https://doi.org/10.26599/AIR.2025.9150054 CAAI Artificial Intelligence Research

A survey on deep learning techniques for image and video feature aggregation

Md Majedul Islam[1] ✉, Sai Prakash Reddy Rao[1], and Selena He[1] ✉

ABSTRACT

Nowadays, deep learning has demonstrated impressive performance in the area of computer vision and pattern recognition, such as objects recognition, videos classification and image segmentation. In particular, convolutional neural networks (CNNs) an advanced deep-learning technique have achieved strong performance in image and video analysis owing to their powerful feature-extraction capabilities.Based on that, current research also make a breakthrough in image and video recognition via aggregating features extracted from various layers of CNNs. Over the last decade, feature fusion strategies have developed from conventional schemes restricted to conditions we need to control over advanced ones which can be achieved with a large number of images or videos. However, due to the rapid development in this field, it is challenging to track and systematically analyze recent advancements. This has inspired us to offer a comprehensive survey of the significant steps taken towards feature aggregation strategies. To better organize and facilitate understanding, we first introduce preliminary knowledge on deep learning. Then this paper focuses on categorizing and reviewing the current strategies from two main aspects: feature aggregation on images and feature aggregation on videos. We further highlight the comparative strengths and limitations among these strategies. Finally, we point out the challenges in this area and motivate further work via proposing future directions on feature aggregation techniques for both images and videos.

KEYWORDS

feature aggregation; feature extraction; deep learning; survey

In recent years, deep learning has achieved great performance in a variety of tasks, such as image classification, object detection, and natural language processing (NLP)[1]. To the best of our knowledge, the core of deep learning technology is that the features extracted from deep learning models are not designed by human beings, but learned from data using a general-purpose procedure.

As we know, one of its branches, convolutional neural networks (CNNs), has shown outstanding performance in image and video recognition. CNNs are composed of convolutional layers and fully connected layers. Recent works have demonstrated that CNNs are able to extract discriminative features at each layer to effectively represent input data. The extracted features possess very useful properties. For instance, they can be extracted directly and efficiently from images which have different sizes or shapes. In addition to that, layer-wise accuracy comparisons[2–5], transferability analysis[6], and representation visualization[4, 7] shows that the low-level layers (usually the first three layers) in CNNs extract features that contain more fine-grained details of objects. However, higher-level layers tend to capture more semantic and abstract features, often beyond human interpretation. Currently, CNN-based image and video analysis have been receiving increasing attention from the community.

However, due to the quality and quantity of images and videos, features learned from each layer are not good enough to promote our model achieving ideal performance in objects computing. For example, as the layers get deeper, the extracted features will lose some important information playing a crucial role, and they may also include significant noise, which can negatively affect the model's performance. On the basis of that, some researchers proposed that features extracted from different layers have mutual support relationships with each other, so we can aggregate features of different layers using various feature fusion strategies, then it is possible to improve performance of our models and help reduce noise simultaneously using the fused features.

The concept of feature fusion is not new to the deep learning era. Foundational work in pattern recognition established key strategies for combining feature sets to improve classification performance. For instance, Yang et al.[8] introduced a distinction between serial feature fusion, the common strategy of concatenating feature vectors, and a novel parallel feature fusion strategy, which used complex vectors to combine features before applying traditional extraction methods like principal component analysis (PCA) and linear discriminant analysis (LDA).While these classical approaches laid the groundwork, they relied on handcrafted or linear projection-based features. The advent of deep learning, particularly CNNs, revolutionized the field by enabling the end-to-end learning of hierarchical features. This survey focuses specifically on the feature aggregation techniques that have been developed within these deep learning architectures, which differ significantly in mechanism and scale from their classical predecessors.

Over the last decade, feature fusion has evolved from earlier stages which are limited to controlled environments to nowadays advanced approaches that can be achieved from large amounts of images to video surveillance. Scientific milestones in feature aggregation are achieved very fast, finally leading to the demise of what used to be better. This rapid development motivates us to

1 Department of Computer Science, Kennesaw State University, Marietta 30060, GA, USA
Address correspondence to Md Majedul Islam, mislam20@students.kennesaw.edu; Selena He, she4@kennesaw.edu

provide a comprehensive survey of the current strategies in features integration. In this survey article, our contributions can be summarized as follows:

- We conduct a systematic review of feature aggregation methods on images and videos based on deep learning techniques and summarize the challenges faced in each setting.
- We provide a detailed analysis of the contributions and novelties of strategies related to feature aggregation conducted on images and videos. Besides, we also compare the similarities and differences among various approaches systematically.
- We suggest four promising future research directions in the area of feature aggregation via analyzing the limitations of existing works, like data diversity, noise reduction, features interpretability and so on. For each direction, we provide a thorough analysis of its deficiency in current work and propose solutions.

Our study covers two areas where feature fusion methods play important role, including image analysis and video computation. Rather than focusing only on their strategies, we compare the differences among them and explore a variety of applications which achieve their goals broadly using feature aggregation, thereby bring new perspective visions. We hope that this survey can serve as a comprehensive reference resource and inspire future researchers in understanding feature fusion and its many potential applications.

The rest of paper is organized as follows, we firstly introduce preliminary knowledge of deep neural networks, especially related to convolutional neural networks in Section 1. Then, in Section 2, we focus on discussing the feature fusion strategies proposed in the fields of image recognition and highlighting the most concerned design components. After that, we explore feature fusion approaches applied in the areas of video detection in Section 3. In Section 4, we discuss the similarities and differences of feature aggregation methods between image and video analysis in detail. After that, we point out some challenges in the areas of feature fusion and propose possible research directions in Section 5. Finally, section 6 concludes the article.

1 Preliminary knowledge

When we examine how neural networks process visual information, it becomes clear that different architectural components serve distinct purposes in feature aggregation. Instead of covering general deep learning concepts, will focus on understanding how these components specifically enable the aggregation strategies discussed throughout this survey.

1.1 How network components enable feature aggregation

Convolutional layers and local feature building: convolutional operations create a natural locality bias - they examine small neighborhoods of pixels and build up understanding piece by piece. This localized processing generates stable descriptors at various scales, which becomes the foundation for aggregation methods like NetVLAD and VLAD encoding[9]. The beauty lies in how early layers focus on details like edges and textures, while deeper layers capture broader semantic meaning. This creates a rich hierarchy where each level offers something different for aggregation strategies to combine, as shown in Fig. 1.

The pooling dilemma and its solutions: traditional pooling faces an interesting challenge: reducing information while preserving what matters most. Max pooling keeps the strongest signals but loses spatial nuance. Average pooling smooths out noise but may dilute important details. This fundamental tension has sparked innovations in learned pooling approaches, where networks decide what to keep based on the specific task. Attention-weighted pooling and geometry-aware methods represent sophisticated attempts to solve this preservation versus reduction problem[10].

Attention as a unifying framework: attention mechanisms provide something powerful - they let networks decide what is important based on the actual data being processed. Rather than treating all features equally, attention creates data-dependent weights that can work across different domains. Whether we are selecting important spatial regions in images or crucial frames in videos, attention offers a flexible way to guide aggregation processes intelligently[11].

Self-supervised learning's contribution: modern self-supervised approaches have revealed something interesting about feature quality. When networks learn through tasks like predicting masked portions or contrasting similar examples, they develop representations that naturally align well across different layers and time points. These features seem inherently more suitable for aggregation because they have learned to be consistent and meaningful without explicit supervision[12].

Extending to video analysis: video processing builds on these spatial concepts while adding temporal complexity. Motion modeling through optical flow helps align features across frames, addressing the challenge that the same object might appear differently over time. Temporal attention mechanisms help networks focus on high-quality frames while managing redundancy. Memory-based approaches maintain context across longer sequences, ensuring that aggregation considers both immediate and distant temporal relationships.

Why this matters: understanding these component roles helps explain why certain aggregation strategies work well in specific contexts. The hierarchical nature of convolutional features makes multi-layer fusion effective. The information-preservation challenges in pooling drive attention-based solutions. The temporal dependencies in videos require specialized approaches that standard image methods can not handle alone.

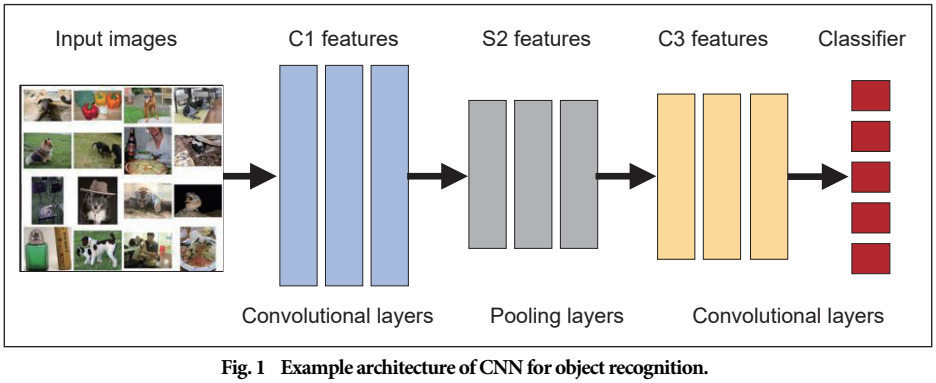

Fig. 1 Example architecture of CNN for object recognition.

2 Feature aggregation on images

In this section, we present image feature aggregation methods grouped into distinct families based on their underlying design principles. The subsections are not arbitrary but rather reflect meaningful categories, such as pooling-based, block-level, attention-driven, dynamic, transformer-inspired, and hybrid strategies. Introducing the methods in this structured way highlights their commonalities and differences, which helps readers see how the fifteen subsections relate to each other and follow the overall logic more smoothly. As we discussed above, CNNs consist of three kinds of neural layers, they are convolutional layers, pooling layers and fully connected layers, each of which play essential roles in image detection. When images are passed through a CNN, the convolutional layers automatically generate multiple feature maps, each encoding different types of information. As we know, low-level layers will extract features including detailed information, however high-level layers extract semantic features people cannot understand. Then, these features will be passed to the fully connected layers to generate one dimensional feature representation of input data.

However, even though image recognition can be completed using the learned features containing useful information, researchers still believe that they have lost much information during the progressing, and also include some noise affecting the performance of models. So, more and more attention has been paid on how to prevent losing information and reduce noise simultaneously by directly handling features extracted from various layers. As a result, feature aggregation strategies are proposed to solve these problems. In Section 2, we briefly conclude the approaches of features integration researchers related to image processing proposed till now.

2.1 Combine local features and global features

Recently, some of research on CNNs indicated that we can categorize the features extracted from CNN into two categories, one is called local feature and the other is called global feature[13–16].

The features extracted from the whole image or from the fully connected layers are called global features or holistic features. However, the features extracted from the convolutional layers are called local features. Global features mainly include more discriminative and semantic information which aim to represent the whole image, whereas local features concentrate on detailed information implicitly suggesting these regions on the image[13].

Deep comprehensive multi-patches aggregation convolutional neural networks (DCMA-CNN), as shown in Fig. 2, proposed by Xie and Hu[13] combined these two types of features into one fused feature to represent the images. They proved that global features and local features have complementary relationships with each other. Then, fused features are generated by aggregating local and global features together. During the process of features integration, more discriminative information is included to present the input data based on the mutual support relationships. Even though fusing local and global features prevents losing crucial information to some extent, however, there is still some space to be improved. Since when they implement the operations of features fusion, they just rudely concatenate them together without filtering the noise which leads to bad influence on the final representations of input data.

Moving beyond simple concatenation, more sophisticated methods have been developed to intelligently align features of different granularities. A prime example is the dual-granularity feature alignment (DFA) approach proposed by Yin et al.[17] for cross-modality person re-identification. This method addresses the challenge of aligning features not only across different image parts, but also across different modalities (e.g., RGB and infrared). Instead of simple fusion, DFA introduces a prototype learning mechanism to map features of different granularities from the same identity to a common, shared prototype in the embedding space. This ensures that coarse global features and fine-grained local features are semantically aligned, allowing the model to learn a representation that is robust to both intra-class variations and cross-modality discrepancies. This work exemplifies a shift towards more structured and goal-oriented fusion strategies that actively model the relationships between different feature sets.

A more refined version of the feature combination is proposed by Lin et al.[18], where they introduce the global and local feature extractor (GLFE) and local temporal aggregation (LTA) for gait recognition. Unlike other models for concatenation, the GLFE module first extracts global and local features separately and only combines them after identifying the most relevant ones as shown in Fig. 3. This structured combination avoids noise and redundant information from deteriorating feature representations. Also, the LTA module maintains spatial and temporal consistency for brief video clips by interpreting expressive motion patterns effectively. Their approach also enhances accuracy using a generalized-mean pooling layer that preserves more spatial resolution than traditional pooling operations. Having settled the complementary roles of global and local features, the following section dives into how intermediate features from various CNN layers can be effectively aggregated to further enhance the image representation.

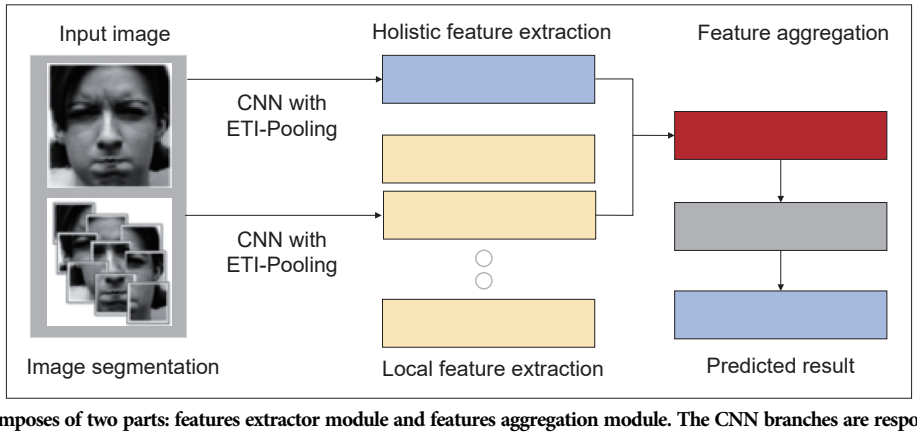

Fig. 2 DCMA-CNN composes of two parts: features extractor module and features aggregation module. The CNN branches are responsible to extract holistic feature and local features, then the extracted features will be fed to feature aggregation module to generate final representation for further process.

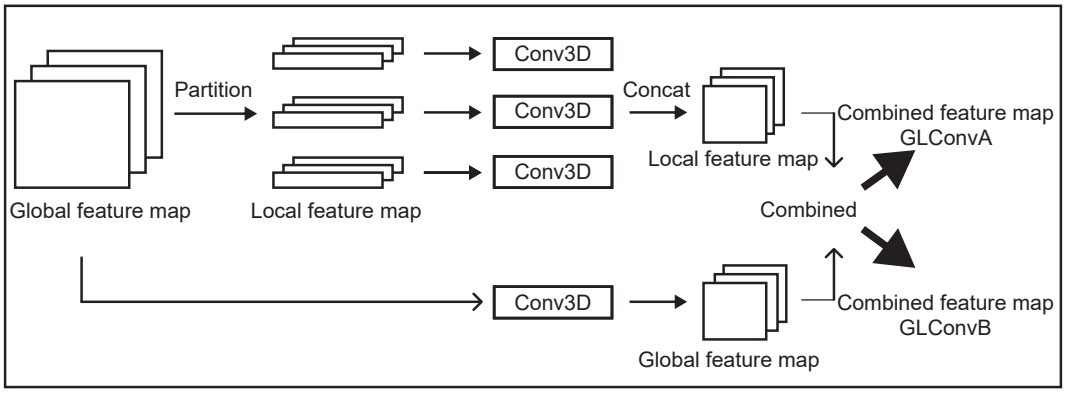

Fig. 3 This image explains how global and local feature extraction and combining to generate feature maps work.

2.2 Aggregate intermediate features extracted from multiple layers of CNN

This subsection reviews methods that aggregate intermediate CNN features to boost discrimination and suppress noise. Approaches either encode and concatenate multi-layer activations or tier layers by depth and fuse them with lightweight normalization/pooling. Currently, CNNs have demonstrated that they have performed superior performance in image computing[19]. And they are hierarchical networks consisting of convolutional layers and dense layers which will generate two kinds of features: one is called Conv-features and the other one is called FC-features[20, 21]. Conv-features which include spatial information are generated from different convolutional layers, however, FC-features include semantic information learned from fully connected layers[22, 23].

As a matter of fact, the information concealed in various layers will significantly increase feature discrimination[24]. Recent works have encoded and concatenated intermediate features using conventional encoding methods, such as BoVW, VLAD, IFK[24–26], or some advanced strategies explored based on them like FACNN, NetVLAD, and EMR[9, 19, 26]. The advanced approaches based on conventional encoding methods are composed of two categories. One aims to aggregate features extracted from all layers and explore the relations of intermediate features across layers. The other one also integrates intermediate features from various layers. But different from the first one, they simply categorize layers into low-level, mid-level and high-level ones, then optimize them with simple strategies like VLAD encoding, L_2 normalization and Average Pooling.

