OmniFD: A Unified Model for Versatile Face Forgery Detection

Haotian Liu, Haoyu Chen, Chenhui Pan, You Hu, Guoying Zhao, and Xiaobai Li

***Abstract*—Face forgery detection encompasses multiple critical tasks, including identifying forged images and videos and localizing manipulated regions and temporal segments. Current approaches typically employ task-specific models with independent architectures, leading to computational redundancy and ignoring potential correlations across related tasks. We introduce OmniFD, a unified framework that jointly addresses four core face forgery detection tasks within a single model, i.e., image and video classification, spatial localization, and temporal localization. Our architecture consists of three principal components: (1) a shared Swin Transformer encoder that extracts unified 4D spatiotemporal representations from both images and video inputs, (2) a cross-task interaction module with learnable queries that dynamically captures inter-task dependencies through attention-based reasoning, and (3) lightweight decoding heads that transform refined representations into corresponding predictions for all FFD tasks. Extensive experiments demonstrate OmniFD's advantage over task-specific models. Its unified design leverages multi-task learning to capture generalized representations across tasks, especially enabling fine-grained knowledge transfer that facilitates other tasks. For example, video classification accuracy improves by 4.63% when image data are incorporated. Furthermore, by unifying images, videos and the four tasks within one framework, OmniFD achieves superior performance across diverse benchmarks with high efficiency and scalability, e.g., reducing 63% model parameters and 50% training time. It establishes a practical and generalizable solution for comprehensive face forgery detection in real-world applications. The source code is made available at https://github.com/haotianll/OmniFD.**

***Index Terms*—Face Forgery Detection, Deepfake, Unified Model, Multi-Task Learning**

I. Introduction

Face forgery detection (FFD) [1], [2] identifies whether a face is authentic or fabricated. With the rapid advancements in face synthesis and manipulation technologies [1], [3], including AI-generated content (AIGC) [4], [5], FFD has become essential for safeguarding against misinformation, identity fraud, and other malicious applications.

FFD comprises several core tasks, including *Image Forgery Classification* [1], [6], [7], *Video Forgery Classification* [8]–[11], *Spatial Forgery Localization* [12]–[14], and *Temporal Forgery Localization* [15]–[18]. Each task focuses on a distinct aspect of forgery analysis. ***Image Forgery Classification*** [19]–[33] determines whether a given face image is real or fake, which is the most fundamental FFD task. ***Video Forgery Classification*** [34]–[43] processes facial video clips to assess their authenticity, which captures temporal dependencies in video sequences to mitigate the limitations of image-based FFD analysis. Instead of providing a simple binary classification, ***Spatial Forgery Localization*** [13], [15], [44]–[48] identifies manipulated pixels in an image, which aims to enhance practical application and improve interpretability,

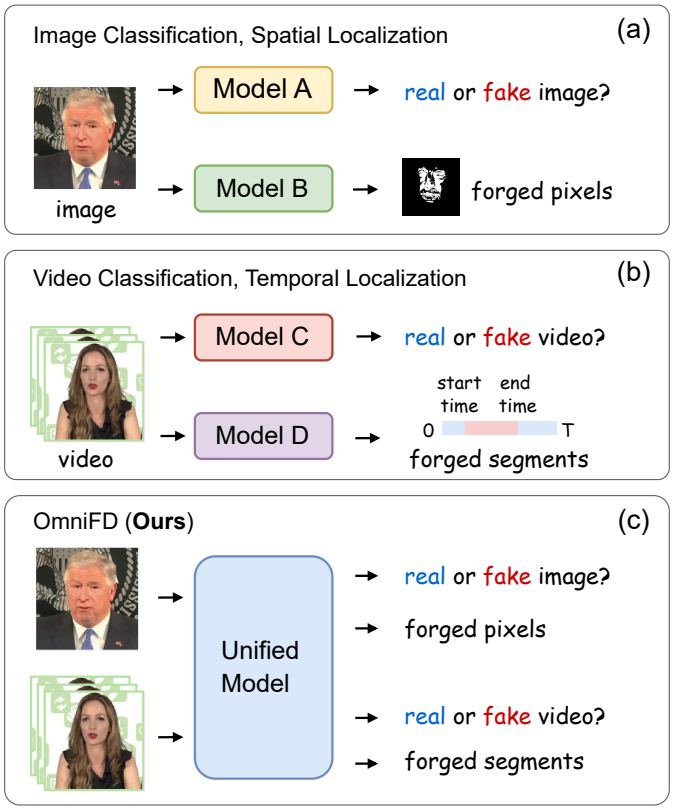
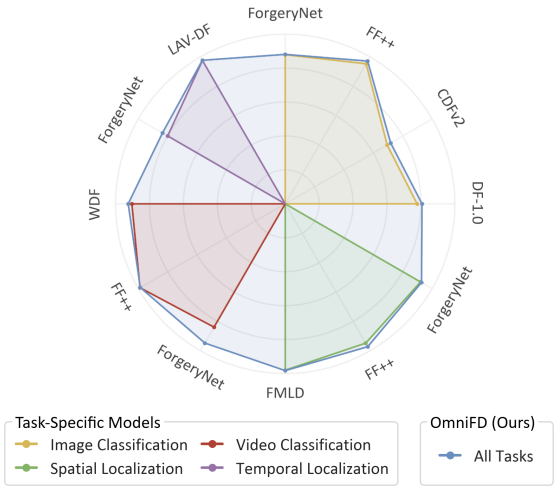


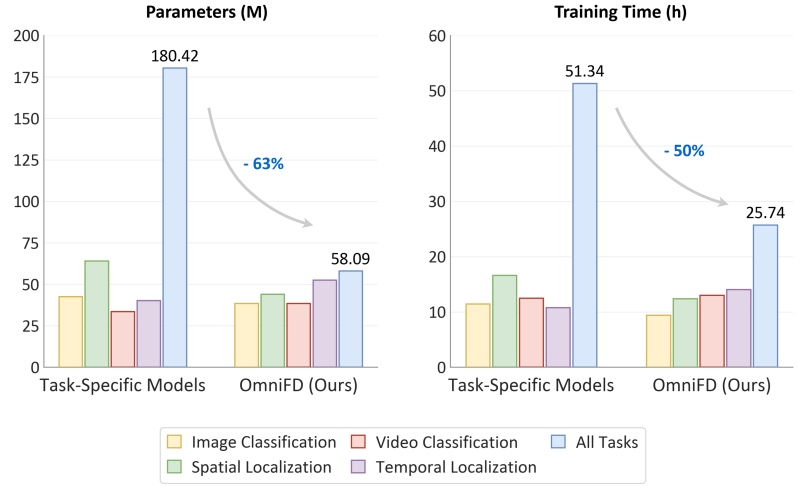
Fig. 1. Illustration of face forgery detection solutions: (a) and (b) task-specific models for the four tasks, i.e., image classification, spatial localization, video classification and temporal localization; (c) we propose a unified model addressing all the four tasks. Our method achieves both superior performance and computational efficiency compared to existing approaches.

assisting human users in understanding and verifying forgeries. In addition, ***Temporal Forgery Localization*** [49]–[54] focuses on detecting manipulated segments within a video clip, where only certain frames are altered. This task addresses the challenge of advanced forgeries in long video clips. Meanwhile, various FFD benchmarks [1], [3], [8], [9], [13], [15], [16], [55]–[58] have been established for these different tasks to evaluate the performance of FFD methods. Despite the swift progress, most existing methods are tailored to one specific task. We summarize the limitations of existing studies as follows:

1) *Redundant and inflexible model designs.* Existing frameworks treat FFD tasks independently, training task-specific models with distinct architectures for each task [1], [8], [58], as described in Figure 1. This fragmented approach leads to redundancy in both model parameters and training costs, especially when multiple tasks are involved. Moreover,



(a) Performance analysis



(b) Efficiency analysis

Fig. 2. Advantages of the proposed unified model against task-specific models. (a) **Performance analysis.** The evaluation covers representative benchmark datasets for each task, including [1], [8], [9], [13], [15], [55], [56], and the latest state-of-the-art methods [30], [33], [41], [42], [44], [48], [54] are selected for comparison. The proposed OmniFD achieves superior performance within a single unified framework on all benchmarks of the four tasks. Detailed results are presented in Tables I, IV-VI. (b) **Efficiency analysis.** Here we compare representative task-specific models with OmniFD in terms of model parameters and training time for handling a single FFD task or all tasks jointly. OmniFD greatly reduces parameter redundancy and computational overhead compared with task-specific models. More detailed results are provided in Table III.

modules specialized for one task are difficult to adapt to others, which limits overall system efficiency.

2) *Potential correlations and benefits among different tasks are overlooked.* For instance, detecting a forged frame in a video can aid in verifying the authenticity of the entire video clip. Although some methods [8], [29], [59] migrate this issue by using additional annotations from related tasks for joint training, they prioritize one primary task and treat others as auxiliary objectives. As a result, the potential of cross-task correlations remains underexploited.

3) *Deployment challenges and poor scalability.* Using multiple specialized models with different architectures [8] for FFD tasks lead to inefficiency in deployment, particularly in resource-constrained environments. This limits their applicability in real-world scenarios that require comprehensive FFD analysis, such as the concurrent detection of forgeries and localization of manipulated regions or segments.

