] K.-K. Maninis, S. Caelles, Y. Chen, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool, "Video object segmentation without temporal information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 6, pp. 1515–1530, Jun. 2019.

[41] H. Xiao, B. Kang, Y. Liu, M. Zhang, and J. Feng, "Online meta adaptation for fast video object segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1205–1217, May 2020.

[42] F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, and A. Sorkine-Hornung, "Learning video object segmentation from static images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3491–3500.

[43] X. Li and C. C. Loy, "Video object segmentation with joint re-identification and attention-aware mask propagation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 90–105.

[44] P. Hu, G. Wang, X. Kong, J. Kuen, and Y.-P. Tan, "Motion-guided cascaded refinement network for video object segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1400–1409.

[45] J. Han, L. Yang, D. Zhang, X. Chang, and X. Liang, "Reinforcement cutting-agent learning for video object segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9080–9089.

[46] S. W. Oh, J.-Y. Lee, K. Sunkavalli, and S. J. Kim, "Fast video object segmentation by reference-guided mask propagation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7376–7385.

[47] L. Zhang, Z. Lin, J. Zhang, H. Lu, and Y. He, "Fast video object segmentation via dynamic targeting network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5581–5590.

[48] H. Lin, X. Qi, and J. Jia, "AGSS-VOS: Attention guided single-shot video object segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3948–3956.

[49] V. Jampani, R. Gadde, and P. V. Gehler, "Video propagation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3154–3164.

[50] K. Xu, L. Wen, G. Li, L. Bo, and Q. Huang, "Spatiotemporal CNN for video object segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1379–1388.

[51] P. Tokmakov, K. Alahari, and C. Schmid, "Learning video object segmentation with visual memory," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4481–4490.

[52] N. Xu, L. Yang, Y. Fan, J. Yang, D. Yue, Y. Liang, B. Price, S. Cohen, and T. Huang, "YouTube-VOS: Sequence-to-sequence video object segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 585–601.

[53] C. Ventura, M. Bellver, A. Girbau, A. Salvador, F. Marques, and X. Giro-i-Nieto, "RVOS: End-to-end recurrent network for video object segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5277–5286.

[54] J. S. Yoon, F. Rameau, J. Kim, S. Lee, S. Shin, and I. S. Kweon, "Pixel-level matching for video object segmentation using convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2186–2195.

[55] S. W. Oh, J.-Y. Lee, N. Xu, and S. J. Kim, "Video object segmentation using space-time memory networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9226–9235.

[56] H. K. Cheng, Y.-W. Tai, and C.-K. Tang, "Rethinking space-time networks with improved memory coverage for efficient video object segmentation," in *Proc. NeurIPS*, 2021, pp. 11781–11794.

[57] X. Lu, W. Wang, M. Danelljan, T. Zhou, J. Shen, and L. Van Gool, "Video object segmentation with episodic graph memory networks," in *Proc. Europeon Conf. Comput. Vis.*, Aug. 2020, pp. 661–679.

[58] Y. Liang, X. Li, N. Jafari, and J. Chen, "Video object segmentation with adaptive feature bank and uncertain-region refinement," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 3430–3441.

[59] H. K. Cheng and A. G. Schwing, "XMem: Long-term video object segmentation with an atkinson-shiffrin memory model," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 640–658.

[60] J. Cheng, Y.-H. Tsai, W.-C. Hung, S. Wang, and M.-H. Yang, "Fast and accurate online video object segmentation via tracking parts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7415–7424.

[61] J. Xu and X. Wang, "Rethinking self-supervised correspondence learning: A video frame-level similarity perspective," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, May 2021, pp. 10075–10085.

[62] G. Loffler, "Perception of contours and shapes: Low and intermediate stage mechanisms," *Vis. Res.*, vol. 48, no. 20, pp. 2106–2127, Sep. 2008.