FACNN aggregates features from multiple pooling layers and injects semantic labels to form a fused, discriminative descriptor for classification. It uses layer-wise normalization and a dedicated aggregation module. FACNN proposed by Lu et al.[19] proved that combing semantic labeled information with fused features which are extracted from multiple layers will perform better performance than previous methods. As shown in Fig. 4, the FACNN model is composed of three components which are backbone pipeline, convolutional features aggregation module and softmax classifier. The convolutional feature aggregation module is responsible for aggregating features extracted from various pooling layers which have been normalized to be same size using L_2 normalization and include semantic information into final discriminative representations. The fused features generated before will be used to do the classification tasks.

EMR builds hierarchical global descriptors by encoding low/mid-level conv features and averaging high-level semantics, followed by normalization for robust retrieval/recognition. In Ref. [26], EMR model proposed by Wang et al. performs ideal performance by directly combining the features extracted from convolutional layers. As outlined in the paper, intermediate convolutional features of the low and mid-level layers are represented by a local descriptor (VLAD) vector[26] and reduced by

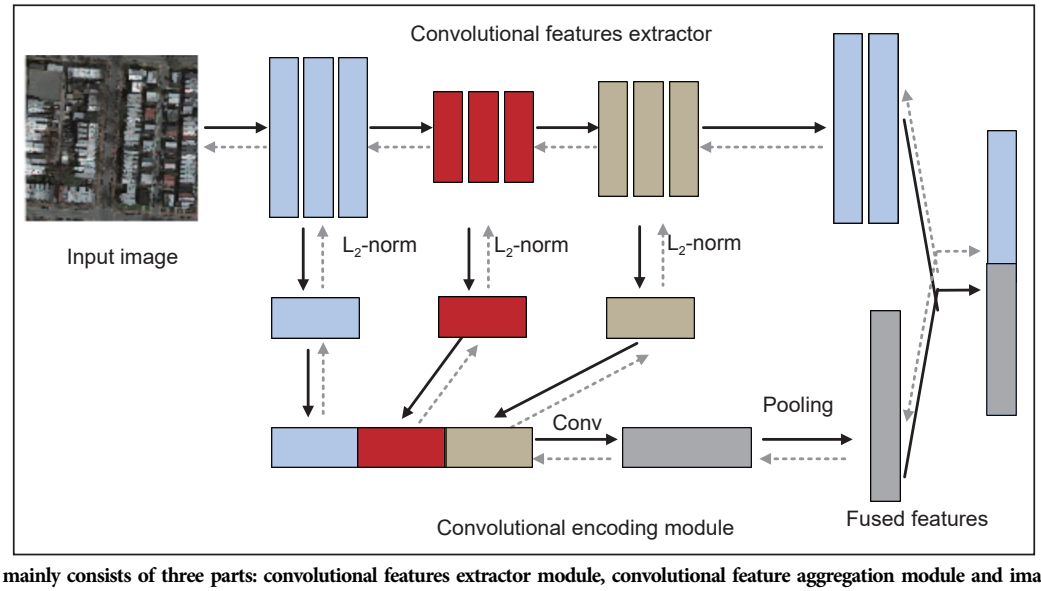

Fig. 4 FACNN mainly consists of three parts: convolutional features extractor module, convolutional feature aggregation module and image classifier. The feature aggregation module is used to aggregate features from convolutional layers and include semantic information into discriminative representations.

principal components analysis to obtain the hierarchical global features. Then, they compute high-level features/semantic features via average pooling. After that, the learned features are normalized to establish new global features, which are also the final discriminative descriptors. The final descriptors will be used to do further processing of images. While integrating intermediate features improves differentiation, following strategies, such as pooling methods, offer an additional mechanism to reduce noise. The next section explores these pooling-based approaches in detail.

2.3 Aggregate features extracted using max pooling and average pooling

This subsection contrasts max vs. average pooling (and hybrids) as noise-aware aggregation: max emphasizes discriminative peaks, average provides smoothing; combining them balances robustness and detail.Except focusing on the categories of features, some research also pay attention to some strategies for noise reductions, like max pooling and average pooling. As we know, most of CNN architectures are organized by convolutional layers, pooling layers and fully connected layers. And with the development of researches on CNNs, people find that max pooling layers are able to extract discriminative features of input images. To the contrary, the average pooling layers are able to make the extracted features become smoothed. Different from previous strategies, except the different approaches of exploring features relationships, they also apply simple approaches of noise reduction during the procession of intermediate features fusion.

MFCNN fuses max-pooled (discriminative) and avg-pooled (smoothed) features across layers to produce a noise-reduced yet expressive representation. MFCNN proposed by Cho et al.[27], as shown in Fig. 5, aggregates discriminative features extracted and smoothed features extracted using max and average pooling separately into one fused feature based on the ideas above. As a result, better performance on image detection of datasets SAR has been achieved using fused features. In Ref. [28], Lee and Nam also conducted similar features fusion strategy as above. In order to extract the most representative features, they apply max pooling over each segment separately and summarize them into single feature vectors by average pooling over it for each layer. The fused features from multiple layers finally lead to highly effective performance. Compared with the approaches which applied terminologies mentioned before, the models with max pooling and average pooling strategies have demonstrated that these two pooling methods can efficiently reduce noise during features extraction processing and achieve more discriminative representations of images. As shown in Table 1, we compare various strategies in terms of the features extracted from all layers in detail.

2.4 Aggregate features extracted from limited regions likely containing specific objects

Region-focused aggregation treats convolutional activations as local descriptors and keeps only the most informative parts of an image to suppress background noise. For example, some further studies using object position, a kind of spatial verification, or a proposed region network (RPN) have shown that local extracted features from restricted areas may boost object-based image detection accuracy[29–32]. Two representative methods are SpoC[33] and CroW[34]. SpoC builds a compact global descriptor by

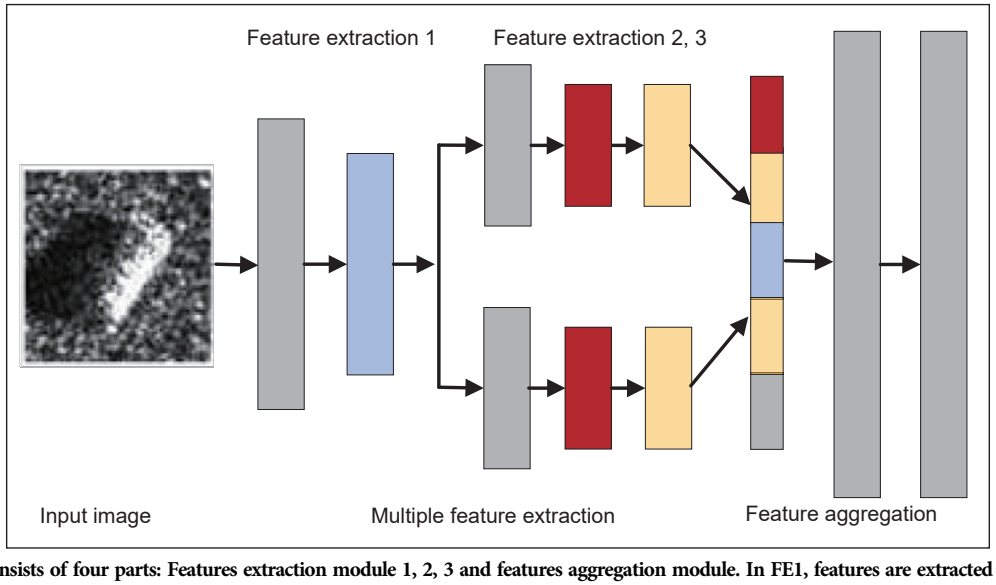

Fig. 5 MFCNN consists of four parts: Features extraction module 1, 2, 3 and features aggregation module. In FE1, features are extracted using max pooling and average pooling together to complete part of noise reduction. Max pooling and average pooing are applied in FE2 and FE3 individually to extract discriminative features and smoothed features. Finally, fused features were generated in the aggregation module.

Table 1 Comparison of features extracted from all layers

Category	Description
No feature relation	DCMA-CNN[13] directly aggregates features from convolutional layers and fully connected layers without analyzing the relations between features.
Feature relation	EMR[26] aggregates intermediate features from different level layers with L_2 normalization and average pooling.
	MFCNN[27] aggregates intermediate features using both max pooling and average pooling during processing.
	FACNN[19] aggregates intermediate features extracted from each pooling layer with pre-processing using L_2 normalization.

summing each feature channel across all spatial locations, optionally with a gentle spatial weight, while CroW improves this by reweighting both spatial positions and channels with simple, non-parametric statistics before pooling[34].

In Ref. [35], Jeong et al. described representing candidate image areas as regional descriptors and then combining them region by region, as shown in Fig. 6. They aggregate dense local features with second-order pooling to form each regional descriptor. This works well across diverse image-retrieval datasets but can have side effects when images are very similar across the whole scene or across partially overlapping regions.

Building on region-focused aggregation, SpoC[33] forms a compact global image descriptor by summing each feature channel over all spatial locations (optionally with a mild spatial weighting such as a center prior), followed by L_2 normalization (and optional PCA whitening). Compared with simple global max/average pooling, SpoC improves retrieval quality while remaining computationally efficient. Strengths: simple, fast, and parameter-free at aggregation time. Limitations: uniform sum pooling can overweight frequent background patterns; performance may drop under heavy clutter or misalignment. CroW extends this idea by applying lightweight spatial and channel reweighting before pooling, improving discriminability with negligible overhead[34].

As people study more deeply, they found that directly aggregating features from regional areas of an image into local descriptors will also include noise affecting performance of models. To overcome these problems, CroW proposed by Kalantidis et al.[34], etc., provides a clear and straightforward approach of using cross-dimensional weighting and integration to construct discriminative image representations efficiently. In CroW, after the model completes extracting features from the last spatial convolutional layer, it will automatically construct final fused features (CroW features) using non-parametric weighting schemes which are applied layer by layer spatially. The CroW framework offers the latest state-of-the-art results with low overhead in image recognition and has obvious qualities to help researchers achieve better insights on the features their models have learned. And it is a valuable scaffold within which to discuss and explore new weighting schemes, at the same time, gives us a clear way to investigate channel and spatial weights independently. Till now, we have discussed some feature aggregation strategies with max pooling and average pooling on whole images.

However, in contrast to those methods, region-by-region approach does not need to obtain all features from the whole image, they only select features with strong activations of various layers with powerful response on more active positions of images.This strategy does not only efficiently reduce noise by occluding unrelated objects from input data, but it also accelerates the speed of image retrieval in reality. As shown in Table 2, we simply conclude the comparison of these two kinds of methods.

2.5 Combine layers into blocks by different features resolution

Block-level aggregation groups layers by resolution/scale and fuses them via skip, iterative, or hierarchical connections to capture multi-scale context without inflating parameters.

Most research pay more attention to features of each layer, however some works proposed that visual acknowledgement calls for rich representations of low to high dimensions, small to large scales and low to high resolutions. Based on that, people design deeper and wider architectures to learn better representations of images, but they also introduce more parameters into computation. So, to overcome these problems, how to better combine layer blocks across a network becomes particularly important. In Fig. 7, it shows some existing works with and without combining layers into blocks. On the basis of the ideas, U-

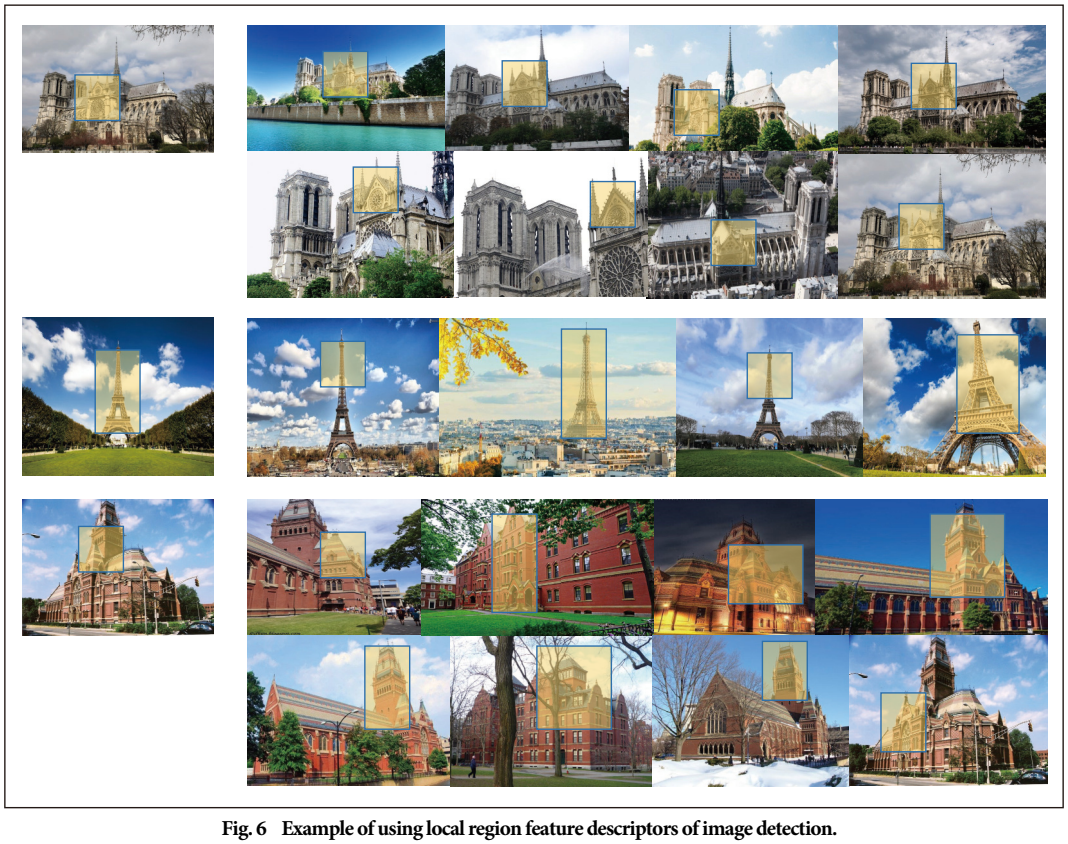

Fig. 6 Example of using local region feature descriptors of image detection.

Table 2 Comparison of pooling schemes with and without region-by-region strategy

Aspect	Pooling	Pooling with region-by-region
Strategy	Extract features from the whole image using pooling schemes	Extract features from active regions using pooling schemes
Advantages	Good for image recognition	More efficient and better for image retrieval
Disadvantages	Features may contain more noise	Less accurate for image recognition

Net proposed by Ronneberger et al.[36] incorporated parts of the network with skip connections as it is generally used for classification and segmentation tasks. Eventually, U-Net achieves ideal performance on different applications, especially for biomedical segmentation with few annotated images.

As skip connections are successfully applied to reduce the depth of networks, features aggregation can be implemented across layers through simple operations[37]. Research related to architectures demonstrates that the integration of deep layers strengthens identification and resolution in comparison to current fusion schemes. Unlike the U-Net suggested by Ronneberger et al.[36], Yu et al.[3] further promoted architectures of networks with various integration blocks to help integrate knowledge across layers. They apply iterative deep aggregation (IDA) and hierarchical deep aggregation (HDA) strategies to allow networks to be more reliable and promote less parameters computing. As shown in Fig. 7, the IDA strategy divides the architecture into blocks according to features resolutions and applies skip connections to fuse scales and resolutions iteratively from low-level to high-level layers. As a result, the shallowest stages are aggregated the most for further processing. However, HDA aggregates blocks and stages across a tree structure and fuse features. Based on that, low-level and high-level layers are integrated together to learn richer combinations, including more of features hierarchy. Their architectures which involve and extend fully connected networks and pyramid networks with hierarchical and iterative skip connections demonstrate that they improve previous representation and resolution via less parameters computing. After discussing aggregation at the block level, we now proceed to state-of-the-art methods that extend beyond the standard CNN architectures, incorporating transformers, dynamic pooling, and many more techniques that improve performance across a various of tasks.

2.6 Recent progress in putting together image features

New study has found creative ways to combine picture features that build on the ones we have already talked about. This makes the field even more advanced.

Transformers for image aggregation. The original designs of transformers were for natural language, but they turn out to be applied very well to handle image features. Transformer models like vision transformer (ViT) and swin Transformer[38] have dramatically changed how image features are fused. Instead of relying solely on local convolution, ViT splits images into patches—each patch acts as an individual token that can communicate with every other patch through self-attention. This mechanism allows for both local and global information to be brought together. Swin Transformer takes a unique approch by using shifted windows to break down and examine images, which helps it efficiently pick up on features at different scales. It goes well with the local feature extraction methods we talked about in Sections 2.1–2.4. Expanding on this concept, hybrid models like feature fusion vision transformer (FFVT), SwinPA-Net, and feature pyramid transformer blend information from multiple image layers or regions. This method enables the models to capture deeper and more relevant context from the images. As a result, these innovations have noticeably improved the accuracy of image classification and recognition, moving well beyond what standard convolutional neural networks can accomplish[39,40].