A unified model [60]–[66] that incorporates multiple modalities and related tasks has attracted attention in visual understanding and natural language processing fields and demonstrated its advantages. The above-mentioned challenges in FFD could be addressed by developing a unified framework to process different tasks. Fundamentally, all four FFD tasks (i.e., image classification, video classification, spatial localization, and temporal localization) aim to identify forgery artifacts, differing only in modalities and annotation granularity. Existing unified models [65], [66] mainly concentrate on tasks within a single modality, or are designed as classification frameworks [60], [61], [64] for multiple modalities in natural scenarios, making them unsuitable for unifying FFD tasks across images and videos. There are preliminary studies [8], [29], [59] on unified FFD frameworks, but they address no more than two tasks and show limited generalization capability.

To address these limitations, we propose OmniFD, a unified framework for face forgery detection. OmniFD is grounded in the principle of multi-task learning [67], which encourages shared representation learning and jointly optimization of multiple objectives. It aims to provide a unified and generalizable solution that enhances both performance and computational efficiency, as illustrated in Figure 2. The contributions of this paper include:

- To our knowledge, it is the first attempt to address all four FFD tasks within a unified solution, eliminating the need for task-specific model designs and improving computational efficiency.
- A unified encoder is designed to unify representation learning across images and videos by modeling spatial and temporal forgeries with shared parameters.
- The framework leverages multi-task learning and cross-task interaction module to facilitate knowledge transfer among related tasks, allowing supervision from a task to reinforce another, such as between image and video classification.
- Extensive experiments on multiple FFD datasets demonstrate that OmniFD achieves state-of-the-art performance with strong generalization across tasks. Its efficiency and scalability make it well-suited for real-world deployment.

II. Related Work

A. Face Forgery Detection

Face forgery detection [1], [2] aims to identify manipulated or synthesized faces in images and videos, which is essential for mitigating misinformation and identity fraud risks. Early studies [1], [19] focused on image classification, using convolutional neural networks (CNNs) [6], [7], [68] to extract features from static images for binary classification.

Some approaches leveraged local textures [20], [21], frequency domain analysis [22]–[24], [69], and inconsistency information [25]–[28] to improve accuracy. With the rise of deep generative models and face manipulation techniques, recent works [30]–[33] emphasize improving generalization against unknown forgery attack types. These image-based methods were also used to process video inputs by extracting frames from video clips and applying image classification models, while overlooking temporal dependencies present in videos. In contrast, other approaches [34]–[43] process entire video clips directly for video classification. Early methods employed recurrent neural networks (RNNs) to identify inconsistencies such as eye-blinking [34], lip motions [35] and biological signals [36]. Subsequent methods [37], [39]–[41], [43] employed 3D CNNs and vision transformers to capture enhanced temporal dependencies for video forgery detection.

While image and video classification methods have shown promising performance and generalization capabilities on forgery benchmarks, they often lack interpretability for human users and fail to meet real-world application needs. To address this, researchers have explored fine-grained forgery detection tasks in spatial and temporal dimensions. Specifically, spatial localization [1], [12] predicts manipulated pixels in an image, while temporal localization [8], [15] detects forged segments in a video clip. These tasks require detailed annotations and improved representation, and specialized benchmarks [13], [15], [16] and methods [13]–[15], [17], [44]–[54] are proposed. As some datasets [1], [8] contain annotations of multiple forgery detection tasks, methods [8], [29], [59] integrate annotations of different tasks, e.g., image classification and spatial localization [29], for joint training. However, they prioritize one task and use others as auxiliary supervision, which limits their generalization to other tasks.

To summarize, current forgery detection methods are usually designed and trained separately for each specific task, limiting cross-task or cross-modal model reuse. While all tasks focus on detecting forgery-related cues at different levels, their interrelations remain underexplored.

B. Unified Models

Unified models [60]–[66] have gained significant attention recently, standing out for their versatility and generalization capability. Unified vision models focus on processing different input modalities or vision tasks using a single model. Existing studies can be categorized into unified architectures and unified learning frameworks.

Unified architectures employ a single backbone to extract generic representations from various modalities, such as images and videos. As a video clip can be viewed as a sequence of image frames, a few methods [61], [70], [71] proposed to adapt 2D CNN or Transformer backbones from image tasks to process video data. OMNIVORE [60] proposed a unified 4D representation for images, videos, and single-view 3D data, and utilized a Transformer model with shared parameters to process all modalities simultaneously. Besides, recent methods [64], [65] also proposed to align various data modalities, such as image, video, text, and audio, into a unified feature space via share Transformers. The unified architectures demonstrate promising performance and generalization capabilities, which also pave the way to process facial images and videos together in face forgery detection.

Unified learning frameworks utilize a single model or formulation to tackle a suite of related vision tasks. Most existing work focuses on tasks within the same modality. For example, object detection and segmentation methods [72], [73] followed a multi-task learning paradigm [67] and jointly trained a shared backbone on a few prediction tasks. Later studies [74]–[76] first unified the annotation format of various tasks and then designed a shared decoder for making predictions. With the rapid progress in large language models (LLMs) [77], [78] and vision-language pretraining [79]–[81], recent methods [62], [63], [66] transform task predictions into text sequence predictions, thereby leveraging LLMs to achieve unified decoding across diverse visual tasks. Nevertheless, the exploration of unified learning frameworks for cross-modal tasks, particularly involving both images and videos, remains limited. While some methods [60], [61], [82] jointly process images and videos, they solely focus on classification. Consequently, unified models for face forgery detection across different tasks and modalities remain unexplored.

III. Method

The proposed OmniFD framework provides a unified FFD solution across image and video modalities. It jointly addresses four FFD tasks within a single architecture. This formulation enables unified representation learning and promotes knowledge transfer among related tasks. The framework consists of three components: 1) a unified encoder that extracts generalized spatiotemporal features from images and videos, 2) a cross-task interaction module that models task dependencies through transformer-based reasoning, and 3) task-oriented decoding heads that produce predictions for four FFD tasks. The model is optimized end-to-end under a multi-task learning paradigm. The remainder of this section is structured as follows: the unified representation across modalities is described in Section III-A, the cross-task interaction module in Section III-B, the task-oriented decoding in Section III-C, and the end-to-end optimization strategy in Section III-D.

A. Unified Representation Across Modalities

Our goal is to utilize a shared backbone for feature extraction across four FFD tasks through multi-task learning. Since all four tasks aim to identify forgery-related artifacts, using a shared backbone promotes unified representation learning and improves parameter efficiency. The main challenge arises from the heterogeneous input modalities, i.e., images and videos, and the different information priorities among the four tasks. Image-related tasks emphasize spatial cues, particularly spatial localization that relies on multi-scale features, whereas video-related tasks must also capture temporal information.

To address this challenge, we introduce a unified encoder that processes both image and video data to extract unified representations. Inspired by [82], the input data is converted into a 4D tensor $\mathbf{X} \in \mathbb{R}^{T \times H \times W \times 3}$ (with $T = 1$ for images and $T > 1$

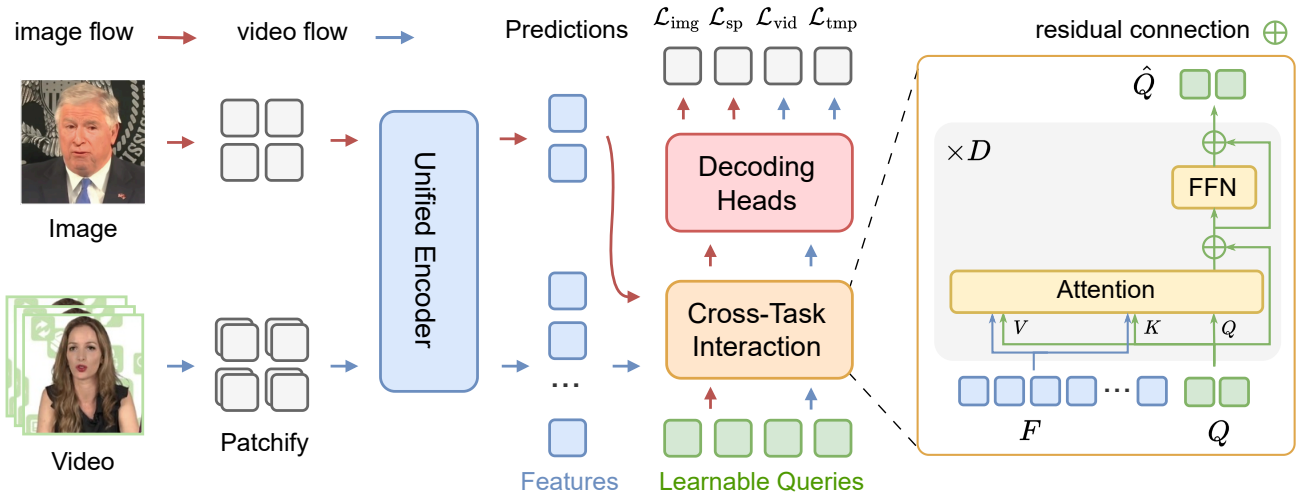


Fig. 3. Overview of the proposed OmniFD framework. (1) The unified encoder processes both image and video inputs to extract shared spatiotemporal features. (2) The cross-task interaction module employs learnable queries that interact with the features to model inter-task dependencies. (3) The decoding heads transform the refined representations into predictions for four FFD tasks. The framework is optimized end-to-end under multi-task learning paradigm.

for videos), divided into smaller patches and then fed into a Swin Transformer [83]. The unified encoder leverages local self-attention together with spatial and temporal positional encodings to produce feature maps that capture fine-grained spatial information and temporal dynamics. Consequently, it generates spatiotemporal features $\{\mathbf{F}^l\}_{l=1}^{L}$ from L backbone stages. Since images and videos share the same parameters during feature extraction, the unified encoder learns generalized representations across both modalities.

