Seeing Through Fog: Towards Fog-Invariant Action Recognition

Enqi Liu[1,2], Liyuan Pan[1,3]*, Zhi Gao[1,2]*, Lingzhi Li[1], Qing Li[2]
[1] Beijing Institute of Technology, Beijing, China
[2] Beijing Institute for General Artificial Intelligence, Beijing, China
[3] Yangtze Delta Region Academy of Beijing Institute of Technology, Jiaxing, China
{enqi.liu, liyuan.pan, zhi.gao, lingzhi.li}@bit.edu, dylan.liqing@gmail.com

Abstract

Foggy conditions are commonly encountered in real-world applications; however, existing action recognition approaches typically assume favorable weather and high-quality video inputs. On foggy days, unpredictable visibility degradation and reduced contrast obstruct the extraction of semantic cues, posing significant challenges for current action recognition methods. In this paper, we mitigate the issues faced in action recognition under foggy conditions by employing two strategies. First, we present FogAct, the first benchmark dataset for foggy action recognition, consisting of paired clean and foggy videos captured with a stereo camera system. The dataset spans 10 scenes and 55 action categories, comprising nearly 10,000 video clips. Second, we propose FogNet, a two-stream CLIP model that discovers fog-invariant semantic information hidden behind the degraded videos. FogNet learns robust representations of foggy videos with guidance from clean videos, effectively capturing shared structural and motion cues between clean and foggy videos. Extensive experiments on FogAct and three other popular datasets demonstrate that our method achieves competitive performance compared with state-of-the-art (SOTA) approaches. Our FogAct and FogNet are given in our project page.

1. Introduction

Action recognition plays a critical role in surveillance [10, 12], autonomous driving [9, 15], and human-computer interaction [14], all requiring accurate action recognition under adverse weather conditions such as fog. Videos captured in foggy conditions, with reduced visibility, blurring, and low contrast, hinder robust feature extraction and complicate the detection of motion details invariant to fog, leading to reduced performance in existing action recognition approaches. This paper seeks to address these challenges and enable accurate foggy action recognition.

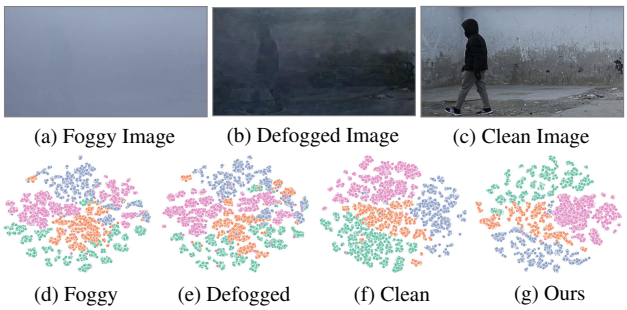


Figure 1. Comparison of foggy, defogged [5], and clean images in FogAct (top row), and corresponding feature distributions (bottom row). The SOTA defogging result still shows residual fog and halo artifacts. Features are extracted via CLIP and visualized using t-SNE. Our learned embeddings are more aligned with clean images, while defogged features show larger intra-class variation and blurred class boundaries.

Existing methods for foggy action recognition [2, 35] follow a two-stage framework. In the first stage, defogging modules are applied to restore relatively clean videos, which are then fed into the second action classification stage. However, these frameworks perform poorly on real-world foggy data as the defogging modules yield suboptimal restorations, ultimately degrading action recognition performance. As shown in Fig. 1b, the defogged example exhibits residual fog and halo artifacts. These issues stem from defogging modules relying on synthetic training data, limiting their performance in handling diverse foggy patterns and environmental variations in real-world conditions. To the best of our knowledge, no real foggy dataset featuring dynamic scenes for action recognition currently exists.

Beyond the error accumulation caused by poorly restored images, we also observe that defogging modules fail to recover action-related semantics satisfactorily. As shown in Fig. 1d and Fig. 1e, despite the application of defogging techniques, the semantic feature distribution demonstrates minimal improvement compared to clean videos (Fig. 1f), both in terms of intra-class compactness and classification

*Corresponding author.