State-space models for vision. Over the past few years, researchers have turned to state-space sequence models like S4 and Mamba because they offer a faster, more computationally efficient way to work with visual data compared to transformers. These models shine when you're dealing with long sequences of images or extremely high-resolution visual situations where traditional transformers might struggle due to their computational demands[41].

These models excel because of how they process information differently. When analyzing image sequences or scanning across various regions of an image, they maintain and refresh their working memory at a rate that grows proportionally with the data size. This is a big advantage over transformers, which become exponentially more expensive as you add more data points. Because of this efficiency, scientists can still extract rich, detailed information from images without requiring the kind of massive computing power that would otherwise be necessary.

For computer vision applications, specialized versions like MambaVLT[42] and VideoMamba[43] have been built to

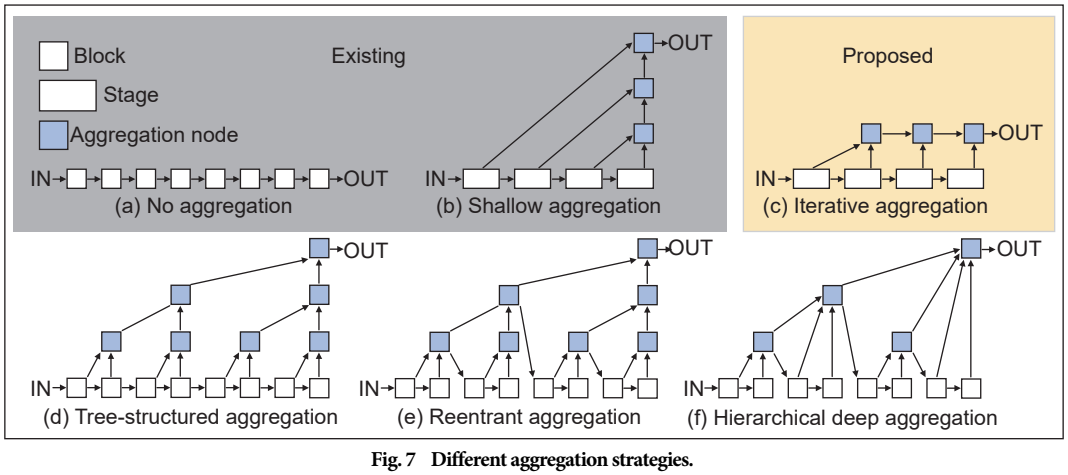

Fig. 7 Different aggregation strategies.

complement the CNN and transformer models that researchers already use. We can essentially stack these state-space models on top of your existing feature extractors, which gives you more flexible and scalable ways to combine different types of visual information. The beauty of this approach is that it remains effective across different types of vision tasks.

Foundation models as feature aggregators. Large-scale foundation models—includ contrastive language-image pretraining (CLIP), distillation with no labels (DINO), and segment anything model (SAM)[44]—serve as versatile feature extractors thanks to extensive pretraining. These models aggregate image features in ways that support prompt-based fusion, adapter-based integration, domain adaptation, and transfer learning. By combining outputs from models like CLIP's visual embeddings, SAM's segmentation results, and DINO's semantic representations—approaches such as Trident—researchers have achieved impressive robustness and adaptability, making it easier to tackle new domains and tasks[45].We summarize the core image-aggregation families—their mechanisms, strengths, and limitations in Table 3.

Self-supervised learning for feature aggregation. With self-supervised learning, you do not need a lot of labels to make a strong model of a picture these days. We have come a long way in self-supervised visual representation learning with contrastive learning tools like MoCo v3 and DINO[48, 49]. These methods figure out how to combine features by picking the ones that agree with each other the most across different improved views of the same picture while still being able to tell them apart from other pictures in the dataset. This method builds on the ones we talked about in Section 2.3 for getting rid of noise.

Dynamic feature aggregation. Dynamic feature aggregation strategies that change based on the input picture have been studied in recent years. Conditional convolution networks use dynamic kernels that depend on the input to group features. This lets the network change how it extracts and groups features based on the content of each image[50]. This method is more flexible and efficient than static methods, and it builds on the ideas of adaptive feature selection that were talked about in Section 2.4.

Multi-scale feature aggregation. To build on the idea of putting layers together into blocks (Section 2.5), more recent research have looked at more complex ways to mix features from various sizes. Transformer architectures and multi-scale feature aggregation are mixed in feature pyramid transformer (FPT) and other ways. This lets the model effectively combine features from various scales and locations[51]. Other works in multi-scale feature aggregation has brought forward methods such as DenserNet, a novel convolutional neural network (CNN) architecture designed for visual localization, which leverages multi-scale feature extraction to improve visual localization and image representation[52]. DenserNet aggregates features at multiple semantic levels, allowing richer representations that enhance keypoint detection even under occlusions and varying illumination. This approach contrasts with earlier methods that relied primarily on a single-scale feature map, thereby missing crucial details across different spatial resolutions. Recent advancements in multi-scale aggregation have introduced the concept of combining convolutional features with transformer-based approaches to extract both local and global feature representations. Techniques like the global and local convolutional layers (GLConv) are used to maintain a separation between global and local features until it is optimal to merge them. This hybrid approach has shown improvement in tasks such as gait recognition and semantic segmentation, especially in environments where both spatial and contextual information are critical[53]. In facial recognition applications, multi-scale feature aggregation has been shown to outperform single-scale approaches by effectively capturing features at different resolutions, leading to better recognition accuracy even under challenging conditions like occlusion or low light. The attention filter mechanism used in DenserNet helps select the most relevant features by suppressing irrelevant details, enhancing the discriminative power of aggregated features for visual localization[52].

Feature aggregation based on attention. Improved attention systems can better combine traits in tasks like recognizing pictures. A network can change how much different feature channels and spatial places are worth during aggregation using methods such as convolutional block attention module (CBAM) and squeeze-and-excite (SE) blocks[54, 55]. These methods build on the ideas of weighted pooling we talked about in Section 2.4 by helping the model focus on the most important characteristics for the job at hand. The S-SELSA module employs semantic similarity through a modified structural similarity index (SSIM) to aggregate global features across frames, which enhances robustness against noise and redundancy[56]. Unlike traditional max and average pooling methods, S-SELSA selects semantically similar frames globally, avoiding over-aggregation by focusing on non-adjacent but similar frames. This approach provides an adaptive mechanism for selecting the most informative features for aggregation.

Recent advancements in attention-based pooling have included the use of reinforcement learning for adaptive pooling region selection. In this method, an agent dynamically adjusts the size and location of pooling regions based on the spatial complexity of the input features. This adaptive pooling method has made big progress in tasks like medical imaging and image super-resolution because it lets the model focus on the most important parts, which makes feature aggregation more effective[53].

Table 3 Comparison of networks, features, and noise reduction techniques

References	Networks	Features	Noise Reduction
BoVW[25], VLAD[26], IFK[24]	Traditional networks	Features without semantic information	No
DCMA-CNN[13]	Deep neural networks	Local and global features	No
FACNN[19], EMR[26], NetVLAD[9]		Intermediate features with semantic information	No
MFCNN[27]		Representative and smoothed features	Max and average pooling
SpoC[33]		Features extracted from limited regions	Sum pooling
CroW[34]			Simple weighting schemes
HDA[3]		Features extracted from layer blocks	Hierarchical deep aggregation
IDA[3]			Iterative deep aggregation

A practical example of attention-based pooling can be seen in medical imaging, where attention mechanisms help focus on the most relevant regions of an image, leading to improved diagnosis accuracy compared to traditional pooling methods. Attention-based pooling has been shown to improve classification accuracy by up to 5% compared to max pooling in benchmark datasets such as CIFAR-10 and ImageNet[56]. This shows that it is good at choosing features that are informative.

These developments in image feature aggregation have made the field much further along and opened new ways to get useful information from pictures and put it together. Some deep learning models have done better at many computer vision jobs since they started using these methods.

Feature aggregating methods have changed over time. They began with simple traditionally methods like BoVW, VLAD, and IFK, but with time they moved on to deep learning models. With CNNs, new approaches to combine features were developed. Later, pooling algorithms, block-level connections, and residual or context-aware designs were included. In the last few years, transformers and hybrid models have pushed the field even further. To make these changes easier to follow, Fig. 8 shows a timeline of the main developments and how each step built on the one before it.

2.7 Discussion about strategies of feature aggregation on images

As we discussed before, feature aggregation strategies on images can be divided into two parts: traditional terminologies, such as BoVW, VLAD, IFK[24–26], and advanced deep learning approaches (especially CNNs), such as FACNN, NetVLAD, EMR[19, 9, 26]. In addition, the CNNs strategies are composed of two parts: aggregating features using layer-wise methodology like DCMA-CNN, MFCNN[13, 27], and fusing features using iterative and hierarchical skip connections like IDA and HDA[3]. As shown in Table 4, we present the comparisons among various feature fusion methods of images in detail.

2.8 Residual feature aggregation for image super-resolution

The residual feature aggregation (RFA) network is an enhanced image super-resolution residual module for feature aggregation which uses skip connections. This was proposed by Liu et al.[57]. Direct forwarding of the features of the prior layer is allowed, with increased representation without the added complexity. The network also includes the enhanced spatial attention (ESA) block that resizes features adaptively with spatial context to allow more discriminative information to be obtained.

The RFA architecture is lightweight in style, where it requires only a 1×1 convolution per every four residual blocks to reduce the computational cost without losing much performance. The method is experimented on benchmarking datasets such as Set5, Set14, B100, Urban100, and Manga109 based on PSNR and SSIM measures. The method can be generalized to extend to video-related tasks using an adaptive selection mechanism, although it is primarily employed to enhance images. The above-mentioned principles of residual feature aggregation and enhanced spatial attention also pave the way for more complex image restoration tasks, such as ultra high-resolution inpainting discussed next.

2.9 Contextual residual aggregation for ultra high-resolution image inpainting

Yi et al.[58] introduce the contextual residual aggregation (CRA) module to improve ultra high-resolution image inpainting. Compared with the RFA super-resolution framework which is primarily concerned with enhancing global detail via light-weight residual connections, the CRA module has a two-stage architecture that first ensures global consistency and then refines local details, making it extremely effective for inpainting tasks. The CRA method combines residual features with contextual variables to facilitate high-quality inpainted regions in the aspects of realistic textures and natural transitions. The CRA's intrinsic spatial attention mechanism helps the model focus on salient locations, ensuring that only the most relevant features are summed. This method significantly reduces computational costs as it focuses on residual features rather than computing all the feature maps again. The CRA module is evaluated on datasets like Places2 and CelebA-HQ, and is shown to outperform state-of-the-art approaches in terms of SSIM and PSNR for image inpainting tasks. The light-weight RFA network, with skip connections and enhanced spatial

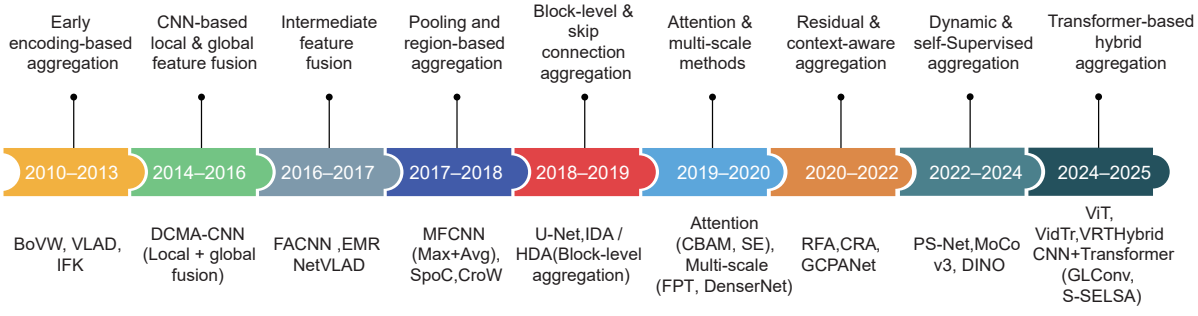

Fig. 8 Timeline of feature aggregation methods in deep learning. The figure highlights the shift from early handcrafted descriptors (BoVW, VLAD, IFK) to CNN-based fusion, pooling strategies, block-level connections, residual/context-aware methods, and most recently, transformer-based hybrid models.

Table 4 Image aggregators comparison

Aggregator	Aggregation mechanism	Advantages	Limitations
CNN[46]	Local convolution / pooling	Local details; fast	Poor global context
Transformer (ViT, Swin[38, 47])	Patch-token self-attention	Global relations; multi-scale	Memory-intensive
State-space (S4, Mamba)	Latent state update	Scalable; robust for long input	Limited interpretability; immature
Foundation Models[45]	Large-scale; cross-modal	Adaptable; zero-/few-shot	Large model size; finetuning required

attention, well improves super-resolution performance. These residual aggregation principles generalize naturally to more difficult restoration tasks, which we consider next in the context of ultra high-resolution image inpainting.

2.10 Global context-aware progressive aggregation for salient object detection

GCPANet integrates global context with progressive aggregation to refine saliency maps: FIA interweaves multi-level features, while hierarchical attention strengthens channel- and spatial-wise responses. The paper by Chen et al.[59] introduces the global context-aware progressive aggregation network (GCPANet), which improves salient object detection by effectively integrating low-level appearance features, high-level semantic features, and global contextual information. This integration is achieved through the feature interweaved aggregation (FIA) module, which progressively refines feature aggregation by interleaving different levels of features, leading to more precise saliency maps. Additionally, the hierarchical attention (HA) module enhances feature selection by adaptively strengthening channel- and spatial-wise responses, ensuring that only the most relevant information is retained for aggregation. These mechanisms collectively enable the network to capture both fine details and broader scene semantics, thereby improving the overall performance of salient object detection.

The paper utilizes ResNet-50 as the backbone for feature extraction and evaluates performance using metrics such as F1 scores and IoU on diverse datasets. While the study primarily focuses on images, the feature aggregation strategies used in GCPANet can be extended to video-based tasks, potentially improving object detection in dynamic environments. Now the following section addresses how feature aggregation can be tailored to boost the detection of small objects.

2.11 Feature aggregation for small object detection

Feature aggregation is crucial for detecting small objects, especially in high-resolution images. Bashir and Wang[60] developed residual feature aggregation (RFA)-a super-resolution network that enhances the representation of small objects by combining residual learning with multi-scale feature aggregation. The method enhances local and global features without inducing over-aggregation through residual connections.

The super-resolution module further improves detection by enhancing the sharpness of fine details, making small objects more identifiable. This approach is particularly useful in remote sensing imagery, where small objects are typically difficult to recognize. Compared to benchmark datasets using Precision, Recall, F1-score, and IoU, this approach promises enhanced small object recognition without diminishing efficiency. Though designed for remote sensing, it can be applied in other areas requiring fine-grained object detection. Next, we explore dense relation distillation with context-aware aggregation—a method aimed at tackling the issue of few-shot object detection.

2.12 Dense relation distillation with context-aware aggregation for few-shot object detection

The paper by Hu et al.[61] introduced dense relation distillation with context-aware aggregation as a novel approach to improving few-shot object detection. The new approach revolves around aggregating valuable features by distilling dense spatial relations and aggregating contextual information from the surrounding objects. Through the exploitation of both local and global features, the new approach enhances the model's object detection ability, especially in the case of sparse labeled data. The method strictly selects and compresses useful spatial relationships, avoiding the danger of overfitting with limited datasets. Not only does this approach integrate local features and global features, but it also pays attention to dense spatial relation distillation, effectively easing overfitting in limited data situations—a distinction from more conventional approaches. The compact relation distillation ensures the model to learn the most beneficial features, making a trade-off between detection performance and computation efficiency. The efficiency of the method is validated by evaluating on the COCO dataset, where it outperforms other state-of-the-art few-shot detection techniques. By effective abstraction of spatially rich relations and context cue aggregation, this approach addresses the limited access to labeled data in few-shot scenarios. The following section applies these aggregations to real-time semantic segmentation, where speed and accuracy are crucial.

2.13 Feature aggregation for real-time semantic segmentation

In real-time semantic segmentation, feature aggregation methods have to balance computation efficiency and segmentation performance. Feature pyramid aggregation network (FPANet) from Ref. [62] proposes a multi-scale feature aggregation approach that exploits a pyramid structure to aggregate features at various scales. The approach enables the model to learn both fine-grained details and high-level context information, which in segmentation tasks, such as autonomous driving and robotics, is crucial.