Typical 2D CNN backbones can also be used to process image and video inputs with shared parameters, as described in [59], [70], but they are less effective than the 4D tensor-based approach. We will make comparisons and discuss in the experiment sections.

B. Cross-Task Interaction Module

Multi-task learning [67] improves generalization by leveraging inductive bias that favors hypotheses useful across related tasks. The inductive bias is manifested through auxiliary training signals, which guide the model to learn representations that capture relationships shared among tasks instead of optimizing each task independently. This inductive bias is crucial in multi-task FFD, where different tasks rely on common manipulation artifacts. The shared cues promote consistent representations across tasks, allowing knowledge from related tasks to reinforce the target task. For instance, detecting a forged frame in a video can aid in verifying the authenticity of the entire video clip. The previous step employs a unified encoder to learn generalized representations, and the current step focuses on explicitly capturing interactions among FFD tasks.

Inspired by [80], [84], we propose a transformer-based cross-task interaction module that employs learnable queries to model relationships among tasks, as illustrated in Figure 3. Specifically, we introduce a set of learnable queries $\mathbf{Q} \in \mathbb{R}^{N\times C}$ that act as shared latent variables capturing dependencies among tasks, where N and C denote the number and dimension of queries, respectively. The queries interact with the feature maps $\{\mathbf{F}^l\}_{l=1}^{L}$ through attention mechanisms and are refined into $\hat{\mathbf{Q}}$ over D transformer layers. The refined queries are optimized to predict each FFD task, which drives them to encode forgery artifacts shared among FFD tasks.

The interaction among FFD tasks is achieved within the cross-task interaction module. The queries and feature maps are concatenated to form keys and values, allowing each query to aggregate information from others and enhance its own representation through the attention mechanism, as illustrated in Figure 3. The process is formulated as:

$$\mathbf{K} = \mathbf{V} = \text{Concat}\big(\mathbf{Q}, \{\mathbf{F}^l\}_{l=1}^{L}\big), \tag{1}$$

$$\hat{\mathbf{Q}} = \text{LN}\big(\mathbf{Q} + \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V})\big), \tag{2}$$

$$\hat{\mathbf{Q}} = \text{LN}\big(\hat{\mathbf{Q}} + \text{FFN}(\hat{\mathbf{Q}})\big), \tag{3}$$

where $\mathbf{K}$, $\mathbf{V}$ are the key and value tensors, LN denotes LayerNorm and FFN is Feedforward Network. Through this process, the refined queries $\hat{\mathbf{Q}} \in \mathbb{R}^{N\times C}$ encode shared forgery patterns across FFD tasks.

Our cross-task interaction module employs a distinct mechanism compared to prior approaches [80], [84], [85]. Existing methods either perform dimension reduction of input sequences [80] or assign a single query to handle one specific prediction task [84], [85], our cross-task interaction process allows learnable queries to adaptively model inter-task representations. These queries jointly address multiple FFD objectives through shared feature reasoning, overcoming the limitations of task-specific parameterization.

C. Task-Oriented Decoding Heads

The refined queries obtained from the cross-task interaction module provide more compact representations compared with the original features. To make the representations discriminative, we utilize the refined queries to predict all four FFD

tasks, allowing supervision signals from multiple objectives to refine their representations. As the training progresses, the refined queries are optimized into embeddings that capture forgery-related artifacts shared across FFD tasks.

Based on this design, we employ four lightweight decoding heads that transform the refined queries into task-specific predictions. In practice, separate decoding heads are used for each task to avoid unnecessary computational overhead when predictions for only one or several tasks are required. As shown in Figure 4, we adopts three decoding processes for the four FFD tasks.

1) Image and Video Classification: For image and video classification, average pooling AvgPool$(\cdot)$ is applied to the refined queries $\hat{\mathbf{Q}} \in \mathbb{R}^{N\times C}$ to obtain a single feature vector, which is then passed through a linear layer Linear$(\cdot)$:

$$p = \text{Linear}(\text{AvgPool}(\hat{\mathbf{Q}})), \tag{4}$$

where p represents the predicted logit for binary classification.

2) Spatial Localization: The spatial localization process first interpolates feature maps $\{\mathbf{F}^l\}_{l=1}^{L}$ to the same spatial resolution, concatenates them, and applies a 1×1 convolution layer to obtain the fused feature $\mathbf{F}_{sp} \in \mathbb{R}^{HW\times C}$. The pixel-wise localization mask is then predicted as:

$$\mathbf{M}_{sp} = \text{Proj}(\mathbf{F}_{sp}\hat{\mathbf{Q}}^{\top}), \tag{5}$$

where Proj$(\cdot)$ denotes a linear projection layer with weights $\mathbf{W}_{sp} \in \mathbb{R}^{N\times1}$, and $\mathbf{M}_{sp} \in \mathbb{R}^{H\times W}$ represents the pixel-wise mask of forged region.

3) Temporal Localization: Similar to typical temporal localization methods [18], [50], the temporal feature $\mathbf{F}_{tmp} \in \mathbb{R}^{T\times C}$ is averaged from the final feature map $\mathbf{F}^L$. It predict forged segments with two outputs: classification scores $\mathbf{S}_{cls} \in \mathbb{R}^{T\times1}$ and boundary regression offsets $\mathbf{S}_{reg} \in \mathbb{R}^{T\times2}$ (start and end time) at each temporal position $t \in \{1, ..., T\}$.

To incorporate the semantic information from the refined queries $\hat{\mathbf{Q}}$, transformer layer is applied to enhance the temporal feature $\mathbf{F}_{tmp}$ as:

$$\mathbf{K} = \mathbf{V} = \text{Concat}\big(\mathbf{F}_{tmp}, \hat{\mathbf{Q}}\big), \tag{6}$$

$$\hat{\mathbf{F}}_{tmp} = \text{LN}\big(\mathbf{F}_{tmp} + \text{Attention}(\mathbf{F}_{tmp}, \mathbf{K}, \mathbf{V})\big), \tag{7}$$

$$\hat{\mathbf{F}}_{tmp} = \text{LN}\big(\hat{\mathbf{F}}_{tmp} + \text{FFN}(\hat{\mathbf{F}}_{tmp})\big). \tag{8}$$

This process resembles Eq. 2, where $\mathbf{F}_{tmp}$ serves as the query. It preserves the temporal dimension necessary for segment prediction at each temporal position. The prediction of forged segments is defined as:

$$\mathbf{S}_{cls} = \text{Conv}_{\text{cls}}(\hat{\mathbf{F}}_{tmp}), \tag{9}$$

$$\mathbf{S}_{reg} = \text{Conv}_{\text{reg}}(\hat{\mathbf{F}}_{tmp}), \tag{10}$$

where Conv_{cls} and Conv_{reg} are 1D convolutional layers.

D. End-to-End Optimization

The OmniFD framework randomly samples annotated facial images and videos to form mini-batches and performs end-to-end optimization via multi-task learning. The overall loss $\mathcal{L}$ is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{img}} + \mathcal{L}_{\text{sp}} + \mathcal{L}_{\text{vid}} + \mathcal{L}_{\text{tmp}}, \tag{11}$$

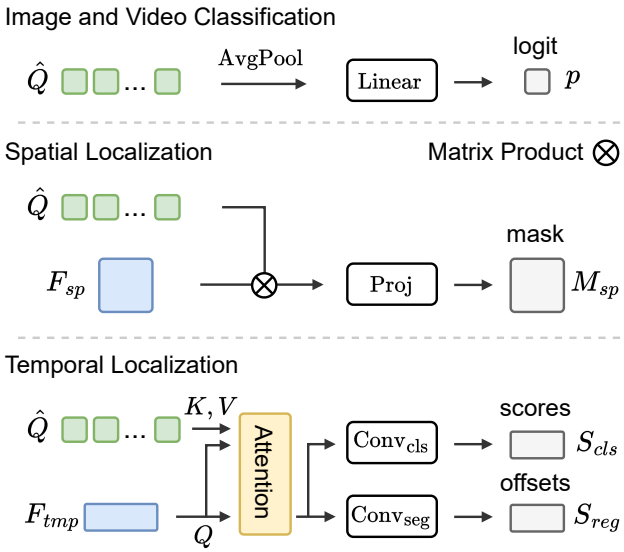


Fig. 4. Illustration of decoding heads for different FFD tasks in OmniFD.

where $\mathcal{L}_{\text{img}}$, $\mathcal{L}_{\text{sp}}$, $\mathcal{L}_{\text{vid}}$, and $\mathcal{L}_{\text{tmp}}$ represent losses for each task, i.e., image classification, spatial localization, video classification, and temporal localization, respectively.