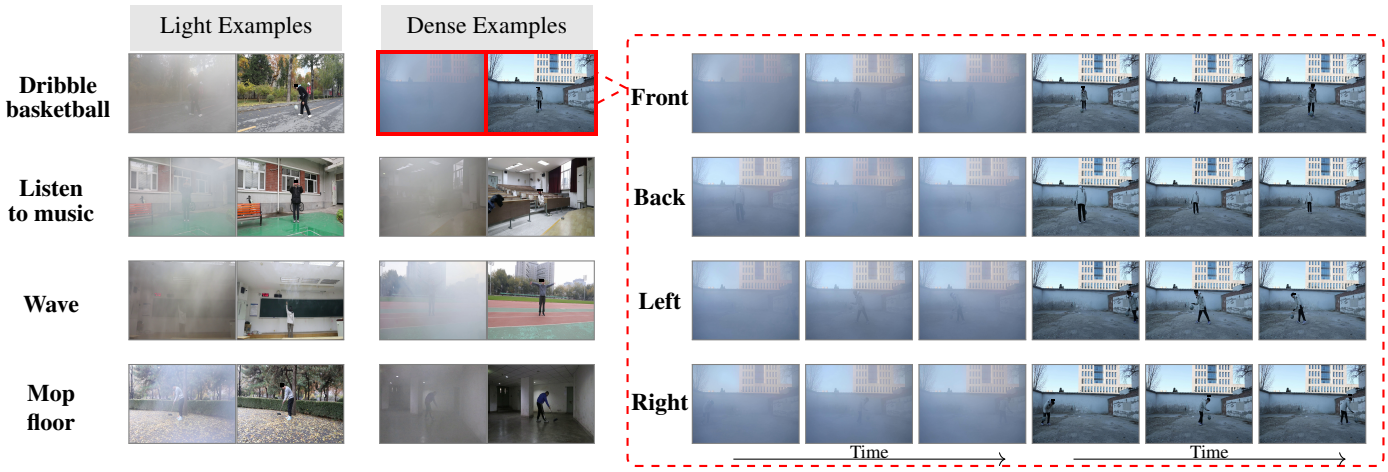


Figure 2. Examples from our FogAct dataset, including four categories. Each category is captured under two fog conditions: light fog examples and dense fog examples. Additionally, each action is recorded from four different perspectives: front, back, left, and right. To illustrate this, we use *'Dribble basketball'* as an example, showing four perspectives at a dense fog intensity level. Each sequence contains three frames sampled along the timeline. The three images on the left show foggy images, while the three on the right display the clean images. Additional examples are provided in the supplementary material.

boundaries. These limitations motivate further exploration to address this problem effectively.

In this paper, to alleviate the aforementioned issue, we present i) FogAct, the first benchmark dataset for foggy action recognition, and ii) FogNet, an end-to-end framework for foggy action recognition.

i) FogAct. Unlike existing methods [2, 35] that employ the atmospheric scattering model (ASM) [24] to simulate fog on datasets such as HMDB-51 [19] and UCF-101 [34], our FogAct dataset is collected from videos captured in real outdoor foggy environments. We designed a stereo video acquisition system that captures clean and foggy videos simultaneously, ensuring frame-wise semantic alignment. Our FogAct dataset consists of 55 distinct action categories captured across 10 different scenes, including construction sites, academic buildings, and classroom playgrounds, totaling 10,000 video clips, as shown in Fig. 2. Compared to existing datasets, FogAct introduces more unpredictable visibility degradation, motion blur, and reduced contrast, which are challenging to reproduce accurately using idealized atmospheric scattering models (ASM).

ii) FogNet. Unlike existing two-stage frameworks, our end-to-end FogNet bypasses traditional defogging by leveraging a two-stream CLIP model to find semantic similarities between foggy and clean videos. To acquire fog-invariant knowledge, FogNet integrates a Fog-Aware Selection (FAS) mechanism, generating semantically meaningful representations from both foggy and clean videos. Note that clean videos are used only during the training phase. A Mutual Enhancement (ME) module ensures complementary improvement, while a Cross-Stream Alignment (CSA) module aligns frames from clean and foggy video pairs, ensuring spatial correspondence and preserving temporal consistency, ultimately boosting recognition performance.