The pyramid aggregation process in FPANet is not over-aggregation but emphasizes the most essential spatial information at each step without redundant feature merging. The model uses lightweight convolutional architecture, which ensures real-time capability without sacrificing accuracy. FPANet is tested on popular datasets like Cityscapes and demonstrates better performance with metrics such as mean intersection over union (mIoU) without sacrificing inference speed. While it is strong, FPANet does not address space-time aggregation or challenges like motion blur or occlusion typical in video segmentation. It is, however, very efficient for static image segmentation where computational expense is a significant factor. FPANet's pyramid aggregation approach demonstrates that balancing fine-grained detail with high-level context is key for real-time segmentation. Shifting our focus, we now consider multi-stage aggregation methods which aim at improving multi-label image recognition.

2.14 Multi-stage feature aggregation for multi-label image recognition

The multi-stage feature aggregation (MSFA) network, proposed by Chen et al.[53], is designed to improve multi-label image classification by aggregating features across stages. Unlike single-stage methods, the multi-stage method adopted here ensures that only the most informative features are propagated forward, with the optimal balance between retaining fine detail and maintaining computational simplicity. The multi-stage architecture allows the model to combine low-level features from shallow layers with high-level features from deeper layers, thus being more capable of extracting fine-grained details as well as overall contextual information. Such a comprehensive feature representation is critical for detecting multiple objects of varying sizes and complexities within an image. The MSFA network utilizes a hierarchical structure that progressively selects and aggregates features at each level, passing only the most relevant information

to the next level. This prevents over-aggregation, with an ideal tradeoff between computational complexity and recognition precision. While there is no use of attention mechanisms in the technique, its successive feature aggregation scheme ensures effective selection of useful features, which is vital to the multi-label recognition problem. The performance of the model is measured in terms of mean average precision (mAP), a typical metric for multi-label image recognition, showing its efficacy in different recognition tasks. Even though the paper does not consider issues such as motion blur or occlusion, the hierarchical aggregation process may be able to counteract such problems.

2.15 Advanced dynamic aggregation

Many papers in feature aggregation focuses on adaptive methods that intelligently balance local detail extraction and global context preservation. In this regard, Ren et al.[63] proposed a shunted self-attention mechanism for vision transformers that fuse tokens at varying scales, effectively promoting both local and global feature fetch. This multi-scale token summarization allows the model to focus on the most informative tokens, thereby enhancing feature representation while maintaining low computational power. The approach strikes an optimal balance between accuracy and efficiency, as evidenced by its improved performance on ImageNet, and holds potential for extension to video analysis through the incorporation of temporal relationships. Complementing this, Chen et al.[64] introduce the pooling strategy network (PS-Net), a learnable pooling strategy that adapts to dynamically select or mix max pooling, average pooling, and generalized pooling for visual semantic embedding tasks. Unlike static pooling approaches, PS-Net adapts to the input data, balancing the preservation of local details and global context. This dynamic selection prevents over-aggregation and enhances computational efficiency by eliminating redundant pooling operations. The flexibility of PS-Net is demonstrated through its successful integration with pre-trained networks like ResNet and VGG, and its efficiency is verified on fine-grained classification benchmarks such as CUB-200-2011, Stanford Cars, and FGVC Aircraft.

Together, these adaptive techniques showcase the evolution of feature aggregation methods: one leveraging self-attention to focus on important tokens in various scales, and the other that employed a dynamic, learnable pooling technique to enhance feature extraction processing. Such approaches not only improve the accuracy-efficiency trade-off in image classification tasks but also provide a robust framework that could be generalized across various visual recognition applications.

A comprehensive summary of these image-based feature aggregation strategies, including their evaluation datasets and representative performance, is presented in Table 5. In Section 3, we extend these aggregation ideas to video and explore ways in which dynamic methods can be generalized to capture both spatial information and temporal dynamics for enhanced video analysis.

Table 5 Comparison of current feature aggregation strategies in images

Model	Family/strategy	Aggregation scope	Dataset	Metric	Representative performance
DCMA-CNN[13]	Hierarchical multipatch aggregation CNN	Multi-patch facial regions	CK+, FER2013, RAF-DB	Accuracy	95.1% (RAF-DB), 99.1% (CK+)
GL-Conv[18]	Global-local feature aggregation (gait recognition)	Global silhouettes + local parts (temporal)	CASIA-B, OU-MVLP	Rank-1 / mAP	98.6% R1 (OU-MVLP), 97.3% (CASIA-B)
FACNN[19]	Feature aggregation CNN (multi-scale)	Convolutional blocks, hierarchical	UC Merced, AID, NWPU-RESISC45	Accuracy	96.8% (AID), 94.2% (NWPU-RESISC45)
EMR[26]	Rich hierarchical feature Aggregation	Multi-level CNN features	UC Merced, WHU-RS19	Accuracy	98.1% (UC Merced), 96.9% (WHU-RS19)
MFCNN[27]	Multiple feature aggregation CNN for SAR ATR	Layer-wise	MSTAR	Accuracy	99.1% (10-class MSTAR)
SPoC[33]	Sum-pooled convolutional features	Global pooling (center prior)	Oxford5K, Holidays, UKB	mAP	65.7% mAP (Oxford), 80.2% (Holidays)
CroW[34]	Cross-dimensional weighted aggregation	Channel- and spatial-weighted pooling	Oxford5K, Paris6K, Holidays	mAP	78.7% (Oxford), +5%–8% over SPoC
NetVLAD[9]	VLAD-inspired end-to-end aggregation	Local convolutional descriptors	Pitts250k, Tokyo24/7	Recall@K	Recall@1 = 71.6% (Pitts250k)
DLA[3]	Deep layer aggregation (IDA + HDA)	Hierarchical + iterative across layers	ImageNet, Cityscapes, CUB, Cars	Top-1 / IoU	22.0% Top-1 error (ILSVRC), 75.9 mIoU (Cityscapes)
RFANet[57]	Residual feature aggregation + ESA	Residual branches with spatial attention	Set5, Set14, Urban100, Manga109	PSNR / SSIM	32.66 dB PSNR (Set5 ×4), 0.950 2 SSIM
CRA[58]	Contextual residual aggregation (for inpainting)	Patch-based residual borrowing	Places2, CelebA-HQ, DIV2K	L1 / FID / SSIM	5.44 L1, 0.884 SSIM, FID = 4.89 (512px Places2)
GCAPANet[59]	Global context-aware progressive aggregation network	Multi-stage global–local progressive fusion	DUTS / ECSSD / HKU-IS	F-measure / MAE	F_β = 0.918 (DUTS), MAE = 0.038
DCNet[61]	Dense relation distillation + context-aware aggregation	Support–query relation + adaptive scale pooling	PASCAL VOC / MS-COCO (few-shot)	mAP	59.6% mAP (VOC novel split, 10-shot), +10% over Meta R-CNN in 1-shot
FPANet[62]	Feature pyramid aggregation net (SeBiFPN + BRM)	Multi-level pyramid fusion (encoder–decoder)	Cityscapes / CamVid	mIoU/FPS (Frames Per Second)	75.9% mIoU (Cityscapes) @39 FPS; 74.7% (CamVid)
MSFA[53]	Multi-stage feature aggregation for multi-label recognition	Cross-scale + local–global fusion (hierarchical loss)	VOC / MS-COCO	mAP / F1	95.2% mAP (VOC), 83.0% mAP (MS-COCO), +5.1% over baseline

3 Feature aggregation in videos

Video feature aggregation methods are organized into nine subsections, each representing a methodological family—such as pooling extensions, temporal block-based, attention-based, dynamic, and hybrid approaches. This structure highlights their conceptual links and prevents the subsections from appearing as isolated techniques.

Feature aggregation in video analysis builds upon the techniques developed for image recognition, adapting them to handle temporal dynamics. Unlike images, videos present unique challenges, such as complex motion, uncontrolled postures, and varying frame quality, making the aggregation process crucial for effective representation. To address these challenges, diverse aggregation strategies, including pooling mechanisms, attention-based weighting, and motion-guided methods, have emerged. The following subsections systematically review these strategies by grouping similar approaches for clear understanding.

3.1 Frame-level aggregation techniques

Frame-level aggregation techniques primarily aim to combine features extracted from individual video frames into cohesive representations that capture relevant spatial and temporal information. These methods often balance simplicity and computational efficiency with the ability to retain critical details across frames. However, directly aggregating frame-level features can introduce substantial noise, especially from low-quality frames, motivating the development of specialized aggregation methods. In the subsequent subsections, we explore several prominent frame-level aggregation strategies, ranging from straightforward pooling operations to advanced methods leveraging geometric relationships and inter-layer interactions.

Frame-level aggregation with average pooling. Recent research indicates that video recognition can be treated as a template-matching problem, where each template consists of multiple frames (images) of the same object, including both high-quality and low-quality frames. Based on this observation, several methods propose aggregating the most discriminative features extracted from high-quality frames to create fused representations for better prediction. However, Hassner et al.[65] demonstrated that simply discarding low-quality frames does not necessarily improve face recognition accuracy, as even low-quality frames may contain valuable features. Consequently, aggregating features from every frame is recommended to avoid information loss[66]. Typically, average pooling methods are employed to produce smoother fused features.

Nevertheless, some studies, such as QAN[67, 68], argue that directly averaging frame-level features introduces considerable noise into the final representation. To address this, they suggest incorporating learned weighting schemes during feature aggregation. Although weighting schemes reduce noise effectively, Gong et al.[69] point out in their component-wise feature aggregation network (C-FAN) that noise often remains at the individual feature-component level. Specifically, they note that each component may contain varying amounts of identity information, and noise across these components is only weakly correlated. Motivated by the observation that intra-class correlations among individual feature components are typically low, C-FAN proposes component-level aggregation to further suppress noise, as illustrated in Fig. 9. Experimental results confirm that C-FAN effectively predicts the quality of feature vectors adaptively, integrating only the most discriminative components for improved video analysis performance.

While average pooling provides a straightforward method for frame-level feature aggregation, its inability to distinguish between frame qualities motivates more sophisticated approaches, as discussed next.

Combine discriminative features from low quality frames with features from high quality frames. As we discussed before, lots of features fusion strategies are proposed to improve models, such as direct elimination of low-quality frames, using learning weight schemes, etc. That means these aggregation methods treat low-quality frames as noise and focus only on high-quality ones[66, 70]. However, even low-quality frames can contain informative signals. Thus, some works argue that every frame should be weighed adaptively, rather than ignored.

To the best of our knowledge, many videos are captured under varying conditions such as lighting, resolution, and pose variations. As suggested by NAN and related research, an ideal algorithm should highlight useful frame features while suppressing irrelevant or noisy ones during feature aggregation, regardless of input data quality[68, 71]. As illustrated in Fig. 10, an attention-based integration network is proposed, which adaptively assigns fine-grained weights to features across all frames in each dimension. Importantly, this network can produce consistent representations efficiently, achieving strong performance with low computational costs and minimal memory requirements.

Although adaptively weighting features significantly enhances

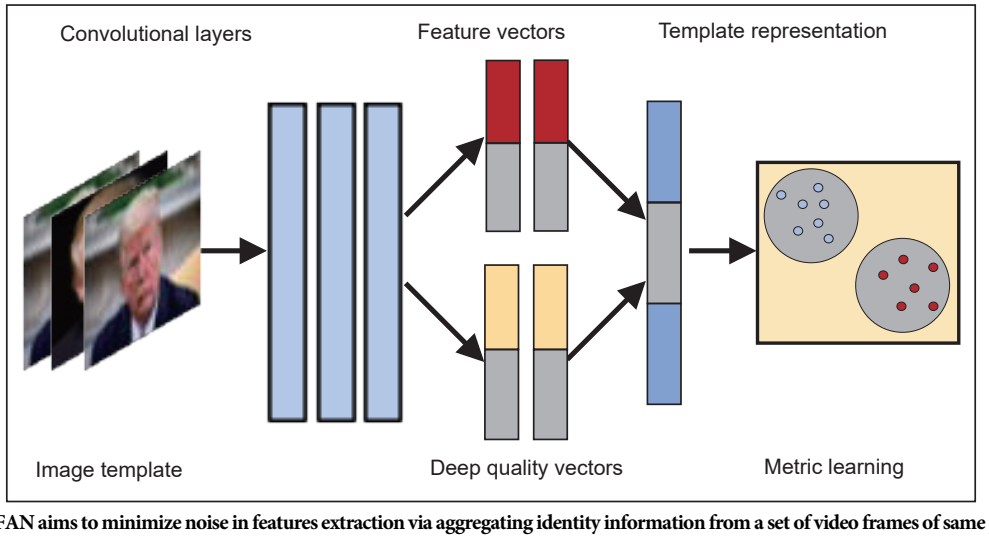

Fig. 9 C-FAN aims to minimize noise in features extraction via aggregating identity information from a set of video frames of same person.

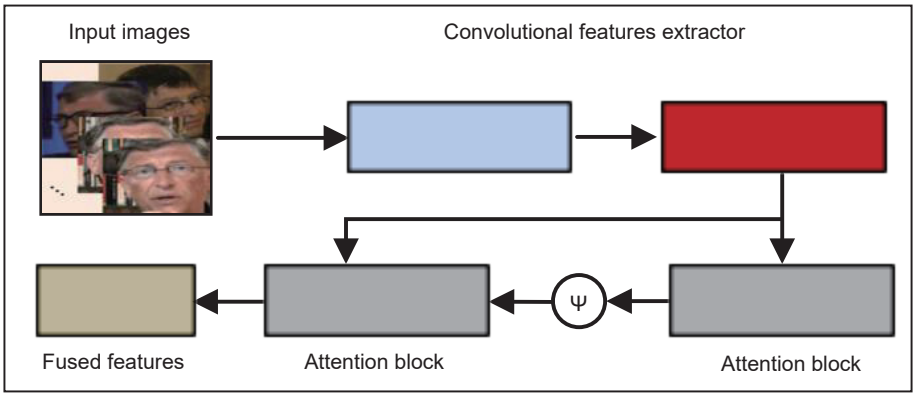

Fig. 10 NAN forms a powerful and discriminative face representation with attention-based aggregation module embedded into CNNs.

video representation, considering the spatial relationships among features can further refine the aggregation process.

Geometry-based feature aggregation. As we know, to enhance video detection effectively, integrating frame-level features using noise-reduction techniques is crucial for preserving discriminative information. Numerous approaches have been proposed for efficiently aggregating features to construct discriminative video representations[66, 68, 70–75]. For instance, common techniques such as max pooling and average pooling help extract robust features while reducing noise during feature aggregation[27, 28, 65–69, 76, 77]. Additionally, enhanced methods utilize learned weighting schemes to integrate frame-level features, primarily based on image quality (resolution)[33, 34, 67, 68, 74, 78]. These approaches, termed "quality-based" methods, assign weights solely based on image resolution but neglect relationships among frames in the feature space.

However, Peng et al.[79] highlights that quality-based approaches overlook spatial relationships within the feature space. To address this, they propose geometry-based feature aggregation (GFA), which incorporates spatial relationships into the aggregation process, yielding more discriminative representations. Specifically, their method defines a reference face image—captured without distortion and under ideal pose and lighting—as the central point in the feature space. Distances between this reference image and other frame features are calculated within the same feature space. This geometric approach aims to ensure that the aggregated features lie closer to the reference object's center, enhancing their discriminative power. As illustrated in Fig. 12, clear distinctions exist between quality-based and geometry-based feature aggregation strategies. Furthermore, geometry-based methods effectively compute the distribution center, achieving balanced feature distributions across frames, even when individual frames lack strong discriminative cues.

Geometry-based approaches effectively leverage spatial relations; however, capturing the complementary nature of inter- and intra-layer features offers additional potential, as explored subsequently.

Inter-feature and intra-feature aggregation. This approach assumes that feature maps from various layers capture complementary levels of information—shallow layers contain detailed texture features, while deeper layers encode semantic abstractions. By fusing these together, FAN achieves improved accuracy over conventional models. Based on that, FAN proposes that feature maps in different layers contain useful information, even though some are considered as noise[80]. FAN generates the most discriminative representations of input data by fusing intra-features and inter-features. Intra-features are regarded as the relationships between feature maps and the relationships between features of different layers are considered as inter-features. The experimental results FAN can achieve better performance compared with existing state-of-the-art models. However, although they paid more attention to aggregating inter-features and intra-features, the noise generated in the intra-features is neglected in their models.

Having discussed frame-level aggregation techniques, the next subsection will explore attention-based aggregation methods that aim to refine the selection of frames and regions by emphasizing their discriminative features.

3.2 Attention-based aggregation techniques

Attention-based aggregation techniques significantly enhance the feature aggregation process by selectively weighing features according to their importance or relevance to a specific task. Leveraging attention mechanisms allows models to focus dynamically on crucial frames or regions within frames while suppressing less informative or noisy features. This selective emphasis greatly improves performance, especially in tasks prone to occlusions, viewpoint variations, or redundancy. The following subsections introduce advanced attention-driven approaches, illustrating how multi-granularity and spatiotemporal attention mechanisms effectively refine feature representations for diverse video analysis applications.