Specifically, $\mathcal{L}_{\text{img}}, \mathcal{L}_{\text{vid}}$ employs standard cross entropy loss for binary classification, $\mathcal{L}_{\text{sp}}$ computes the pixel-wise cross entropy loss, while $\mathcal{L}_{\text{tmp}}$ combines focal loss for classification and DIoU loss for regression, as defined in [18]. Following common practice, $\mathcal{L}_{\text{sp}}$ and $\mathcal{L}_{\text{tmp}}$ are computed only from forged samples, while real samples are excluded from loss calculation.

Note that the unified design of OmniFD enables training with annotations from either a single task or any combination of tasks, without requiring both image and video data or full annotations for all four tasks. The experimental results on multiple FFD datasets validate this flexibility.

IV. Experiments

A. Experimental Setup

1) Datasets: We evaluate our approach on five publicly available face forgery detection datasets.

ForgeryNet [8] is a large-scale and comprehensive dataset for face forgery detection, comprising over 2.9 million static images and 220k video clips. It includes 15 distinct image manipulation techniques and 8 distinct video manipulation techniques. The image and video subsets contain different subjects, forming two separate datasets. ForgeryNet provides detailed annotations for all four FFD tasks: the image subset supports image classification and spatial localization, while the video subset supports video classification and temporal localization. Following [8], 2.3 million images serve as the training set and 150k as the validation set, while 140k videos are used for training and 18k for validation.

FaceForensics++ (FF++) [1] is a representative dataset for video classification, containing 1,000 real videos, each altered by four forgery methods. It is divided into training, validation, and test sets with 720, 140, and 140 videos, respectively.

Since it provides annotations of manipulated regions and every frame in the fake videos is modified, FF++ is also adopted for image classification and spatial localization by extracting frames from the videos, as described in [1], [13].

WildDeepfake (WDF) [9] contains 7,314 deepfake video clips collected from the internet, covering diverse scenes, identities, and activities for video forgery classification. The dataset is split into 6,508 training and 806 testing samples [9].

FMLD [13] is a spatial forgery localization dataset comprising 120k synthetic and attribute-manipulated facial images generated using four GAN-based methods. The training, validation, and test sets follow an 8:1:1 split defined in [13].

LAV-DF [15] is an audio-visual deepfake dataset containing manipulated segments in video or audio modalities. It includes 78k training videos and 26k testing videos. We follow [15] and use the subset that excludes audio-only manipulation videos for evaluating temporal localization.

Seven additional datasets are used as test sets to evaluate the cross-domain generalization capability of image classification models, including *CelebDF-v1 (CDFv1)* [55], *CelebDF-v2 (CDFv2)* [55], *DeeperForensics-1.0 (DF-1.0)* [56], *DeepFakeDetection (DFD)* [1], *DFDC* [57], *DFDCP* [57], and *FaceShifter (Fsh)* [3]. Following the protocol introduced in DeepfakeBench [58], we construct image classification datasets based on them by extracting facial frames from videos when necessary.

Experiments are conducted on these datasets. ForgeryNet dataset serves as the unified training source for all four FFD tasks. Its annotations across image and video modalities are used to evaluate OmniFD as a unified multi-task learning framework (Section IV-B). FF++, WDF, LAV-DF, and FMLD datasets are employed to assess OmniFD's generalization performance on individual tasks, whereas seven additional datasets (CDFv1, CDFv2, DF-1.0, DFD, DFDC, DFDCP, and Fsh) are used to evaluate its cross-domain generalization in image classification (Section IV-E).

2) Evaluation Metrics: We adopt commonly used evaluation metrics for each FFD task following established protocols. Specifically, image classification is evaluated using Accuracy (Acc) and Area Under the ROC Curve (AUC) metrics, as in [1], [8]. For video classification, we report video-level Accuracy (Acc) and AUC following [8]. For spatial localization, two levels of Intersection over Union are used for ForgeryNet: IoU (threshold 0.1) and IoU_{diff} (threshold 0.01) [8], while Pixel-wise Binary Classification Accuracy (PBCA) and Inverse Intersection Non-Containment (IINC) are employed for other datasets [1], [13], consistent with [12], [13]. For temporal localization, Average Precision (AP), mean Average Precision (mAP) and Average Recall (AR) metrics are utilized, following the evaluation protocols in [8], [15].

3) Implementation Details: For data processing, we follow [1], [29], [59] and utilize RetinaFace [86] to detect and extract facial regions from images and videos in ForgeryNet dataset. The cropped faces are then resized to a resolution of 224×224. For videos of ForgeryNet, we randomly sample 32 frames per video with a temporal stride of 4 during training. Following [1], [8], [15], we apply standard data augmentation techniques during training, including geometric transformations (random resizing, cropping, and horizontal flipping) and random perturbations (image compression and color distortion).

The proposed OmniFD is implemented with PyTorch and MMEngine [87]. We adopt the Swin Transformer [83] as the unified encoder architecture, initialized with pretrained weights from ImageNet-1k [88] and Kinetics-400 [89], following [60]. The cross-task interaction module uses 64 learnable queries ($N = 64$) and 3 transformer layers ($D = 3$). During training, OmniFD jointly processes image and video data and constructs each mini-batch with 128 static images and 8 video clips from ForgeryNet [8]. Training is conducted on four NVIDIA Tesla V100 GPUs using the AdamW optimizer. The weight decay is set to 0.0001. The initial learning rate is set to 0.0001, followed by a cosine decay scheduler. We train OmniFD for 50k iterations to ensure convergence. The same training settings are used for all comparison methods unless specified otherwise.

B. Comparison with State-of-the-Art

We compare the proposed OmniFD with state-of-the-art FFD methods on the ForgeryNet [8] dataset, as shown in Table I and Figure 2(a). The comparison includes task-specific methods and multi-task learning methods across four FFD tasks, i.e., video classification, temporal localization, image classification, and spatial localization. OmniFD utilizes the same parameters for all tasks, except for the decoding heads. Experimental results show that OmniFD achieves state-of-the-art performance across all four tasks. Overall, OmniFD with Swin-S achieves the best performance on all four tasks (with the second-best mAP on temporal localization and accuracy on image classification), outperforming all single-task and dual-task baseline methods.

We further compare OmniFD with the conventional multi-task learning approach [67] and the unified OMNIVORE model [66]. The multi-task learning approach uses annotations from the same modality (e.g., image classification and spatial localization), whereas OMNIVORE processes images and videos together for classification. Specifically, the multi-task learning approach is implemented with a Swin-T backbone integrated with task-specific heads: linear layers for image and video classification, ActionFormer [18] for temporal localization, and FCN [90] for spatial localization. Compared with these methods, OmniFD achieves superior performance and supports multiple FFD tasks across different modalities in a single model. These results indicate its effectiveness as a unified framework for face forgery detection.

C. Advantage of Multi-Task Integration

OmniFD is flexible and versatile, which can be used for all four tasks, fewer tasks, or only one task to suit different application scenarios. To explicitly demonstrate the advantages of fusing multiple FFD tasks, we test OmniFD in different training settings and compare the performance, i.e., OmniFD is trained on one task, a pair of tasks, and all four tasks. The experiment is conducted on the ForgeryNet dataset, and the results are shown in Table II. The four FFD tasks, i.e., video classification, temporal localization, image classification, and

TABLE I
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON FORGERYNET [8] VIDEO AND IMAGE VALIDATION SETS. RESULTS MARKED WITH † ARE CITED FROM [8], [41], [59]. BOLD AND UNDERLINED INDICATE THE BEST AND SECOND-BEST RESULTS.