[63] C. E. Connor, S. L. Brincat, and A. Pasupathy, "Transformation of shape information in the ventral pathway," *Current Opinion Neurobiol.*, vol. 17, no. 2, pp. 140–147, Apr. 2007.

[64] C. F. Altmann, H. H. Bülthoff, and Z. Kourtzi, "Perceptual organization of local elements into global shapes in the human visual cortex," *Current Biol.*, vol. 13, no. 4, pp. 342–349, Feb. 2003.

[65] J. Taylor and Y. Xu, "Representation of color, form, and their conjunction across the human ventral visual pathway," *NeuroImage*, vol. 251, May 2022, Art. no. 118941.

[66] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014.

[67] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," 2017, *arXiv:1710.10196*.

[68] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Sep. 2019, pp. 4401–4410.

[69] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.

[70] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition. in computer vision and pattern recognition," in *Proc. CVPR*, 2016.

[71] N. Xu, L. Yang, Y. Fan, D. Yue, Y. Liang, J. Yang, and T. Huang, "YouTube-VOS: A large-scale video object segmentation benchmark," 2018, *arXiv:1809.03327*.

[72] J. Valmadre, L. Bertinetto, J. F. Henriques, R. Tao, A. Vedaldi, A. Smeulders, P. H. S. Torr, and E. Gavves, "Long-term tracking in the wild: A benchmark," in *Proc. ECCV*, 2018, pp. 670–685.

[73] Q. Zhou, X. Liang, K. Gong, and L. Lin, "Adaptive temporal encoding network for video instance-level human parsing," 2018, *arXiv:1808.00661*.

[74] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3192–3199.

[75] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017, *arXiv:1705.06950*.

[76] D. F. Fouhey, W.-c. Kuo, A. A. Efros, and J. Malik, "From lifestyle vlogs to everyday interactions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4991–5000.

[77] B. Miao, M. Bennamoun, Y. Gao, and A. Mian, "Self-supervised video object segmentation by motion-aware mask propagation," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2022, pp. 1–6.

[78] J. Son, "Contrastive learning for space-time correspondence via self-cycle consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2022, pp. 14679–14688.

[79] L. Li, T. Zhou, W. Wang, L. Yang, J. Li, and Y. Yang, "Locality-aware inter-and intra-video reconstruction for self-supervised correspondence learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 8719–8730.

[80] R. Li, S. Zhou, and D. Liu, "Learning fine-grained features for pixel-wise video correspondences," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 9632–9641.

[81] N. Mayer, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 4040–4048.

[82] Y. Hu, R. Wang, K. Zhang, and Y. Gao, "Semantic-aware fine-grained correspondence," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 97–115.

**JIN-MO CHOI** received the B.S. and M.S. degrees in computer science from Hanyang University, Republic of Korea, in 2003 and 2005, respectively. He is currently pursuing the Ph.D. degree with the Electronics and Telecommunications Research Institute (ETRI), University of Science and Technology, Republic of Korea. His research interests include computer vision, image classification, object detection, character generation, face recognition, and segmentation.

**BLAGOVEST I. VLADIMIROV** received the B.S. degree in industrial automation from the Technical University of Sofia, Bulgaria, and the M.E. and Ph.D. degrees in systems engineering from the Nagoya Institute of Technology, Japan, in 2008. Since 2010, he has been a Researcher with the Electronics and Telecommunications Research Institute (ETRI), South Korea. His research interests include machine learning, cognitive systems, and their practical applications.

**SANGJOON PARK** received the B.S. and M.S. degrees in electronics engineering from Kyungpook National University, in 1988, and 1990, respectively, and the Ph.D. degree from the Department of Computer Science, North Carolina State University, in 2006. From 1990 to 2001, he was a Senior Researcher with the Agency for Defense Development (ADD). He is currently a Principal Researcher with the Defense ICT Convergence Research Section, Electronics and Telecommunications Research Institute (ETRI), South Korea. He is also an Adjunct Professor with the University of Science and Technology, South Korea. His current research interests include metaverse training systems, indoor positioning, and human augmentation technologies.