MG-RAFA: multi-granularity reference-aided attentive feature aggregation. Zhang et al.[81] introduced a novel approach, multi-granularity reference-aided attentive feature aggregation (MG-RAFA), to improve the task of video-based person re-identification (reID), which involves matching the same person across different video clips captured at various times and locations.

One of the main problems in video-based person re-identification is redundancy and variation in video frames. Since videos contain multiple frames of the same individual, there is often significant repetition of information across time. Another major issue is occlusion and motion blur, which occur frequently in real-world surveillance videos. When a person is partially obstructed or moving rapidly, key identity features might be lost or distorted. Many existing ReID methods struggle with these situations because they aggregate features indiscriminately, leading to degraded performance in recognizing individuals under occlusions or blurring.

To address these problems, the author introduced MG-RAFA framework. This method enhances the aggregation of spatio-temporal features by introducing attention mechanisms that selectively emphasize important frames and regions. Instead of treating all frames equally, the model assigns different attention weights based on how informative each feature is concerning the

global context. A key innovation in MG-RAFA is the use of reference feature nodes (S-RFNs) to model global relations. Instead of using all the original feature nodes, which can be computationally expensive, the model constructs a small but representative reference set that captures the structural patterns of an individual across frames. By relating each feature position to these reference nodes, the attention mechanism learns a more holistic representation of identity while reducing computational complexity.

MG-RAFA has demonstrated superior results on datasets such as Kinetics and UCF101, where it outperformed conventional methods in person re-identification and action recognition tasks. It captures relationships at different levels of granularity, using large regions for coarse details and smaller regions for finer details, allowing for adaptive and precise attention learning[81].

MG-RAFA significantly improves person re-identification through attentive mechanisms, yet integrating hierarchical spatiotemporal aggregation can address broader challenges, as we see in the following subsection.

PSTA: pyramid spatial-temporal aggregation. The pyramid spatial-temporal aggregation (PSTA) framework propsed by Wang et al.[82] also aimed at improving video-based person re-identification (ReID). PSTA aggregates frame-level features in a hierarchical manner to progressively fuse both short-term and long-term spatial-temporal information. This allows the model to handle challenges like occlusion, background clutter, and viewpoint variations more effectively. The framework also includes a spatial-temporal aggregation module (STAM), which uses spatial reference attention (SRA) and temporal reference attention (TRA) to enhance discriminative features and suppress irrelevant ones.

Similar approaches to PSTA have used contextual feature analysis to enhance the aggregation of spatial-temporal features. For instance, temporal context enhanced feature aggregation uses attention weights based on the temporal context of neighboring frames to refine features, effectively managing the temporal order and improving the detection of subtle changes over time[53].

PSTA is particularly effective in applications like sports video analysis, where maintaining temporal continuity is crucial for accurately detecting and tracking fast-moving objects. The inclusion of SRA, and TRA enhances discriminative features while suppressing irrelevant information, improving model performance.

The hierarchical approach of PSTA emphasizes both short and long-term features, while multi-correspondence attention mechanisms can further improve feature alignment, which is presented next.

MuCAN: multi-correspondence aggregation network for video super-resolution. Li et al.[83] proposed the multi-correspondence aggregation network (MuCAN) for video super-resolution, that aggregates features of multiple frames from a video with a multi-head cross-frame attention mechanism. The method efficiently aligns temporal and spatial features by alleviating artifacts such as occlusion and motion blur using attention on stable regions between frames. MuCAN strikes a balance between speed and accuracy through its lightweight backbone and enhanced attention mechanisms. It preserves fine details and learns global context by fusing local and global features from nearby frames. The method is evaluated on benchmark datasets like REDS and Vid4, showing that it is appropriate for high-resolution video applications. While attention-based techniques effectively address redundancy, motion-based aggregation approaches further improve accuracy by explicitly modeling the temporal dynamics of video content. The next subsection discusses the motion-based aggregation techniques in video analysis.

3.3 Motion-based feature aggregation

Motion-based feature aggregation methods explicitly incorporate motion information, such as optical flow or temporal shifts, into the aggregation process to accurately represent dynamic actions in video sequences. Recognizing and modeling motion allows these approaches to mitigate temporal inconsistencies and enhance the coherence of aggregated features across frames. By carefully considering movement patterns and object trajectories, motion-guided aggregation strategies substantially improve performance in tasks that rely heavily on accurate temporal alignment, as detailed in the subsections that follow.

Flow-guided feature aggregation (FGFA). Human actions in video sequences are three dimensional (3D) spatio-temporal signals which characterize visual appearance and dynamical motions of the involved objects[84]. However, because of the dynamical motions, the features of the same entity instance are not spatially compatible. A naïve features fusion may also inhibit the performance which implies that modeling the motion is very important in the procession of learning[85]. Also, camera motions are required to be analyzed in details suggested by recent works[86, 87].

In Ref. [86], Li et al. demonstrated that local spatio-temporal information is important to classify large-scale videos and CNN architectures can benefit from weakly-labeled data for powerful features. In particular, the spatio-temporal information astonishingly compact to connectivity of architectures in time sequence, which indicates the temporal information are connected motions play important roles in videos analysis. Flow-guided features aggregation (FGFA) leverages temporal coherence on the level of features[85]. At the same time, FGFA boosts performance of each frame via fusing neighboring features along the motion tracks, as shown in Fig. 11. Particularly, to generate the pre-frame feature maps, they apply data extractor into networks frame by frame. After that, the optimal flow networks are applied to measure the movements between the neighboring frames and the reference frames[88]. In their works, the feature maps of the neighboring frames are distorted by flow motion to the reference frames. The warped feature maps are then fused according to the adaptive weighting mechanism as well as their own feature maps on the reference frames. FGFA has been demonstrated that it can improve quality of features, and it would be complimentary to existing box-level frameworks for better performance[88–91]. Even though FGFA can achieve ideal performance, however, more lightweight networks.While FGFA effectively integrates neighboring frames based on optical flow, further refinements through fine-grained motion feature aggregation can significantly enhance detection accuracy, as illustrated subsequently.

Motion-based feature aggregation. Recent research has introduced innovative methods to capture motion information in videos. The motion feature aggregation (MFA) technique, developed by Ref. [92], combines coarse-grained motion features (CGMF) for drastic shifts and fine-grained motion features (FGMF) for small shifts in video. This approach has shown particular promise in video-based person re-identification tasks, effectively handling challenges such as occlusion and motion blur. The fine-grained motion learning (FGML) method within MFA acts as an attention mechanism to correctly identify moving parts, enhancing the model's ability to focus on the most informative aspects of motion.

Building upon motion-based approaches, recent studies suggest

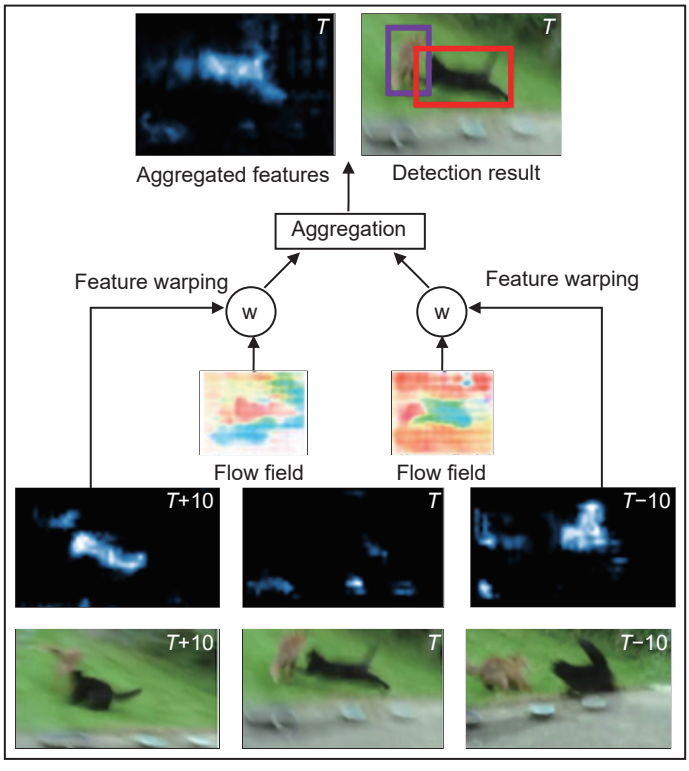

Fig. 11 Flow-guided features aggregation (FGFA). Feature activations at reference frame *T* are low, however, the nearby frames *T*–10 and *T*+10 have high activations. FGFA aggregates these three parts together to improve features of reference frames

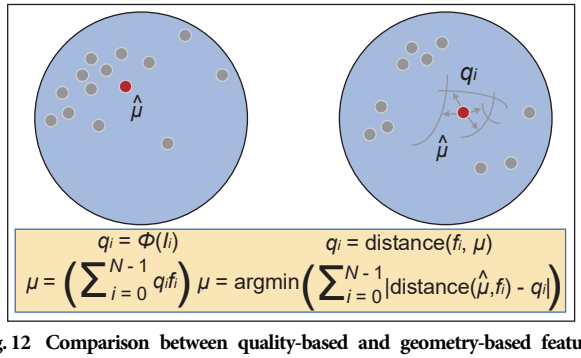

Fig. 12 Comparison between quality-based and geometry-based features aggregation strategies. Quality-based strategy, which is the left one, just considers the features of each frame, whereas the right geometry-based one considers both images quality and their geometric locations.

that transformer-based methods provide superior capabilities in modeling both short and long-range temporal dependencies without explicit convolutional operations which will be discussed next.

Video transformers for temporal aggregation. Computer vision researchers have developed transformer architectures tailored for video processing, understanding that video analysis is fundamentally different from static image work because you have to track how things change over time. Models such as video restoration Transformer (VRT), VidTr, and TimeSformer handle this by examining relationships between different parts of each frame, while simultaneously keeping track of how these visual elements shift and move as the video plays out. VRT incorporates an interesting combination of temporal reciprocal self-attention and parallel warping methods. Put simply, the model figures out which parts of different frames correspond to each other, then aligns them so they can work together more effectively. TimeSformer uses a completely different approach - it processes spatial information and temporal information in separate steps during its attention operations, making it much more efficient when dealing with several video segments simultaneously. When researchers test these methods on practical applications - whether that's figuring out what actions are happening in a video or cleaning up poor-quality footage both techniques deliver strong performance[42,93].

State-space models for video aggregation. Video-focused state-space models like MambaVLT and VideoMamba have adapted efficient linear-state processing methods to handle the challenge of connecting information across long video sequences. These systems continuously update their internal representations as they move through both the temporal dimension (time) and spatial dimensions (within frames). The clever part about this architecture is that it can handle full-length video clips without running into the massive computational bottlenecks that plague other approaches when dealing with longer video sequences.

Foundation video models for aggregation. Large-scale foundation models designed for video include VideoMAE and modified versions of TimeSformer. These models employ masked auto-encoding training approaches and combined space-time attention mechanisms. This combination enables them to understand videos effectively even without task-specific training, making them valuable for transfer learning applications where you need to adapt to new video analysis tasks. We contrast representative video aggregators and their trade-offs in Table 6.

3.4 Transformer-based aggregation techniques

Transformer-based aggregation techniques represent a significant evolution beyond conventional convolutional methods by leveraging self-attention mechanisms to capture long-range spatiotemporal dependencies across video frames. Unlike traditional approaches that often struggle with complex motions and misalignments, transformer architectures excel at dynamically modeling interactions within and across frames, adapting flexibly to temporal and spatial variations.

Table 6 Video aggregator comparison

Model/method	Temporal span	Mechanism	Pros	Cons
VRT, VidTr	Whole clip	Global attention; warping	High accuracy; flexible	Memory; complexity
TimeSformer	Full video	Divided space/time attention	Efficient; multi-scale	Needs pretraining
Mamba / State-space	Long clip	Latent state updates	Scalable; robust over frames	New; less explored
Foundation (VideoMAE)	Global	Masked spatio-temporal fusion	Zero-shot; adaptable	Large model; heavy HW

These methods provide robust and precise aggregation capabilities, especially beneficial in video restoration and analysis tasks, where maintaining accurate feature alignment is critical. The following subsections illustrate prominent transformer-based strategies, highlighting their effectiveness in addressing common challenges in video processing tasks.

Transformer-based aggregation: the VRT approach. Video restoration aims to restore high-quality (HQ) frames from low-quality (LQ) video sequences. Traditional methods, such as sliding window-based or recurrent architectures, face limitations such as inefficiency of sliding window methods, limited Temporal Modeling and handling misaligned frames due to motion or blur.

Transformer-based aggregation: the VRT approach. Video restoration aims to restore HQ frames from LQ video sequences. Traditional methods, such as sliding window-based or recurrent architectures, face limitations such as inefficiency of sliding window methods, limited Temporal Modeling and handling misaligned frames due to motion or blur. Liang et al.[45] propsed a video restoration Transformer or VRT, a multi-scale method that employs temporal reciprocal self-attention (TRSA) and parallel warping to extract information, align information across frames, and fuse information across frames, respectively, for parallel computation. The proposed method is employed on a range of video restoration problems: video super-resolution, deblurring, denoising, frame enhancement, and space-time super-resolution.

VRT introduced the TRSA module, which divides the video into small clips and applies reciprocal attention to align and fuse information across frames. This mechanism ensures that features from different frames are effectively integrated, improving motion estimation and feature alignment.

Another crucial component of VRT is parallel warping, which further enhances frame alignment by adapting features from neighboring frames without relying solely on traditional optical flow-based warping techniques. This approach helps mitigate motion artifacts and preserves fine details across frames.

The VRT framework successfully addresses frame alignment challenges; however, convolution-free transformers offer an alternative perspective, which is discussed next.

VidTr: Video Transformer without convolutions. Some reserchers introduced enhancements to temporal aggregation that employ temporal filtering with adaptive gating. This approach dynamically weights different temporal segments based on their contribution to action detection. The use of adaptive gating has proven effective in video classification and recognition tasks, where different parts of a video may contribute unevenly to the overall classification objective[45].

VidTr represents a novel transformer-based approach that removes convolutional operations and instead uses spatio-temporal separable-attention to model video data[94]. By aggregating spatial and temporal features via self-attention, VidTr can handle challenges like motion blur and occlusion more effectively compared to traditional CNNs. This makes it an efficient alternative for capturing long-range dependencies across video sequences.

VidTr leverages self-attention layers to model spatiotemporal dependencies without relying on convolutions, offering an efficient alternative to conventional CNN-based approaches. Other transformer-based models that eliminate convolutions have integrated hybrid attention mechanisms. These mechanisms combine spatial and temporal attention in a way that enhances the model's ability to differentiate between moving objects and background noise. This hybrid attention-based approach improves overall accuracy in video object detection, particularly in challenging environments with high motion blur or clutter[53].

VidTr leverages scalable attention mechanisms to effectively aggregate features across video frames without downsampling, preserving important fine-grained details. This allows VidTr to achieve more efficient feature aggregation for tasks such as video object detection and image retrieval, focusing on capturing long-range dependencies between frames[94].

VidTr provides efficiency by entirely removing convolutions, yet integrating guided deformable attention can further optimize performance in restoration tasks, explored in the following subsection.

Video restoration Transformer with guided deformable attention. Recurrent video restoration Transformer (RVRT)[95], a cutting-edge framework designed for video restoration tasks like super-resolution, denoising, and deblurring. The architecture leverages recurrent feature refinement (RFR) modules and guided deformable attention (GDA) for efficient feature aggregation and alignment between neighbouring video clips. The proposed RVRT combines the benefits of both recurrent and transformer models, enabling the processing of local neighbouring frames in parallel while propagating global information recurrently.

RVRT addresses challenges in video restoration, such as temporal correspondence and alignment, by using GDA, which dynamically aggregates information from multiple locations across video clips. This results in better handling of large motions and complex degradations. Extensive evaluations on benchmark datasets demonstrate the state-of-the-art performance of RVRT, balancing computational efficiency and accuracy, and showing its robustness across various video restoration tasks.

Transformer-based methods have significantly advanced feature aggregation; however, temporal-based aggregation techniques, which are discussed next, focus specifically on leveraging temporal dependencies to enhance motion-sensitive tasks, such as action recognition and video super-resolution.

3.5 Temporal-based feature aggregation techniques

Temporal-based feature aggregation techniques specifically emphasize capturing meaningful temporal relationships and dependencies across multiple video frames. By integrating both short-term and long-term temporal contexts, these methods aim to model dynamic changes effectively, providing more accurate and stable representations across sequences. Particularly beneficial for applications like video super-resolution, action recognition, and object detection, temporal-based strategies ensure that feature aggregation accurately reflects temporal coherence and variations.

The following subsections detail advanced methods utilizing adaptive and deformable temporal aggregation, demonstrating significant improvements over traditional techniques.