Method	Backbone	Video Classification		Temporal Localization		Image Classification		Spatial Localization	
		Acc	AUC	AP@0.5	mAP	Acc	AUC	IoU	IoU_{diff}
SlowFast† [10]	SlowFast-50	88.78	93.88	-	-	-	-	-	-
VideoSwin [11]	Swin-T	84.22	91.73	-	-	-	-	-	-
UniFormer [38]	UniFormer-S	82.59	89.16	-	-	-	-	-	-
TALL [40]	Swin-B	82.38	87.73	-	-	-	-	-	-
UVIF† [59]	ResNet-101	86.57	94.42	-	-	-	-	-	-
MINTIME-XC† [41]	Xception	87.64	94.25	-	-	-	-	-	-
DFD-FCG [43]	ViT-L	84.13	91.73	-	-	-	-	-	-
Xception† [7]	Xception	-	-	68.29	62.83	-	-	-	-
SlowFast+BSN† [17]	3D ResNet-50	-	-	82.25	**73.42**	-	-	-	-
ActionFormer [18]	SlowFast-101	-	-	74.73	55.81	-	-	-	-
TriDet [50]	Swin-T	-	-	74.08	59.69	-	-	-	-
Re2TAL [51]	SlowFast-101	-	-	74.69	56.03	-	-	-	-
UMMAFormer [54]	ResNet-50	-	-	84.37	66.89	-	-	-	-
AdaTAD [52]	VideoMAE-B	-	-	78.06	60.96	-	-	-	-
EfficientNet† [6]	EfficientNet-B0	-	-	-	-	79.86	89.31	-	-
Xception† [7]	Xception	-	-	-	-	80.78	90.12	-	-
ResNeSt† [68]	ResNeSt-101	-	-	-	-	**82.06**	<u>91.02</u>	-	-
GramNet† [20]	ResNet-18	-	-	-	-	80.89	90.20	-	-
F3-Net† [22]	Xception	-	-	-	-	80.86	90.15	-	-
M2TR [29]	EfficientNet-B4	-	-	-	-	81.41	90.67	-	-
LAA-Net [33]	EfficientNet-B4	-	-	-	-	81.74	90.92	-	-
Xception+Reg† [8]	Xception	-	-	-	-	-	-	89.55	67.57
Xception+Unet† [8]	Xception	-	-	-	-	-	-	95.99	79.71
HRNet† [14]	HRNet-w48	-	-	-	-	-	-	96.27	88.73
UperNet [45]	HRNet-w48	-	-	-	-	-	-	<u>98.70</u>	91.99
IML-ViT [46]	ViT-B	-	-	-	-	-	-	98.64	94.32
Mesorch [47]	MiT-B3	-	-	-	-	-	-	98.49	93.96
SparseViT [48]	SparseViT	-	-	-	-	-	-	98.62	94.40
multi-task (video) [67]	Swin-T	84.64	91.63	77.68	60.29	-	-	-	-
multi-task (image) [67]	Swin-T	-	-	-	-	79.31	89.28	98.69	<u>94.80</u>
OMNIVORE-T [60]	Swin-T	87.43	94.98	-	-	78.45	88.60	-	-
OMNIVORE-S [60]	Swin-S	<u>89.66</u>	<u>96.31</u>	-	-	80.83	90.33	-	-
OmniFD **(Ours)**	Swin-T	88.75	95.44	<u>84.71</u>	65.66	80.08	89.52	98.69	94.76
OmniFD **(Ours)**	Swin-S	**90.10**	**96.57**	**87.32**	<u>68.24</u>	<u>81.89</u>	**91.10**	**98.72**	**94.93**

spatial localization, are denoted as *Video*, *Temporal*, *Image*, and *Spatial*, respectively.

From Table II, three main findings can be concluded:

1) Cross-modal learning provides significant performance gains. Joint training of image and video classification (*Image+Video*) leads to a clear improvement in video classification, achieving a +4.63% gain in accuracy and +3.77% in AUC over the video classification baseline (row 1).

2) Fine-grained temporal and spatial cues enhance higher-level forgery recognition. Specifically, joint training of video classification and temporal localization (*Video+Temporal*) yields +0.95% improvement in video classification accuracy, while the *Image+Spatial* setting increases image classification accuracy by +0.40%.

3) Full-task integration delivers the best overall improvement. OmniFD trained jointly on all four tasks (row 8) achieves the best overall performance, reaching 88.75% accuracy in video classification and competitive results across other tasks.

The results demonstrate beneficial interactions among related and cross-modal tasks. Tasks with fine-grained annotations, such as temporal or spatial localization, provide detailed supervision that facilitates knowledge transfer and improves higher-level classification performance. In cross-modal learning, large-scale image data offer diverse visual priors that strengthen representation learning in the video domain, leading to more generalizable and transferable features. These benefits are further amplified when all tasks are trained jointly, allowing OmniFD to learn generalized representations across tasks and modalities.

The spatial localization task shows limited improvement in the multi-task setting, primarily because other tasks lack supervision at a finer annotation level. Overall, OmniFD's unified multi-task framework effectively exploits complementary information from FFD tasks, achieving robust and generalizable forgery detection across diverse scenarios.

D. Efficiency Analysis

The efficiency of OmniFD stems from its unified architecture, which allows a single model to handle multiple FFD tasks during both training and inference. In contrast, task-specific approaches use separate models for each task, leading

TABLE II
THE EFFECT OF DIFFERENT TASK COMBINATIONS IN OMNIFD ON FORGERYNET [8] VALIDATION SET.

Setting	Video Classification		Temporal Localization		Image Classification		Spatial Localization	
	Acc	AUC	AP@0.5	mAP	Acc	AUC	IoU	IoU_{diff}
Video	84.07	91.71	-	-	-	-	-	-
Temporal	-	-	77.91	60.93	-	-	-	-
Image	-	-	-	-	79.58	89.42	-	-
Spatial	-	-	-	-	-	-	**98.71**	**95.06**
Video + Temporal	85.02	91.63	79.06	61.10	-	-	-	-
Image + Spatial	-	-	-	-	79.98	**89.94**	**98.71**	94.85
Video + Image	88.70	**95.48**	-	-	79.91	89.87	-	-
All Tasks	**88.75**	95.44	**84.71**	**65.66**	**80.08**	89.52	98.69	94.76

TABLE III
EFFICIENCY COMPARISON OF OMNIFD AND TASK-SPECIFIC MODELS WHEN PERFORMING ONE TO FOUR FFD TASKS. RESULTS ARE BASED ON A 224×224 IMAGE OR 32-FRAME VIDEO ON A V100 GPU.

	Task	Param (M)	Flops (G)	Latency (ms)	Training Time (h)
Task-Specific	*Image* (*F3-Net*)	42.52	10.95	9.51	11.45
	Spatial (*UperNet*)	64.04	45.40	11.89	16.63
	Video (*SlowFast*)	33.65	50.57	40.63	12.47
	Temporal (*UMMAFormer*)	40.21	138.95	86.41	10.79
	Image + Spatial	106.56	56.35	21.40	28.08
	Video + Temporal	73.86	189.53	127.04	23.26
	All Tasks	180.42	245.87	148.44	51.34
OmniFD (**Ours**)	*Image*	38.42	7.21	11.86	9.39
	Spatial	44.01	11.51	13.47	12.39
	Video	38.42	121.84	52.11	13.02
	Temporal	52.50	122.71	62.43	14.06
	Image + Spatial	44.01	11.51	13.47	12.80
	Video + Temporal	52.50	122.71	62.42	14.09
	All Tasks	58.09	134.21	75.89	25.74

to redundant parameters and extra computation when multiple tasks are involved.

Table III presents the quantitative results of efficiency analysis. We compare OmniFD (Swin-T) with four representative task-specific models, i.e., F3-Net [22], UperNet [45], SlowFast [10], and UMMAFormer [54] (from Table I). We report the number of parameters, FLOPs, inference latency, and training time for processing one to four FFD tasks. All metrics are measured using a 224×224 image or a 32-frame video input on an NVIDIA V100 GPU. The results indicate that OmniFD achieves comparable efficiency to task-specific models (rows 1–4) when trained on a single task, while its advantage becomes more pronounced as more tasks are involved. Since task-specific models have distinct architectures, multiple forwardings are required to make predictions for different FFD tasks. For instance, when processing all four tasks simultaneously (row 7), OmniFD reduces up to 63.40% parameters, 45.41% FLOPs, 48.87% latency, and 49.86% training time compared to task-specific models. Figure 2(b) further illustrates the efficiency improvements. These results demonstrate OmniFD's superiority in computational efficiency and its practicality for real-world deployment.

E. Generalization of the OmniFD Architecture

The OmniFD is designed as a general-purpose architecture capable of generalizing across diverse FFD tasks and datasets. Beyond leveraging multi-task data for joint training, OmniFD can also address each task independently using the same architecture. To evaluate this generalization capability of OmniFD, we compare OmniFD with state-of-the-art methods across four FFD tasks on representative datasets beyond ForgeryNet, as summarized in Tables IV–VII and Figure 2(a). To ensure fair comparisons, we adopt identical evaluation protocols and use the same training data as other methods when training OmniFD on each task or dataset.

Table IV presents the comparison results of image classification. We follow the same data pre-processing and cross-dataset evaluation protocol introduced in DeepfakeBench [58], train OmniFD (Swin-S) on the FF++ (c23) [1] dataset and evaluate the AUC performance across FF++ and other seven datasets: CDFv1 [55], CDFv2 [55], DF-1.0 [56], DFD [1], DFDC [57], DFDCP [57], and Fsh [3]. The results show that OmniFD achieves the best performance in both intra-dataset (98.52% AUC) and cross-dataset (78.14% average AUC) evaluation, demonstrating strong generalization capability in image forgery classification. Similarly, Tables V, VI, and VII summarize the results on spatial localization, temporal localization, and video classification, respectively. Following the protocols in [15], [42], [44], we evaluate spatial localization on FF++ and FMLD [13] datasets, video classification on FF++ and WDF [9] datasets, and temporal localization on LAV-DF [15] dataset. OmniFD consistently achieves state-of-the-art performance across all FFD tasks over existing methods.

These results show that when OmniFD is trained on a single dataset for an individual task, it achieves state-of-the-art performance across benchmarks, demonstrating strong generalization capability and the simplicity of its unified architecture.

F. Ablation Studies

1) Effectiveness of the Core Components: We analyze the effectiveness of OmniFD's core components, i.e., unified encoder and cross-task interaction module, and the ablation results with Swin-T are presented in Table VIII.

The results show that both the unified encoder and cross-task interaction module are important for OmniFD's unified modeling. Compared to removing the unified encoder and training each task separately (row 1), using a unified encoder with shared weights for four tasks (row 3) brings significant performance gains, e.g., +4.72% video classification accuracy

TABLE IV
COMPARISON RESULTS AND CROSS-DATASET GENERALIZATION ON REPRESENTATIVE IMAGE CLASSIFICATION BENCHMARKS.

Method	**Intra**	**Cross Dataset**							
	FF++	CDFv1	CDFv2	DF-1.0	DFD	DFDC	DFDCP	Fsh	Avg.
Xception [7]	96.37	77.94	73.65	83.41	81.63	70.77	73.74	62.49	74.80
EfficientNet [6]	95.67	79.09	74.87	83.30	81.48	69.55	72.83	61.62	74.68
F3-Net [22]	95.92	70.93	67.86	55.31	76.55	63.26	69.42	65.53	66.98
RECCE [28]	96.21	76.77	73.19	79.85	81.19	71.33	74.19	60.95	73.92
UCF [30]	97.05	77.93	75.27	82.41	80.74	71.91	75.94	64.62	75.55
OmniFD (**Ours**)	**98.52**	**79.45**	**77.32**	**84.87**	**81.97**	**76.49**	**76.32**	**70.56**	**78.14**

TABLE V
COMPARISON RESULTS OF SPATIAL LOCALIZATION ON FF++ [1] AND FMLD [13] TESTING SETS.

Method	FF++		FMLD	
	PBCA ↑	IINC ↓	PBCA ↑	IINC ↓
DFFD [12]	94.85	4.57	98.72	3.31
Locate [13]	95.77	3.62	99.06	2.53
SAM [76]	92.97	4.25	97.29	3.40
SAM+LoRA [91]	93.12	4.78	98.06	3.51
DADF [44]	96.64	3.21	99.26	2.64
OmniFD (**Ours**)	**98.60**	**1.91**	**99.46**	**1.55**

TABLE VI
COMPARISON RESULTS OF TEMPORAL LOCALIZATION ON LAV-DF [15] TESTING SUBSET.

Method	AP@0.5	AP@0.75	AP@0.95	AR@10	AR@20
BA-TFD [53]	85.20	47.06	0.29	59.32	61.19
TadTR [49]	83.48	63.57	5.44	71.38	72.42
TriDet [50]	80.71	60.93	2.91	67.63	67.64
BA-TFD+ [15]	96.82	86.47	3.90	79.15	79.60
UMMAFormer [54]	98.83	95.95	30.11	92.32	92.65
OmniFD (**Ours**)	**98.96**	**96.81**	**59.69**	**97.08**	**97.47**

TABLE VII
COMPARISON RESULTS OF VIDEO CLASSIFICATION ON FF++ [1] AND WDF [9] TESTING SETS.

Method	FF++		WDF	
	Acc	AUC	Acc	AUC
MADD [21]	97.60	99.29	82.62	90.71
RECCE [28]	97.06	99.32	83.25	92.02
UAL [31]	98.13	99.56	84.16	92.56
SFDG [32]	98.19	99.53	84.41	92.57
DFGaze [42]	**98.80**	**99.91**	86.10	93.00
OmniFD (**Ours**)	98.57	99.76	**87.10**	**94.76**

TABLE VIII
ABLATION STUDY OF THE CORE COMPONENTS IN THE OMNIFD FRAMEWORK.

Unified Encoder	Cross-Task Interaction	**Video**	**Temporal**	**Image**	**Spatial**
		Acc	mAP	Acc	IoU_{diff}
-	✓	84.03	60.93	79.58	**95.06**
✓	-	88.41	64.22	79.50	94.77
✓	✓	**88.75**	**65.66**	**80.08**	94.76

and +4.73% temporal localization mAP, demonstrating the benefit of unified representation learning with one shared encoder. Meanwhile, row 2 removes the cross-task interaction module and replaces it with four vanilla task-specific decoders. Specifically, we use average pooling with linear layers for video and image classification, convolutional layers for spatial localization, and remove the transformer layer in temporal localization, making predictions of each task directly from the features without queries of the cross-task interaction module. Compared to this setting, OmniFD with cross-task interaction module shows clear gains, e.g., +1.44% mAP in temporal localization and +0.58% accuracy in image classification, indicating that it improves knowledge transfer across tasks and promotes a unified decoding process.

2) Performance of Different Backbones: We perform ablation experiments to evaluate OmniFD with different backbone architectures, as shown in Table IX. The backbones include representative CNN models, ResNet [92] and ConvNeXt [93], as well as the 4D tensor-based Swin Transformer [83]. We train four separate OmniFD models on each FFD task as baselines and compare them with the jointly trained model.

The main finding is that the jointly trained OmniFD consistently improves overall performance over the baselines across all backbones, significantly improving video classification accuracy by at least 4%. The backbones exhibit performance variations when trained separately for each task, mainly due to differences in architectural design. Notably, the Swin Transformer, with its 4D tensor modeling and window-based attention, proves more effective for unified modeling in OmniFD than CNN-based backbones. Among backbones of the same architecture (rows 5–8), the Swin-S variant achieves higher overall performance than Swin-T, indicating that larger-capacity models provide stronger generalization and enhance OmniFD's representation learning across FFD tasks. In summary, these results demonstrate the effectiveness of unified 4D tensor modeling and the scalability of OmniFD across backbone architectures.

G. Visualization

1) Attention Map: We visualize the attention maps between the learnable queries $\mathbf{Q}$ and input features $\mathbf{F}^L$ to analyze the mechanism of learnable queries in the cross-task interaction module, as shown in Figure 5. The comparison includes four types of forged faces and two real faces from ForgeryNet [8]. The visualization results show that different queries attend to distinct facial regions, from global facial contours to local attributes such as the eyes and mouth. For the same query, forged and real samples often exhibit different activation

TABLE IX
ABLATION OF DIFFERENT BACKBONES FOR OMNIFD ON THE FORGERYNET [8] VALIDATION SET. THE SUBSCRIPT DENOTES IMPROVEMENT OVER THE CORRESPONDING BASELINE.

Method	**Video**	**Temporal**	**Image**	**Spatial**
	Acc	mAP	Acc	IoU_{diff}
ResNet-50 [92]	80.00	56.57	79.04	94.75
OmniFD (**Ours**)	$85.51_{\uparrow 5.5}$	$57.57_{\uparrow 1.0}$	$78.63_{\downarrow 0.4}$	$94.39_{\downarrow 0.4}$
ConvNeXt-T [93]	82.02	61.25	80.01	95.17
OmniFD (**Ours**)	$86.44_{\uparrow 4.4}$	$64.37_{\uparrow 3.1}$	$80.10_{\uparrow 0.1}$	$94.87_{\downarrow 0.3}$
Swin-T [83]	84.03	60.93	79.58	95.06
OmniFD (**Ours**)	$88.75_{\uparrow 4.7}$	$65.66_{\uparrow 4.7}$	$80.08_{\uparrow 0.5}$	$94.76_{\downarrow 0.3}$
Swin-S [83]	84.89	63.32	81.42	95.13
OmniFD (**Ours**)	$90.10_{\uparrow 5.2}$	$68.24_{\uparrow 4.9}$	$81.89_{\uparrow 0.5}$	$94.93_{\downarrow 0.2}$

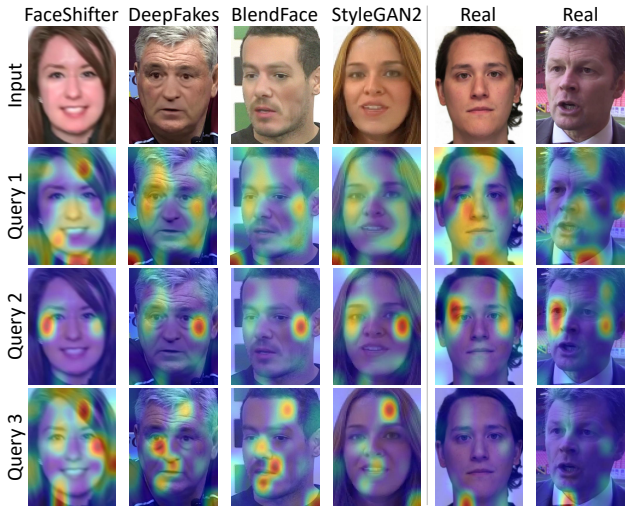


Fig. 5. Visualization of attention maps between learnable queries $\hat{\mathbf{Q}}$ and features on forged and real faces. Samples of four types of forged faces and two real faces from ForgeryNet [8] are included for comparison. The three bottom rows (Query 1, 2, and 3) show that different queries learn to focus different facial attributes. Comparing the four left columns with the right two columns, it shows that the activation patterns of real faces differ from those of the forged faces.