Adaptive spatio-temporal feature aggregation for video super-resolution using deformable 3D convolutions. Traditional video super-resolution (SR) methods often rely on sequential motion compensation and spatial feature extraction, which limits their ability to effectively utilize spatio-temporal dependencies. Many approaches employ optical flow-based motion estimation, but inaccuracies in motion alignment can lead to artifacts and a lack of temporal coherence in reconstructed video frames. Additionally, standard 3D convolutional networks, while capable of capturing spatial and temporal relationships, struggle with fixed receptive fields, making them inadequate for handling complex motion variations across frames.

To address these challenges, Ying et al.[96] proposed the deformable 3D convolution network (D3Dnet), which integrates deformable convolution with 3D convolution to enable both motion-aware feature extraction and spatio-temporal modeling in a unified manner. Unlike standard 3D convolution, which operates on a fixed sampling grid, deformable 3D (D3D) convolution adaptively adjusts its receptive field using learnable offsets, allowing the network to focus on relevant regions with greater flexibility. This method enhances both motion compensation and spatial feature extraction by dynamically adapting to variations in object motion. The proposed framework processes a sequence of video frames through residual deformable 3D convolution blocks, ensuring efficient information fusion across time. The model is trained using a mean square error (MSE) loss function to minimize the difference between the predicted super-resolved frames and the ground truth. By fully leveraging spatio-temporal dependencies in a more adaptive manner, the proposed method significantly enhances the clarity and coherence of video super-resolution results.

Adaptive deformable 3D convolutions enhance video super-resolution by flexibly modeling spatiotemporal relationships, yet explicitly modeling temporal excitation can further refine action recognition, as detailed next.

TEA: temporal excitation and aggregation. TEA-temporal excitation and aggregation for action recognition[97] is a framework that introduces a novel method to improve temporal modeling for action recognition in videos. The approach comprises two key modules, the motion excitation (ME)module which captures short-range motion by calculating temporal differences between adjacent frames at the feature level, then selectively enhancing motion-sensitive channels in the spatiotemporal features. This improves motion encoding without relying on computationally expensive optical flow and the multiple temporal aggregation (MTA) module used to model long-range temporal relationships, this module extends local convolutions through hierarchical residual architecture. This allows features to be aggregated across multiple frames, enlarging the receptive field and enabling the capture of long-range temporal dynamics without adding extra computational complexity.

Together, these modules form the TEA block, which is inserted into a ResNet backbone to create an efficient and effective model for action recognition. The TEA network shows good performance across several benchmark datasets such as Kinetics, Something–Something, HMDB51, and UCF101, achieving competitive results at lower computational costs. Figure 13 shows the TEA framework in which the sparse sampling strategy is adopted to sample T frames from videos. The 2D ResNet is utilized as the backbone, and the ME and MTA modules are inserted into each ResNet block to form the TEA block. The simple temporal pooling is applied to average action predictions for the entire video. The TEA network demonstrates a significant improvement in both short- and long-range temporal modeling, making it essential for video-based tasks like intelligent surveillance, autonomous driving, and entertainment.

While TEA efficiently captures both short-range and long-range temporal relationships for action recognition, further enhancing temporal context through attention-driven frame selection can significantly boost the performance in video object detection tasks, as detailed in the subsequent approach.

Temporal context enhanced feature aggregation for video object detection. He et al.[98] proposed a temporal context enhanced aggregation approach in their TCENet model, which aggregates features from adjacent frames with consideration of their temporal context. This approach enhances the features of the reference frame using attention weights and corrects temporal order to help surmount challenges like appearance changes and occlusions, especially for objects in rare poses or partially

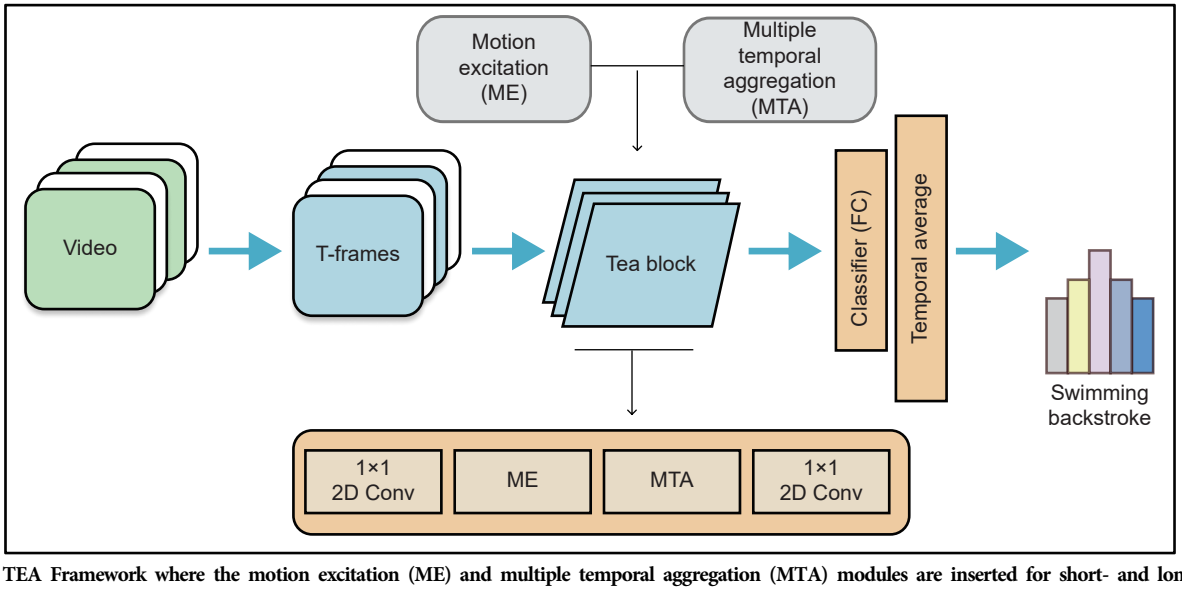

Fig. 13 TEA Framework where the motion excitation (ME) and multiple temporal aggregation (MTA) modules are inserted for short- and long-range temporal modeling.

occluded. The model employs a dynamically selecting temporal stride predictor that selects frames to aggregate based on temporal dynamics for effective and non-redundant feature extraction.

Furthermore, TCENet also comes with the DeformAlign module, which deforms features across frames at the pixel level precisely to address motion shifts in aggregating features. Over-aggregation is avoided by only summing up critical frames, lowering computational expense. The model is evaluated using the ImageNet VID dataset, employing mean average precision (mAP) as the primary metric, to demonstrate its ability to improve video object detection performance.

TCENet significantly enhances temporal feature aggregation through dynamically selecting and aligning frames. Nevertheless, integrating multi-level aggregation techniques that combine short-term details with long-term context, as illustrated by CompFeat, provides further performance improvements in tasks such as video instance segmentation.

CompFeat: comprehensive feature aggregation for video instance segmentation. Fu et al.[99] introduced CompFeat, a comprehensive feature aggregation framework tailored for video instance segmentation. The framework incorporates a multi-level temporal aggregation module, which fuses short-term spatial details with long-term temporal context to enhance segmentation performance. By integrating feature alignment techniques, CompFeat ensures consistency across frames, addressing challenges like motion blur and occlusion. Additionally, adaptive temporal attention dynamically prioritizes informative frames, reducing noise and redundant features for optimal segmentation.

CompFeat is evaluated on YouTube-VIS and Cityscapes-VPS benchmarks, demonstrating superior segmentation accuracy while maintaining computational efficiency through its lightweight design. The model leverages pre-trained backbones like ResNet for feature extraction, further refining its feature aggregation capabilities. While primarily focused on video instance segmentation, its methodology could extend to broader applications, such as surveillance and autonomous systems, where robust temporal feature aggregation is crucial.

While temporal aggregation techniques efficiently capture short-term and long-term temporal dynamics, the integration of memory-based global context provides a broader semantic perspective, as illustrated by the MEGA approach in the subsequent section.

3.6 Memory enhanced global-local aggregation (MEGA)

Traditional methods often struggle with aggregating features due to motion blur and occlusion[100]. Memory enhanced global-local aggregation (MEGA) network overcomes this by aggregating features from not only just local frames, but also global semantic cues. Introduction of the long-range memory module enables a much broader and extensive aggregation process by which key frames are allowed to access information from previous frames already processed. This technique helps to tackle problems such as motion blur and occlusion through improved feature localization, thereby enhancing overall detection accuracy.

Moreover, the MEGA network draws on the power of temporal information across frames to improve video object detection. Different from most methods that are based on just a few reference frames, MEGA explores larger temporal contexts, reusing more precomputed features stored in LRM and thereby enabling the network to build much richer feature representations.

Although MEGA significantly improves detection accuracy by considering broader temporal contexts, addressing redundancy in feature aggregation through structural and semantic similarity metrics, as seen in R-CNN based methods, further enhances object detection performance.

3.7 Feature extraction in video object detection using R-CNN

In video sequences, moving objects frequently experience distortions, including changes in shape, occlusion by other objects, and blurring due to fast motion. These factors significantly reduce the accuracy of traditional object detection models, which primarily rely on static image-based feature extraction techniques. Additionally, a major challenge in video object detection is the redundancy present in consecutive frames, leading to excessive and unnecessary computations when aggregating frame features. Traditional object detection models often use adjacent frames for feature aggregation, which may result in redundant calculations without effectively improving detection performance. To overcome these challenges You et al.[56] presented improved faster R-CNN framework integrated with feature pyramid networks (FPN) and dynamic region aware convolution (DRConv) to enhance feature extraction capabilities. The authors proposed the structural similarity-based sequence level semantics aggregation (S-SELSA) module, an advanced feature aggregation technique that combines semantic and structural similarity metrics to improve object detection performance. Unlike conventional approaches that rely solely on temporal proximity for feature aggregation, S-SELSA selects frames with the highest similarity based on semantic attributes and feature map correlations, thereby reducing redundancy and improving detection accuracy.

R-CNN approaches effectively mitigate redundancy through intelligent frame selection; however, explicitly extracting and aggregating features from keyframes for summarization purposes offers a complementary and efficient way to reduce redundancy, as discussed next.

3.8 Keyframe extraction and feature-based aggregation for video summarization

Recent advancements in unsupervised video summarization leverage deep learning-based feature extraction and clustering techniques to identify keyframes while minimizing redundancy. Jadon and Jasim[102] proposed an unsupervised video summarization framework that extracts keyframes using a combination of traditional vision-based filters (SIFT, histograms) and deep learning embeddings (ResNet16).

The proposed framework employs a combination of keyframe extraction techniques and video skimming to generate more fluid and human-friendly summaries. Initially, the video is decomposed into frames, and various feature extraction methods such as uniform sampling, image histograms, scale-invariant feature transform (SIFT), and deep learning-based ResNet16 embeddings are applied to identify visually significant frames. To group similar frames and select the most representative ones, clustering methods like K-Means and Gaussian mixture models (GMMs) are utilized. Instead of abrupt transitions, the framework incorporates video skimming, where short sequences around the selected keyframes are retained to ensure smoother transitions, making the summary more coherent.

While keyframe extraction optimizes summarization through selective aggregation, dynamically adapting aggregation strategies based on motion and spatial deformation provides enhanced accuracy for video object detection, as elaborated in the subsequent dynamic aggregation method.

3.9 Dynamic feature aggregation for video object detection

Cui[103] presents dynamic feature aggregation (DFA) in the paper,

an approach to increasing video object detection efficiency at a low amount of computation. DFA contains two components: vanilla dynamic aggregation (VDA), where the frames are picked from motion, and deformable dynamic aggregation (DDA), where the frames are deformed based on the situation. This approach circumvents over-aggregation by putting a primary focus on keyframes, especially in cases involving motion blur. The method also utilizes pre-trained models like SELSA and FGFA for better spatial-temporal information acquisition. DFA is established to be effective on the ImageNet VID benchmark, indicating its applicability for real-world applications of video detection.

4 Discussion

Feature aggregation is important in both image and video analysis for enhancing model performance through the complementary information aggregation from multiple layers. While the principles of feature integration apply similarly to both domains, they differ considerably when it comes to complexity of temporal processing, redundancy handling, and architecture designs. This section focuses on image-on-video relationships of the feature aggregation; similarities in their mechanisms and differences between image and video feature aggregation.

4.1 Relationships between image and video for feature aggregation

Currently, we have discussed features aggregation strategies of images and videos separately. We summarize the key video-based methods, their temporal spans, and their performance on standard benchmarks in Table 7. Besides, we also displayed detailed discussions of them from four different aspects, as shown in Tables 5 and 7.

Similarities between image and video feature aggregation. As we mentioned before, since videos consist of lots of frames and each frame can be regarded as one image of specific objects, so we can cast videos as templates (collections of images of same objects). Based on that, this means we can analyze videos using the strategies applied on images.

For images, many approaches have been proposed to fuse extracted features into more representative and noise-resistant forms. These include leveraging complementary local and global features[13], combining max pooling (for salient features) with average pooling (for noise reduction)[27], and restricting aggregation to highly activated regions[33, 34]. However, from the perspective of videos, one of the most important things is how to better promote the speed and accuracy of videos tracking or detection using the redundancies among different frames. Currently, lots of works focus on the approaches of noise reduction which are treated as ideal ways to handle redundancies, like attention-based weights learning schemes[68], 1×1 convolutional layer with average pooling[80], and geometry-aware quality metric mechanism[79] and so on.

As we discussed at the beginning, we can apply the strategies of still images into the pre-processing of videos to augment the quality of final representations. Also, the noisy reduction approaches of videos analysis can be used in images recognition before the steps of features fusion. In addition, videos tracking or

Table 7 Comparison of current feature aggregation strategies in videos

Model	Family/strategy	Temporal span	Dataset	Metric	Representative performance
QAN[67]	Quality-aware network (learned quality scores)	Whole video / set-level	IJB-A / YTF / iLIDS-VID / PRID2011	Accuracy / Rank-1	96.2% YTF, +14.6% Rank-1 on iLIDS-VID
C-FAN[69]	Component-wise feature aggregation	Whole video / template	IJB-S / YTF / IJB-A	Verification, Rank-1	96.5% YTF, +3%–5% over average pooling
NAN[68]	Neural aggregation network (attention blocks)	Whole video / set-level	IJB-A / YTF / Celebrity-1 000	TAR / Rank-1	95.7% YTF, TAR@FAR=0.01 = 0.933 (IJB-A)
FAN[80]	Multi-layer CNN feature aggregation (inter + intra feature)	Per-frame (real-time tracking)	OTB50 / OTB100	Precision / success	79.0% precision (OTB100), 24 FPS
FGFA[85]	Flow-guided feature aggregation (optical flow + adaptive weights)	Short–mid (neighbor frames)	ImageNet VID	mAP	76.3% mAP (ResNet-101), +2.9% over baseline
GFA[79]	Geometry-guided feature aggregation	Whole video / template	IJB-A / YTF	Verification, Rank-1	97.2% YTF accuracy, superior to quality pooling
MG-RAFA[81]	Attention (multi-granularity reference aided)	Mid/long sequence	MARS	Rank-1 / mAP	90.0% R1, 85.9% mAP
PSTA[82]	Pyramid spatial-temporal attention	Hierarchical (short–long)	MARS / DukeMTMC-VID	Rank-1	91.5% R1 (MARS), 98.3% (Duke)
SELSA[56]	Sequence-level semantic aggregation (global context)	Global across video	ImageNet VID	mAP	80.5% mAP (ResNet-101)
MEGA[100]	Memory-enhanced global-local aggregation	Global + local + long memory	ImageNet VID	mAP	85.4% mAP (ResNeXt-101), best to-date
DFA[103]	Dynamic feature aggregation (vanilla + deformable)	Adaptive (slow vs. fast motion)	ImageNet VID	mAP / FPS	46.6% mAP with 31%–76% speedup vs FGFA/SELSA
TEA[97]	CNN + temporal excitation/aggregation	Short + long	Kinetics-400	Top-1 Acc	76.1% Top-1
RVRT[95]	Transformer (recurrent + GDA)	Long, clip-to-clip	Vimeo-90K	PSNR / SSIM	37.2 dB PSNR
VidTr[94]	Transformer (separable attention)	Long global	Kinetics-400	Top-1 Acc	79.1% Top-1
VRT[45]	Transformer (multi-scale TRSA + warping)	Long	REDS / Vimeo-90K	PSNR / SSIM	+2.1 dB over SOTA
D3Dnet[96]	Deformable 3D convolution	Short (7 f)	Vid4	PSNR	28.5 dB PSNR

detection benefit from one phenomenon that videos can be cast as one template, so we also can adopt some similar ways to augment the diversity of still images, like data augmentation, data segmentation and data transfer learning[104–106].