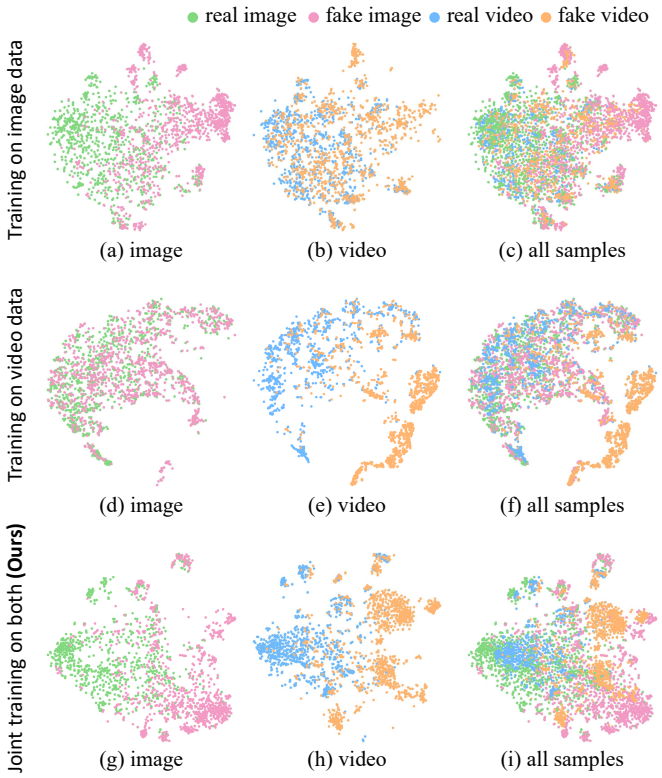


Fig. 6. T-SNE visualization of features in OmniFD under three settings: training on image data (a)-(c), training on video data (d)-(f), and joint training on both (g)-(i). Each row corresponds to a different training setting, and the three columns show t-SNE embeddings for image, video, and all samples from the same model. Single-modality training fails to achieve cross-modal generalization, while joint training yields compact clusters with consistent real–fake boundaries across image and video modalities.

patterns, where certain queries (e.g., Query 3) provide strong cues for assessing facial authenticity. These results indicate that the learnable queries capture forgery-related artifacts and encode them into generalized representations useful for face forgery detection.

2) T-SNE Visualization: We further analyze the feature distribution and cross-modal alignment in OmniFD via t-SNE visualization, as shown in Figure 6. We consider three training settings: training on image data, training on video data, and joint training on both. We extract the t-SNE embeddings of the last-stage features $\mathbf{F}^L$ using the same model on both image and video samples from ForgeryNet [8].

As shown in Figure 6(a)–(c), when trained only on image data, the model separates real and fake images, but video samples remain entangled. A similar pattern appears in Figure 6(d)–(f), where real and fake videos are separated while image samples are intermixed. In contrast, jointly training OmniFD on image and video data, i.e., Figure 6(g)–(i), produces compact and well-aligned clusters for both modalities. Figure 6(i) shows clearer boundaries between real and fake samples across modalities, whereas those in Figure 6(c) and (f) are vague. These results demonstrate that single-modality training fails to learn representations that generalize across modalities, making joint supervision essential for learning modality-invariant representations and improving cross-modal alignment.

V. CONCLUSION

We present OmniFD, the first unified model that simultaneously tackles four core face forgery detection tasks: image classification, video classification, spatial localization, and temporal localization. OmniFD eliminates the need for task-specific models by employing a unified encoder to extract generalized representations across different modalities and tasks. With the proposed cross-task interaction module and multi-task optimization, OmniFD effectively exploits inter-task correlations, achieving state-of-the-art performance on multiple benchmarks while demonstrating its efficiency and scalability. This provides a unified framework for practical and generalizable face forgery detection. For future work, we will develop foundational models for face forgery detection from a data scaling perspective. Additionally, we will explore how

to integrate multimodal data such as audio and text for more comprehensive face forgery analysis.

REFERENCES

[1] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 1–11.

[2] F. Juefei-Xu, R. Wang, Y. Huang, Q. Guo, L. Ma, and Y. Liu, "Countering malicious deepfakes: Survey, battleground, and horizon," *Int. J. Comput. Vis. (IJCV)*, vol. 130, no. 7, pp. 1678–1734, 2022.

[3] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Advancing high fidelity identity swapping for forgery detection," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2020, pp. 5074–5083.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, June 2022, pp. 10 684–10 695.

[6] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *Int. Conf. Learn. Represent. (ICLR)*. PMLR, 2019, pp. 6105–6114.

[7] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2017, pp. 1251–1258.

[8] Y. He, B. Gan, S. Chen, Y. Zhou, G. Yin, L. Song, L. Sheng, J. Shao, and Z. Liu, "Forgerynet: A versatile benchmark for comprehensive forgery analysis," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2021, pp. 4360–4369.

[9] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, "Wilddeepfake: A challenging real-world dataset for deepfake detection," in *ACM Int. Conf. Multimedia (ACMMM)*, 2020, pp. 2382–2390.

[10] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 6202–6211.

[11] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022, pp. 3202–3211.

[12] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, "On the detection of digital face manipulation," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2020, pp. 5781–5790.

[13] C. Kong, B. Chen, H. Li, S. Wang, A. Rocha, and S. Kwong, "Detect and locate: Exposing face manipulation by semantic-and noise-level telltales," *IEEE Trans. Inf. Forensics Secur. (TIFS)*, vol. 17, pp. 1741–1756, 2022.

[14] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, vol. 43, no. 10, pp. 3349–3364, 2020.

[15] Z. Cai, S. Ghosh, A. Dhall, T. Gedeon, K. Stefanov, and M. Hayat, "Glitch in the matrix: A large scale benchmark for content driven audio-visual forgery detection and localization," *Comput. Vis. Image Underst. (CVIU)*, vol. 236, p. 103818, 2023.

[16] Z. Cai, S. Ghosh, A. P. Adatia, M. Hayat, A. Dhall, T. Gedeon, and K. Stefanov, "Av-deepfake1m: A large-scale llm-driven audio-visual deepfake dataset," in *ACM Int. Conf. Multimedia (ACMMM)*, 2024, pp. 7414–7423.

[17] T. Lin, X. Zhao, H. Su, C. Wang, and M. Yang, "Bsn: Boundary sensitive network for temporal action proposal generation," in *Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[18] C.-L. Zhang, J. Wu, and Y. Li, "Actionformer: Localizing moments of actions with transformers," in *Eur. Conf. Comput. Vis. (ECCV)*. Springer, 2022, pp. 492–510.

[19] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," in *IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*. IEEE, 2018, pp. 1–7.

[20] Z. Liu, X. Qi, and P. H. Torr, "Global texture enhancement for fake face detection in the wild," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2020, pp. 8060–8069.

[21] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional deepfake detection," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2021, pp. 2185–2194.

[22] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," in *Eur. Conf. Comput. Vis. (ECCV)*. Springer, 2020, pp. 86–103.

[23] J. Li, H. Xie, J. Li, Z. Wang, and Y. Zhang, "Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2021, pp. 6458–6467.

[24] C. Miao, Z. Tan, Q. Chu, H. Liu, H. Hu, and N. Yu, "F2trans: High-frequency fine-grained transformer for face forgery detection," *IEEE Trans. Inf. Forensics Secur. (TIFS)*, vol. 18, pp. 1039–1051, 2023.

[25] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face x-ray for more general face forgery detection," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2020, pp. 5001–5010.

[26] X. Dong, J. Bao, D. Chen, T. Zhang, W. Zhang, N. Yu, D. Chen, F. Wen, and B. Guo, "Protecting celebrities from deepfake with identity consistency transformer," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022, pp. 9468–9478.

[27] K. Shiohara and T. Yamasaki, "Detecting deepfakes with self-blended images," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022, pp. 18 720–18 729.

[28] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang, "End-to-end reconstruction-classification learning for face forgery detection," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022, pp. 4113–4122.

[29] J. Wang, Z. Wu, W. Ouyang, X. Han, J. Chen, Y.-G. Jiang, and S.-N. Li, "M2tr: Multi-modal multi-scale transformers for deepfake detection," in *Proc. Int. Conf. Multimed. Retr. (ICMR)*, 2022, pp. 615–623.

[30] Z. Yan, Y. Zhang, Y. Fan, and B. Wu, "Ucf: Uncovering common features for generalizable deepfake detection," in *Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 22 412–22 423.

[31] Y. Wu, X. Song, J. Chen, and Y.-G. Jiang, "Generalizing face forgery detection via uncertainty learning," in *ACM Int. Conf. Multimedia (ACMMM)*, 2023, pp. 1759–1767.

[32] Y. Wang, K. Yu, C. Chen, X. Hu, and S. Peng, "Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023, pp. 7278–7287.

[33] D. Nguyen, N. Mejri, I. P. Singh, P. Kuleshova, M. Astrid, A. Kacem, E. Ghorbel, and D. Aouada, "Laa-net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2024, pp. 17 395–17 405.

[34] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing ai created fake videos by detecting eye blinking," in *IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*. IEEE, 2018, pp. 1–7.

[35] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2021, pp. 5039–5049.