Despite the differences in temporal structure, image and video feature aggregation techniques share several foundational principles. Both domains aim to extract and combine spatially and semantically rich features to generate more discriminative representations. For example, in image analysis, strategies like DCMA-CNN and FACNN aggregate global and local features or intermediate layer outputs to enhance classification performance. Similarly, in videos, attention-based models like MG-RAFA and PSTA focus on selecting discriminative spatial and temporal features across frames. Another commonality is the use of pooling mechanisms such as max pooling, average pooling, or learned pooling strategies to reduce noise and dimensionality. Whether applied to image patches or individual video frames, these pooling operations help suppress irrelevant information and highlight the most informative signals. Additionally, transformer-based approaches like ViT for images and VRT or VidTr for videos showcase how attention mechanisms can be generalized to both spatial and temporal domains.

Differences between image and video feature aggregation. As shown in Fig. 14, we also present one overview of differences between image feature aggregation and video features aggregation. Firstly, since videos consist of lot of frames, how to balance the features from low quality frames and high quality frames is very important. As a result, the final fused features will include discriminative information as much as possible and simultaneously reduce noise to most extent. Secondly, videos analysis does not only require accuracy, but the models also can track the objects in real-time way. In another word, some aggregation schemes that need minimal memory and perform efficiently are necessary for videos tracking tasks. Nevertheless, there is no need for still images analysis to consider memory consuming and real-time detecting in most situations.

The most significant difference between image and video feature aggregation lies in the temporal dimension. Image aggregation focuses purely on spatial information extracted from static frames. In contrast, video aggregation must handle temporal redundancy, motion alignment, and frame quality variation, requiring more complex modeling of dependencies across time.

While image aggregation strategies like NetVLAD and EMR encode spatial relationships from fixed feature maps, video approaches like FGFA, TEA, and MuCAN must account for object motion, optical flow, and variable frame quality over time. Moreover, real-time constraints are more prominent in video tasks—demanding lightweight and memory-efficient architecture such as MEGA or CompFeat. Video aggregation often involves long-term memory and attention to temporal continuity; a necessity rarely encountered in static image tasks. Therefore, while the concept of feature fusion is shared, video-based models must go beyond spatial context and address dynamic environmental factors, making their aggregation strategies inherently more complex and application-specific. Finally, although we can augment our images data to enhance the diversity, however, compared to videos computation, we don't necessarily consider the redundancies, blur problems, out of focus in images recognition.

5 Current challenges and future directions

Although features aggregation on images and videos analysis have achieved a series of remarkable results, the studies are still in the primary stage where various challenges need to be solved. Firstly, input data of images recognition does not only lack diversity, but their quantity and quality are also low. Secondly, much noise is still included in the fused features during procession of the

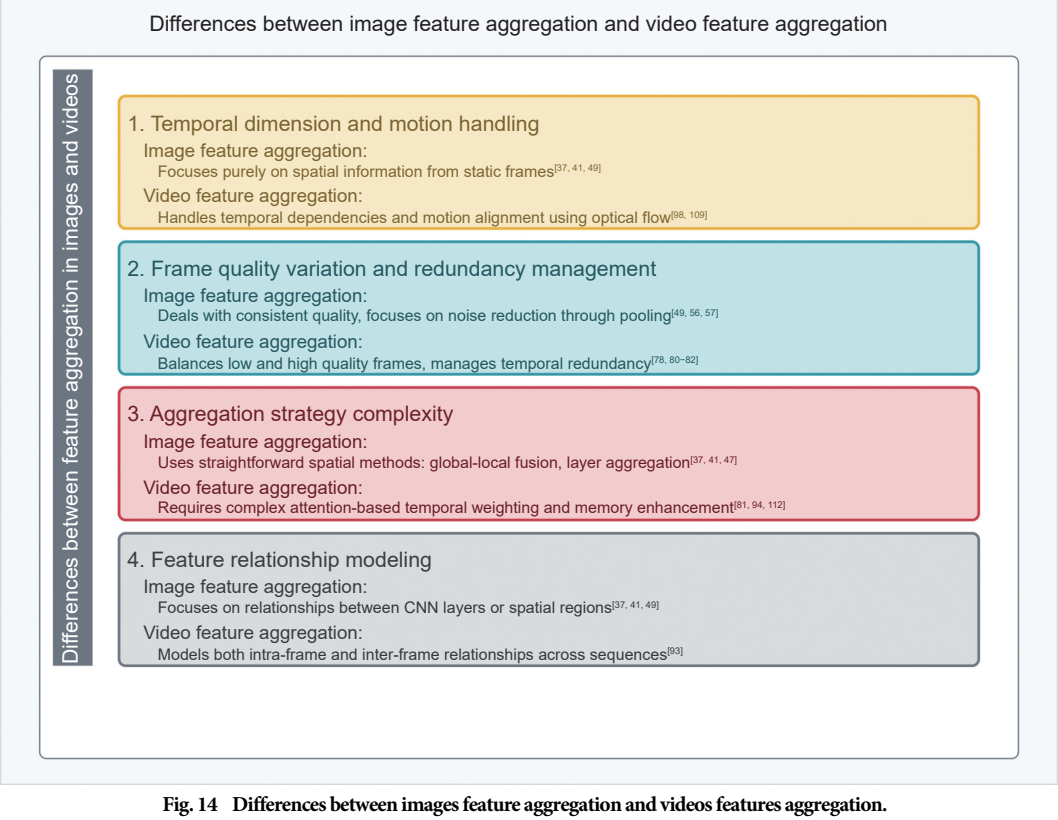

Fig. 14 Differences between images feature aggregation and videos features aggregation.

features integration. Thirdly, how to distinguish useful features which can be aggregated into the final representations is still challenging. In addition, how to implement features fusion strategies more efficiently with less memory consuming in videos tracking is still waiting to be solved. In this chapter, we propose some solutions to these challenges.

5.1 Quantity, diversity, and quality of the input data on images recognition

Currently, due to the large capacity of the deep learning methods, we can achieve models with satisfied performance using similar networks and general training parameters. In another word, the most critical factor that determines the performance of models is not the network parameters and various hyperparameters, but the data. What kinds of data we feed to the networks, what kinds of features the networks will learn. Data quantity or diversity, as we know, can ensure the generalization ability of models. However, data of good quality can promote the models to extract features with high accuracy. As a result, the quantity, diversity and quality of data are the essential conditions for models to be robust and accurate. However, the methods proposed above, they only concentrate on how to aggregate features from different layers to improve performance of models without considering the problems of data including quantity, diversity and quality.

A consistent limitation across existing works is the reliance on large, high-quality, and well-annotated datasets. In practice, datasets often suffer from class imbalance, label noise, or domain bias (e.g., lighting variations, occlusion, resolution shifts), especially in video datasets that span different environments or subjects. For instance, video aggregation techniques such as FGFA or TEA rely heavily on well-aligned frame sequences. Any distortion in data, due to jitter, motion blur, or camera shifts, can drastically reduce performance.

Furthermore, image-based methods that fuse global and local features (e.g., DCMA-CNN, FACNN) may overfit to dominant object patterns or backgrounds in the training data. To overcome this, self-supervised learning, domain generalization, and synthetic data augmentation techniques (e.g., MoCo v3, DINO) should be integrated more deeply into feature aggregation pipelines. These approaches can reduce dependency on human-labeled data while improving model robustness across diverse real-world conditions.

To solve the problems, we point out some approaches to promote the data diversity and quantity which are presented from two aspects: one part is to handle the input data directly, the other part is to deal with the architecture of models. Data augmentation[104], one solution to the problems of limited data, aims to help prevent overfitting and other affections. As we know, geometric transformations, random erasing, color space augmentations, features space augmentation, neural style transfer and meta-learning are used in data augmentation most frequently[107]. After applying augmentation methods, more comprehensive data will be generated which can prevent overfitting, increase generalization capabilities, presumably regularize models, balance classes and so on to help promote performance of models.

5.2 Noise reduction in the fused features

With the development of objects recognition with deep learning, people propose that the noise included in images or videos will seriously affect the performance of models. As Ref. [108] mentioned, in real-life applications, researchers have to face noise with the images themselves, like low resolution, bad illumination or combination of different factors, etc. As a result, much attention has been paid to noise reduction in object detection. As we mentioned before, lots of approaches have been used to reduce noise during the processing of features integration, like adapting weights mechanism to aggregate features with high weights, max pooling to obtain more semantic information, or average pooling to make features become more smoothed and so on. However, all of them only focus on how to efficiently aggregate features extracted from various layers and which features can be aggregated together with simple and basic technologies.

Based on the discussions above, we propose that applying semantic segmentation or instance segmentation as a natural step in the progression of objects analysis will reduce affections of noise from the root. Generally speaking, most of images or videos will not only include the objects we want to detect, but they also will contain other objects which will generate interference on our models. However, as we know, semantic segmentation is originally located in classification tasks used to predict which are the target objects in one image[105]. And it is able to segment all the objects in one image and locate the objects we need, such as the experiments stated in Refs. [109–112]. Therefore, it will help improve noise reduction in feature aggregation and promote the models to achieve better performance.

The majority of the Section 2 models (e.g., multi-level aggregation, spatial attention-based fusion, and intermediate feature combination techniques) tend to suppress noisy or irrelevant features naturally through purifying representations layer by layer. Yet they may rely only on pooling operations or attention weights, possibly not sufficient to remove all redundancy or low-quality activations. Future research can continue to enhance noise reduction by implementing more dynamic filtering methods that adapt to content-based image features during the process of fusion.

5.3 Features interpretability

Currently, CNNs have achieved superior performance in the field of computer vision and patterns, like object detection and classification. With continuous in-depth study on CNNs, lots of research demonstrate that low-level layers contain detailed information, whereas high-level layers include more semantic and abstract features people can not understand. Current works have already applied these acknowledgements to obtain better performance. However, they still do not have a clear perception that what features of high-level layers the networks learn, which ones promote their networks, which ones affect their fused features and so on.

For example, while models like EMR, S-SELSA, or VRT aggregate features effectively, they do not explicitly explain the contribution of each feature map or temporal slice. Recent methods have started incorporating attention visualization, saliency maps, and activation clustering, but these are often post-hoc and not inherently integrated into the training pipeline.

Therefore, features interpretability will help us improve features fusion strategies. Till now, more and more attention has been paid on features interpretability[113, 114] and also, lots of approaches have been proposed, such as saliency map and activation maximization[115, 116]. Future research should focus on embedding interpretability modules directly within aggregation architectures. This includes attention heatmaps that dynamically evolve with training, semantic segmentation overlays of important regions, and dimensionality-reduced embeddings that highlight critical transitions in videos or structural cues in images. With the help of methodologies of features interpretability, we strongly believe future research will prevent missing useful features and include

noise in the procession of features aggregation.

5.4 Detection efficiency and memory consumption on videos detection

Video-based detection tasks inherently demand significant computational resources due to the high dimensionality of input data and the temporal dependencies between frames. Unlike image-based detection, where inference is conducted on a single frame, video analysis must process multiple frames, often hundreds per second, requiring not only efficient feature aggregation, but also optimized memory utilization. To generate a compact representation of videos, an efficient aggregation scheme that needs minimal memory and performs efficiently is necessary. Currently, most research analyze videos via casting videos into variety of templates. Also, some of them improve their models using attention mechanisms, adaptive weights mechanism or others techniques to compute the relationships among different frames. However, since the approaches of videos analysis have to compute features frame by frame, so it also results in consuming too much memory and bringing down the efficiency of models to some extent. To improve detection efficiency, future research should focus on advancements of lightweight architectures with selective attention mechanisms or memory-constrained processing. This not only reduces processing time but also curtails GPU/TPU memory usage during inference.

6 Conclusions

In this article, we conduct a comprehensive review of feature aggregation strategies in the field of images and videos analysis. We provide detailed comparisons among various approaches of these strategies applied on them individually. Then, we provide evaluations and comparisons of feature integration methods between images and videos computation in details. Finally, we suggest four promising future directions in the area of feature aggregation.

Acknowledgements

This research was supported in part by the U.S. National Science Foundation under Grant No. CNS-2244450.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Article History

Received: 2 January 2025; Revised: 7 October 2025; Accepted: 18 December 2025

References

[1] Zhao, B.; Feng, J.; Wu, X.; Yan, S. A survey on deep learning-based fine-grained object classification and semantic segmentation. *International Journal of Automation and Computing* Vol. 14, No. 2, 119–135, 2017.

[2] Donahue, J.; Jia, Y.; Vinyals, O.; Hoffman, J.; Zhang, N.; Tzeng, E.; Darrell, T. DeCAF: A deep convolutional activation feature for generic visual recognition. In: Proceedings of the 31st International Conference on Machine Learning, 647–655, 2014.

[3] Yu, F.; Wang, D.; Shelhamer, E.; Darrell, T. Deep layer aggregation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2403–2412, 2018.

[4] Zeiler, M. D.; Fergus, R. Visualizing and understanding convolutional networks. *arXiv preprint* arXiv: 1311.2901, 2014.

[5] Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3431–3440, 2015.

[6] Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How transferable are features in deep neural networks? In: Proceedings of the Advances in Neural Information Processing Systems, 3320–3328, 2014.

[7] Yu, F.; Koltun, V.; Funkhouser, T. Dilated residual networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 636–644, 2017.

[8] Yang, J.; Yang, J. Y.; Zhang, D.; Lu, J. F. Feature fusion: Parallel strategy vs. serial strategy. *Pattern Recognition* Vol. 36, No. 6, 1369–1381, 2003.

[9] Arandjelovic, R.; Gronat, P.; Torii, A.; Pajdla, T.; Sivic, J. NetVLAD: CNN architecture for weakly supervised place recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5297–5307, 2016.

[10] Psomas, B.; Kakogeorgiou, I.; Karantzalos, K.; Avrithis, Y. Keep it SimPool: Who said supervised transformers suffer from attention deficit? In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 5327–5337, 2024.

[11] Muralidhara, S.; Hashmi, K. A.; Pagani, A.; Liwicki, M.; Stricker, D.; Afzal, M. Z. Attention-guided disentangled feature aggregation for video object detection. *Sensors* Vol. 22, No. 21, Article No. 8583, 2022.

[12] Tschannen, M.; Djolonga, J.; Ritter, M.; Mahendran, A.; Houlsby, N.; Gelly, S.; Lucic, M. Self-supervised learning of video-induced visual invariances. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 13803–13812, 2020.

[13] Xie, S.; Hu, H. Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks. *IEEE Transactions on Multimedia* Vol. 21, No. 1, 211–220, 2019.

[14] Dai, Y.; Liu, X.; Dong, S.; Yang, L. Group emotion recognition based on global and local features. *IEEE Access* Vol. 7, 111617–111624, 2019.

[15] Zhou, Y.; Liu, Y.; Han, G.; Zhang, Z. Face recognition based on global and local feature fusion. In: Proceedings of the IEEE Symposium Series on Computational Intelligence, 2771–2775, 2019.

[16] Yan, K.; Mei, S.; Ma, M.; Yan, F. Fusing deep local and global features for remote sensing image scene classification. In: Proceedings of the IGARSS - IEEE International Geoscience and Remote Sensing Symposium, 3029–3032, 2019.

[17] Yin, J.; Ma, Z.; Xie, J.; Nie, S.; Liang, K.; Guo, J. Dual-granularity feature alignment for cross-modality person re-identification. *Neurocomputing* Vol. 511, No. C, 78–90, 2022.

[18] Lin, B.; Zhang, S.; Yu, X. Gait recognition via effective global-local feature representation and local temporal aggregation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 14628–14636, 2022.

[19] Lu, X.; Sun, H.; Zheng, X. A feature aggregation convolutional neural network for remote sensing scene classification. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 57, No. 10, 7894–7906, 2019.

[20] Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1–9, 2015.

[21] Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C. Y.; Berg, A. C. SSD: Single shot MultiBox detector. In: *Computer Vision – ECCV 2016. Lecture Notes in Computer Science, Vol. 9906*. Leibe, B.; Matas, J.; Sebe, N.; Welling, M. Eds. Springer Cham, 21–37, 2016.

[22] Cheng, G.; Yang, C.; Yao, X.; Guo, L.; Han, J. When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs. *IEEE Transactions on*

Geoscience and Remote Sensing Vol. 56, No. 5, 2811–2821, 2018.
[23] Liang, Y.; Monteiro, S. T.; Saber, E. S. Transfer learning for high resolution aerial image classification. In: Proceedings of the IEEE Applied Imagery Pattern Recognition Workshop, 2017.
[24] Li, E.; Xia, J.; Du, P.; Lin, C.; Samat, A. Integrating multilayer features of convolutional neural networks for remote sensing scene classification. *IEEE Transactions on Geoscience and Remote Sensing* Vol. 55, No. 10, 5653–5665, 2017.
[25] Hu, F.; Xia, G. S.; Hu, J.; Zhang, L. Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. *Remote Sensing* Vol. 7, No. 11, 14680–14707, 2015.
[26] Wang, G.; Fan, B.; Xiang, S.; Pan, C. Aggregating rich hierarchical features for scene classification in remote sensing imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* Vol. 10, No. 9, 4104–4115, 2017.
[27] Cho, J. H.; Park, C. G. Multiple feature aggregation using convolutional neural networks for SAR image-based automatic target recognition. *IEEE Geoscience and Remote Sensing Letters* Vol. 15, No. 12, 1882–1886, 2018.
[28] Lee, J.; Nam, J. Multi-level and multi-scale feature aggregation using pretrained convolutional neural networks for music auto-tagging. *IEEE Signal Processing Letters* Vol. 24, No. 8, 1208–1212, 2017.
[29] Tolias, G.; Sicre, R.; Jégou, H. Particular object retrieval with integral max-pooling of CNN activations. *arXiv preprint* arXiv: 1511.05879, 2015.
[30] Gordo, A.; Almazán, J.; Revaud, J.; Larlus, D. Deep image retrieval: Learning global representations for image search. *arXiv preprint* arXiv: 1604.01325, 2016.
[31] Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *arXiv preprint* arXiv: 1506.01497, 2015.
[32] Forcén, J. I.; Pagola, M.; Barrenechea, E.; Bustince, H. Aggregation of deep features for image retrieval based on object detection. In: Pattern Recognition and Image Analysis. Lecture Notes in Computer Science, Vol. 11867. Morales, A.; Fierrez, J.; Sánchez, J. S.; Ribeiro, B. Eds. Springer Cham, 553–564, 2019.
[33] Babenko, A.; Lempitsky, V. Aggregating deep convolutional features for image retrieval. *arXiv preprint* arXiv: 1510.07493, 2015.
[34] Kalantidis, Y.; Mellina, C.; Osindero, S. Cross-dimensional weighting for aggregated deep convolutional features. *arXiv preprint* arXiv: 1512.04065, 2016.
[35] Jeong, D. J.; Choo, S.; Seo, W.; Cho, N. I. Regional deep feature aggregation for image retrieval. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1737–1741, 2017.
[36] Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing, 234–241, 2015.
[37] He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. *arXiv preprint* arXiv: 1603.05027, 2016.
[38] Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L.; et al. Swin transformer V2: Scaling up capacity and resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11999–12009, 2022.
[39] Tang, H.; Liu, D.; Shen, C. Data-efficient multi-scale fusion vision transformer. *Pattern Recognition* Vol. 161, Article No. 111305, 2025.
[40] Wang, J.; Yu, X.; Gao, Y. Feature fusion vision transformer for fine-grained visual categorization. *arXiv preprint* arXiv: 2107.02341, 2021.
[41] Huang, Y.; Chen, Y.; Wu, C.; Song, B.; Wang, H. Image super-resolution reconstruction network based on enhanced swin transformer via alternating aggregation of local-global features. *arXiv preprint* arXiv: 2401.00241, 2023.
[42] Liu, X.; Zhou, L.; Zhou, Z.; Chen, J.; He, Z. MambaVLT: Time-evolving multimodal state space model for vision-language tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8731–8741, 2025.
[43] Li, K.; Li, X.; Wang, Y.; He, Y.; Wang, Y.; Wang, L.; Qiao, Y. VideoMamba: State space model for efficient video understanding. *arXiv preprint* arXiv: 2403.06977, 2024.
[44] Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W. Y.; et al. Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 3992–4003, 2024.
[45] Liang, J.; Cao, J.; Fan, Y.; Zhang, K.; Ranjan, R.; Li, Y.; Timofte, R.; Van Gool, L. VRT: A video restoration transformer. *IEEE Transactions on Image Processing* Vol. 33, 2171–2182, 2024.
[46] Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Communications of the ACM* Vol. 60, No. 6, 84–90, 2017.
[47] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint* arXiv: 2010.11929, 2020.
[48] Chen, X.; Fan, H.; Girshick, R.; He, K. Improved baselines with momentum contrastive learning. *arXiv preprint* arXiv: 2003.04297, 2020.
[49] Caron, M.; Touvron, H.; Misra, I.; Jegou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 9630–9640, 2022.
[50] Yang, B.; Bender, G.; Le, Q. V.; Ngiam, J. CondConv: Conditionally parameterized convolutions for efficient inference. In: Proceedings of the Neural Information Processing Systems, 2019.
[51] Zhang, D.; Zhang, H.; Tang, J.; Wang, M.; Hua, X.; Sun, Q. Feature pyramid transformer. *arXiv preprint* arXiv: 2007.09451, 2020.
[52] Liu, D.; Cui, Y.; Yan, L.; Mousas, C.; Yang, B.; Chen, Y. DenserNet: Weakly supervised visual localization using multi-scale feature aggregation. *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 35, No. 7, 6101–6109, 2021.
[53] Chen, J.; Xu, F.; Zeng, T.; Li, X.; Chen, S.; Yu, J. MSFA: Multi-stage feature aggregation network for multi-label image recognition. *IET Image Processing* Vol. 18, No. 7, 1862–1877, 2024.
[54] Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7132–7141, 2018.
[55] Woo, S.; Park, J.; Lee, J. Y.; Kweon, I. S. CBAM: Convolutional block attention module. *arXiv preprint* arXiv: 1807.06521, 2018.
[56] You, H.; Lu, Y.; Tang, H. Improved feature extraction and similarity algorithm for video object detection. *Information* Vol. 14, No. 2, Article No. 115, 2023.
[57] Liu, J.; Zhang, W.; Tang, Y.; Tang, J.; Wu, G. Residual feature aggregation network for image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2356–2365, 2020.
[58] Yi, Z.; Tang, Q.; Azizi, S.; Jang, D.; Xu, Z. Contextual residual aggregation for ultra high-resolution image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7505–7514, 2020.
[59] Chen, Z.; Xu, Q.; Cong, R.; Huang, Q. Global context-aware progressive aggregation network for salient object detection. *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 34, No. 7, 10599–10606, 2020.
[60] Bashir, S. M. A.; Wang, Y. Small object detection in remote sensing images with residual feature aggregation-based super-resolution and object detector network, *Remote Sensing*, vol. 13,

no. 9, pp. 1854, 2021.
[61] Hu, H.; Bai, S.; Li, A.; Cui, J.; Wang, L. Dense relation distillation with context-aware aggregation for few-shot object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10180–10189, 2021.
[62] Wu, Y.; Jiang, J.; Huang, Z.; Tian, Y. FPANet: Feature pyramid aggregation network for real-time semantic segmentation. *Applied Intelligence* Vol. 52, No. 3, 3319–3336, 2022.
[63] Ren, S.; Zhou, D.; He, S.; Feng, J.; Wang, X. Shunted self-attention via multi-scale token aggregation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10843–10852, 2022.
[64] Chen, J.; Hu, H.; Wu, H.; Jiang, Y.; Wang, C. Learning the best pooling strategy for visual semantic embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 15784–15793, 2021.
[65] Hassner, T.; Masi, I.; Kim, J.; Choi, J.; Harel, S.; Natarajan, P.; Medioni, G. Pooling faces: Template based face recognition with pooled face images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 127–135, 2016.
[66] Rao, Y.; Lin, J.; Lu, J.; Zhou, J. Learning discriminative aggregation network for video-based face recognition. In: Proceedings of the IEEE International Conference on Computer Vision, 3801–3810, 2017.
[67] Liu, Y.; Yan, J.; Ouyang, W. Quality aware network for set to set recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4694–4703, 2017.
[68] Yang, J.; Ren, P.; Zhang, D.; Chen, D.; Wen, F.; Li, H.; Hua, G. Neural aggregation network for video face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5216–5225, 2017.
[69] Gong, S.; Shi, Y.; Kalka, N. D.; Jain, A. K. Video face recognition: Component-wise feature aggregation network (C-FAN). In: Proceedings of the International Conference on Biometrics, 2020.
[70] Goswami, G.; Vatsa, M.; Singh, R. Face verification via learned representation on feature-rich video frames. *IEEE Transactions on Information Forensics and Security* Vol. 12, No. 7, 1686–1698, 2017.
[71] Liu, Z.; Hu, H.; Bai, J.; Li, S.; Lian, S. Feature aggregation network for video face recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop, 990–998, 2019.
[72] Cevikalp, H.; Triggs, B. Face recognition based on image sets. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2567–2573, 2010.
[73] Li, H.; Hua, G.; Shen, X.; Lin, Z.; Brandt, J. Eigen-PEP for video face recognition. In: *Computer Vision--ACCV 2014. Lecture Notes in Computer Science, Vol. 9005*. Cremers, D.; Reid, I.; Saito, H.; Yang, M. -H. Eds Cham Springer, 17–33, 2015.
[74] Crosswhite, N.; Byrne, J.; Stauffer, C.; Parkhi, O.; Cao, Q.; Zisserman, A. Template adaptation for face verification and identification. *Image and Vision Computing* Vol. 79, 35–48, 2018.
[75] Deng, J.; Guo, J.; Zhang, D.; Deng, Y.; Lu, X.; Shi, S. Lightweight face recognition challenge. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop, 2638–2646, 2019.
[76] Sun, Y.; Wang, X.; Tang, X. Deep learning face representation from predicting 10, 000 classes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1891–1898, 2014.
[77] Parkhi, O. M.; Vedaldi, A.; Zisserman, A. Deep face recognition. *Deep Face Recognition* 2015.
[78] Rao, Y.; Lu, J.; Zhou, J. Attention-aware deep reinforcement learning for video face recognition. In: Proceedings of the IEEE International Conference on Computer Vision, 3951–3960, 2017.
[79] Peng, B.; Jin, X.; Wu, Y.; Li, D. Geometry guided feature aggregation in video face recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop, 2670–2677, 2019.
[80] Zhang, L.; Dong, Y.; Wu, Y. Multi-layer CNN features aggregation for real-time visual tracking. In: Proceedings of the 24th International Conference on Pattern Recognition, 2404–2409, 2018.
[81] Zhang, Z.; Lan, C.; Zeng, W.; Chen, Z. Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10404–10413, 2020.
[82] Wang, Y.; Zhang, P.; Gao, S.; Geng, X.; Lu, H.; Wang, D. Pyramid spatial-temporal aggregation for video-based person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 12006–12015, 2022.
[83] Li, W.; Tao, X.; Guo, T.; Qi, L.; Lu, J.; Jia, J. MuCAN: Multi-correspondence aggregation network for video super-resolution. *arXiv preprint* arXiv: 2007.11803, 2020.
[84] Sun, L.; Jia, K.; Yeung, D. Y.; Shi, B. E. Human action recognition using factorized spatio-temporal convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, 4597–4605, 2016.
[85] Zhu, X.; Wang, Y.; Dai, J.; Yuan, L.; Wei, Y. Flow-guided feature aggregation for video object detection. In: Proceedings of the IEEE International Conference on Computer Vision, 408–417, 2017.
[86] Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Li, F. F. Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1725–1732, 2014.
[87] Wang, H.; Kläser, A.; Schmid, C.; Liu, C. -L. Action recognition by dense trajectories. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, 3169–3176, 2011.
[88] Dosovitskiy, A.; Fischer, P.; Ilg, E.; Häusser, P.; Hazirbas, C.; Golkov, V.; van der Smagt, P.; Cremers, D.; Brox, T. FlowNet: Learning optical flow with convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, 2758–2766, 2016.
[89] Kang, K.; Ouyang, W.; Li, H.; Wang, X. Object detection from video tubelets with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 817–825, 2016.
[90] Han, W.; Khorrami, P.; Le Paine, T.; Ramachandran, P.; Babaeizadeh, M.; Shi, H.; Li, J.; Yan, S.; Huang, T. S. Seq-NMS for video object detection. *arXiv preprint* arXiv: 1602.08465, 2016.
[91] Brox, T.; Bruhn, A.; Papenberg, N.; Weickert, J. High accuracy optical flow estimation based on a theory for warping. In: *Computer Vision-ECCV 2004. Lecture Notes in Computer Science, Vol. 3024*. Pajdla, T.; Matas, J. Eds Chem Springer, 25–36, 2004.
[92] Gu, X.; Chang, H.; Ma, B.; Shan, S. Motion feature aggregation for video-based person re-identification. *IEEE Transactions on Image Processing* Vol. 31, 3908–3919, 2022.
[93] Shi, Y.; Dong, M.; Xu, C. Harnessing vision foundation models for high-performance, training-free open vocabulary segmentation. *arXiv preprint* arXiv: 2411.09219, 2024.
[94] Zhang, Y.; Li, X.; Liu, C.; Shuai, B.; Zhu, Y.; Brattoli, B.; Chen, H.; Marsic, I.; Tighe, J. VidTr: Video transformer without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 13557–13567, 2022.
[95] Liang, J.; Fan, Y.; Xiang, X.; Ranjan, R.; Ilg, E.; Green, S.; Cao, J.; Zhang, K.; Timofte, R.; Gool, L. V. Recurrent video restoration transformer with guided deformable attention. *arXiv preprint* arXiv: 2206.02146, 2022.
[96] Ying, X.; Wang, L.; Wang, Y.; Sheng, W.; An, W.; Guo, Y. Deformable 3D convolution for video super-resolution. *IEEE Signal Processing Letters* Vol. 27, 1500–1504, 2020.
[97] Li, Y.; Ji, B.; Shi, X.; Zhang, J.; Kang, B.; Wang, L. TEA: Temporal excitation and aggregation for action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 906-915, 2020. [LinkOut]

[98] He, F.; Gao, N.; Li, Q.; Du, S.; Zhao, X.; Huang, K. Temporal context enhanced feature aggregation for video object detection. *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 34, No. 7, 10941–10948, 2020.

[99] Fu, Y.; Yang, L.; Liu, D.; Huang, T. S.; Shi, H. CompFeat: Comprehensive feature aggregation for video instance segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 35, No. 2, 1361–1369, 2021.

[100] Chen, Y.; Cao, Y.; Hu, H.; Wang, L. Memory enhanced global-local aggregation for video object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 10334–10343, 2020.

[101] Jadon, S.; Jasim, M. Unsupervised video summarization framework using keyframe extraction and video skimming. In: Proceedings of the IEEE 5th International Conference on Computing Communication and Automation, 140–145, 2020.

[102] Cui, Y. DFA: Dynamic feature aggregation for efficient video object detection. *arXiv preprint* arXiv: 2210.00588, 2022.

[103] Shorten, C.; Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *Journal of Big Data* Vol. 6, No. 1, Article No. 60, 2019.

[104] Garcia-Garcia, A.; Orts-Escolano, S.; Oprea, S.; Villena-Martinez, V.; Martinez-Gonzalez, P.; Garcia-Rodriguez, J. A survey on deep learning techniques for image and video semantic segmentation. *Applied Soft Computing* Vol. 70, 41–65, 2018.

[105] Pan, S. J.; Yang, Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* Vol. 22, No. 10, 1345–1359, 2010.

[106] Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random erasing data augmentation. *arXiv preprint* arXiv: 1708.04896, 2017.

[107] Koziarski, M.; Cyganek, B. Image recognition with deep neural networks in presence of noise–Dealing with and taking advantage of distortions. *Integrated Computer-Aided Engineering* Vol. 24, No. 4, 337–349, 2017.

[108] Wen, S.; Dong, M.; Yang, Y.; Zhou, P.; Huang, T.; Chen, Y. End-to-end detection-segmentation system for face labeling. *IEEE Transactions on Emerging Topics in Computational Intelligence* Vol. 5, No. 3, 457–467, 2021.

[109] Khan, K.; Attique, M.; Syed, I.; Gul, A. Automatic gender classification through face segmentation. *Symmetry* Vol. 11, No. 6, Article No. 770, 2019.

[110] Khan, K.; Attique, M.; Syed, I.; Sarwar, G.; Irfan, M. A.; Khan, R. U. A unified framework for head pose, age and gender classification through end-to-end face segmentation. *Entropy* Vol. 21, No. 7, Article No. 647, 2019.

[111] Benini, S.; Khan, K.; Leonardi, R.; Mauro, M.; Migliorati, P. Face analysis through semantic face segmentation. *Signal Processing: Image Communication* Vol. 74, 21–31, 2019.

[112] Zhang, Q. S.; Zhu, S. C. Visual interpretability for deep learning: A survey. *Frontiers of Information Technology & Electronic Engineering* Vol. 19, No. 1, 27–39, 2018.

[113] Chakraborty, S.; Tomsett, R.; Raghavendra, R.; Harborne, D.; Alzantot, M.; Cerutti, F.; Srivastava, M.; Preece, A.; Julier, S.; Rao, R. M.; et al. Interpretability of deep learning models: A survey of results. In: Proceedings of the IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation, 2018.

[114] Olah, C.; Satyanarayan, A.; Johnson, I.; Carter, S.; Schubert, L.; Ye, K.; Mordvintsev, A. The building blocks of interpretability. *Distill* Vol. 3, No. 3, Article No. e10, 2018.

[115] Olah, C.; Mordvintsev, A.; Schubert, L. Feature visualization. *Distill* Vol. 2, No. 11, Article No. e7, 2017.