[36] U. A. Ciftci, I. Demir, and L. Yin, "Fakecatcher: Detection of synthetic portrait videos using biological signals," *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 2020.

[37] Z. Gu, Y. Chen, T. Yao, S. Ding, J. Li, F. Huang, and L. Ma, "Spatiotemporal inconsistency learning for deepfake video detection," in *ACM Int. Conf. Multimedia (ACMMM)*, 2021, pp. 3473–3481.

[38] K. Li, Y. Wang, G. Peng, G. Song, Y. Liu, H. Li, and Y. Qiao, "Uniformer: Unified transformer for efficient spatial-temporal representation learning," in *Int. Conf. Learn. Represent. (ICLR)*, 2022.

[39] H. Wang, Z. Liu, and S. Wang, "Exploiting complementary dynamic incoherence for deepfake video detection," *IEEE Trans. Circuits Syst. Video Technol. (TCSVT)*, 2023.

[40] Y. Xu, J. Liang, G. Jia, Z. Yang, Y. Zhang, and R. He, "Tall: Thumbnail layout for deepfake video detection," in *Int. Conf. Comput. Vis. (ICCV)*, October 2023, pp. 22 658–22 668.

[41] D. A. Coccomini, G. K. Zilos, G. Amato, R. Caldelli, F. Falchi, S. Papadopoulos, and C. Gennaro, "Mintime: multi-identity size-invariant video deepfake detection," *IEEE Trans. Inf. Forensics Secur. (TIFS)*, 2024.

[42] C. Peng, Z. Miao, D. Liu, N. Wang, R. Hu, and X. Gao, "Where deepfakes gaze at? spatial–temporal gaze inconsistency analysis for video face forgery detection," *IEEE Trans. Inf. Forensics Secur. (TIFS)*, vol. 19, pp. 4507–4517, 2024.

[43] Y.-H. Han, T.-M. Huang, K.-L. Hua, and J.-C. Chen, "Towards more general video-based deepfake detection through facial component guided adaptation for foundation model," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2025, pp. 22 995–23 005.

[44] Y. Lai, Z. Luo, and Z. Yu, "Detect any deepfakes: Segment anything meets face forgery detection and localization," in *Chin. Conf. Biom. Recognit.* Springer, 2023, pp. 180–190.

[45] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 418–434.

[46] X. Ma, B. Du, Z. Jiang, A. Y. A. Hammadi, and J. Zhou, "Iml-vit: Benchmarking image manipulation localization by vision transformer," *arXiv preprint arXiv:2307.14863*, 2023.

[47] X. Zhu, X. Ma, L. Su, Z. Jiang, B. Du, X. Wang, Z. Lei, W. Feng, C.-M. Pun, and J.-Z. Zhou, "Mesoscopic insights: orchestrating multi-scale & hybrid architecture for image manipulation localization," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 39, no. 10, 2025, pp. 11 022–11 030.

[48] L. Su, X. Ma, X. Zhu, C. Niu, Z. Lei, and J.-Z. Zhou, "Can we get rid of handcrafted feature extractors? sparsevit: Nonsemantics-centered, parameter-efficient image manipulation localization through spare-coding transformer," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 39, no. 7, 2025, pp. 7024–7032.

[49] X. Liu, Q. Wang, Y. Hu, X. Tang, S. Zhang, S. Bai, and X. Bai, "End-to-end temporal action detection with transformer," *IEEE Trans. Image Process. (TIP)*, vol. 31, pp. 5427–5441, 2022.

[50] D. Shi, Y. Zhong, Q. Cao, L. Ma, J. Li, and D. Tao, "Tridet: Temporal action detection with relative boundary modeling," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023, pp. 18 857–18 866.

[51] C. Zhao, S. Liu, K. Mangalam, and B. Ghanem, "Re2tal: Rewiring pretrained video backbones for reversible temporal action localization," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023, pp. 10 637–10 647.

[52] S. Liu, C.-L. Zhang, C. Zhao, and B. Ghanem, "End-to-end temporal action detection with 1b parameters across 1000 frames," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2024, pp. 18 591–18 601.

[53] Z. Cai, K. Stefanov, A. Dhall, and M. Hayat, "Do you really mean that? content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization," in *Int. Conf. Digit. Image Comput.: Tech. Appl. (DICTA)*. IEEE, 2022, pp. 1–10.

[54] R. Zhang, H. Wang, M. Du, H. Liu, Y. Zhou, and Q. Zeng, "Ummaformer: A universal multimodal-adaptive transformer framework for temporal forgery localization," in *ACM Int. Conf. Multimedia (ACMMM)*, 2023, pp. 8749–8759.

[55] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2020, pp. 3207–3216.

[56] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2020, pp. 2889–2898.

[57] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The deepfake detection challenge (dfdc) dataset," *arXiv preprint arXiv:2006.07397*, 2020.

[58] Z. Yan, Y. Zhang, X. Yuan, S. Lyu, and B. Wu, "Deepfakebench: a comprehensive benchmark of deepfake detection," in *Adv. Neural Inform. Process. Syst. (NeurIPS)*, 2023, pp. 4534–4565.

[59] H. Liu, C. Pan, Y. Liu, G. Zhao, and X. Li, "Unified video and image representation for boosted video face forgery detection," in *Eur. Conf. Artif. Intell. (ECAI)*, 2024, pp. 673–680.

[60] R. Girdhar, M. Singh, N. Ravi, L. Van Der Maaten, A. Joulin, and I. Misra, "Omnivore: A single model for many visual modalities," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022, pp. 16 102–16 112.

[61] A. Piergiovanni, W. Kuo, and A. Angelova, "Rethinking video vits: Sparse video tubes for joint image and video learning," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023, pp. 2214–2224.

[62] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang, "Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," in *Int. Conf. Mach. Learn. (ICML)*. PMLR, 2022, pp. 23 318–23 340.

[63] T. Chen, S. Saxena, L. Li, T.-Y. Lin, D. J. Fleet, and G. E. Hinton, "A unified sequence interface for vision tasks," *Adv. Neural Inform. Process. Syst. (NeurIPS)*, vol. 35, pp. 31 333–31 346, 2022.

[64] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Imagebind: One embedding space to bind them all," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023, pp. 15 180–15 190.

[65] J. Wang, Y. Ge, R. Yan, Y. Ge, K. Q. Lin, S. Tsutsui, X. Lin, G. Cai, J. Wu, Y. Shan *et al.*, "All in one: Exploring unified video-language pre-training," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023, pp. 6598–6608.

[66] J. Wang, D. Chen, C. Luo, B. He, L. Yuan, Z. Wu, and Y.-G. Jiang, "Omnivid: A generative framework for universal video understanding," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2024, pp. 18 209–18 220.

[67] R. Caruana, "Multitask learning," *Mach. learn.*, vol. 28, pp. 41–75, 1997.

[68] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha *et al.*, "Resnest: Split-attention networks," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2022, pp. 2736–2746.

[69] Z. Sun, N. Ruan, and J. Li, "Ddl: Effective and comprehensible interpretation framework for diverse deepfake detectors," *IEEE Trans. Inf. Forensics Secur. (TIFS)*, 2025.

[70] J. Lin, C. Gan, and S. Han, "Tsm: Temporal shift module for efficient video understanding," in *Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 7083–7093.

[71] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *Int. Conf. Mach. Learn. (ICML)*, vol. 2, no. 3, 2021, p. 4.

[72] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2961–2969.

[73] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2018, pp. 7482–7491.

[74] R. Hu and A. Singh, "Unit: Multimodal multitask learning with a unified transformer," in *Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 1439–1449.

[75] B. Yan, Y. Jiang, J. Wu, D. Wang, P. Luo, Z. Yuan, and H. Lu, "Universal instance perception as object discovery and retrieval," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023, pp. 15 325–15 336.

[76] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023, pp. 4015–4026.

[77] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[78] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[79] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Int. Conf. Mach. Learn. (ICML)*. PMLR, 2021, pp. 8748–8763.

[80] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Adv. Neural Inform. Process. Syst. (NeurIPS)*, vol. 35, pp. 23 716–23 736, 2022.

[81] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Int. Conf. Mach. Learn. (ICML)*. PMLR, 2023, pp. 19 730–19 742.

[82] H. Duan, Y. Zhao, Y. Xiong, W. Liu, and D. Lin, "Omni-sourced webly-supervised learning for video recognition," in *Eur. Conf. Comput. Vis. (ECCV)*. Springer, 2020, pp. 670–688.

[83] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 10 012–10 022.

[84] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Eur. Conf. Comput. Vis. (ECCV)*. Springer, 2020, pp. 213–229.

[85] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira, "Perceiver: General perception with iterative attention," in *Int. Conf. Mach. Learn. (ICML)*. PMLR, 2021, pp. 4651–4664.

[86] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2020, pp. 5203–5212.

[87] M. Contributors, "MMEngine: Openmmlab foundational library for training deep learning models," 2022.

[88] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis. (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[89] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[90] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2015, pp. 3431–3440.

[91] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.

[92] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2016, pp. 770–778.

[93] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "Convnext v2: Co-designing and scaling convnets with masked autoencoders," in *IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2023, pp. 16 133–16 142.