Our main contributions are as follows:

- We introduce FogAct, the first fog dataset with benchmarking for human action recognition, with $\approx$ 10K dynamic clean-foggy paired videos across 55 classes.
- We propose FogNet, the first end-to-end framework to leverage pre-trained vision-language models for foggy action recognition.
- Extensive experiments on real-world and synthetic datasets demonstrate that our method outperforms existing state-of-the-art approaches. Our code and dataset will be made available to facilitate reproducible research.

2. Related Work

Action Recognition. Unlike detection and segmentation [1, 13], foggy action recognition remains underexplored. Existing methods largely rely on handcrafted embeddings [3, 35], which are domain-specific and poorly adapt to complex, dynamic environments. They use synthetic datasets with idealized fog, missing real degradations like blur and contrast. Existing works fall into two categories. The first uses visual features with linear classifiers for direct action prediction [18, 22, 30, 43, 47]. The second leverages multimodal cues [4, 23, 38, 39, 44], using text to enrich visual representations. However, both rely on clean-scene distributions and struggle under fog. In contrast, our method extracts fog-invariant features, bridging domain gaps and enhancing robustness in foggy conditions.

Defogging and Foggy Datasets. Fog degrades visual features via atmospheric scattering. Defogging methods include physical and learning-based approaches. Physical

Table 1. Comparison of existing defogging datasets based on several key attributes. 'I&O' indicates whether the datasets include indoor or outdoor scenes. 'S&V' specifies whether the datasets contain single images or videos. 'Dyn.' denotes whether the dataset includes actions that evolve over time. 'Real' refers to whether foggy images are captured in real-world foggy conditions. 'Mul.' indicates multiple foggy intensity levels. 'Pair' highlights the inclusion of clean counterparts. 'Pers.' represents the number of perspectives. Finally, 'Anno.' specifies whether the datasets are annotated with action recognition labels.

Dataset	Venue	Scale	Resolution	I&O	S&V	Dyn.	Real	Mul.	Pers.	Pair	Anno.
Foggy Zurich [31]	ECCV'18	3808	1920 ×1080	O	S	✗	✓	✓	1	✗	✗
SOTS [20]	TIP'18	500	640 ×480	I	S	✗	✗	✓	1	✓	✗
NH-HAZE [8]	CVPR'20	55	1600 ×1200	O	S	✗	✓	✗	1	✓	✗
BeDDE [50]	TIP'20	208	1800 ×1200	O	S	✗	✓	✗	1	✓	✗
ACDC [33]	ICCV'21	4006	1920 ×1080	O	S	✗	✓	✗	1	✗	✗
RHVD [7]	QoMEX'21	403	1600 ×900	O	V	✗	✓	✗	1	✗	✗
REVIDE [49]	CVPR'21	1982	2708 ×1800	I	V	✗	✓	✓	1	✓	✗
VIREDA [11]	ASPAI'22	6	1670 ×1080	I	S	✗	✓	✓	1	✓	✗
HazeWorld [45]	CVPR'23	5084	1229 ×700	O	S	✗	✗	✓	1	✓	✗
Ours	2025	9724	1920 ×1080	I&O	V	✓	✓	✓	4	✓	✓

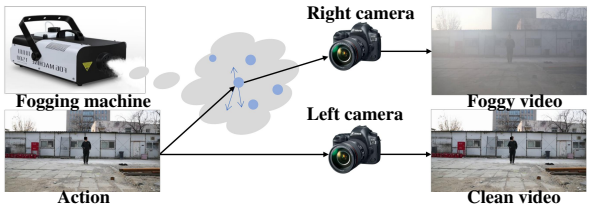


Figure 3. Overview of the Stereo Video Acquisition System. It includes two Canon DSLR cameras configured for stereo imaging, and a professional fogging machine. The lights from a scene point are reflected and refracted by blue fog particles, resulting in a foggy video in the right camera. In contrast, light directly forms a clean video in the left camera.

models (e.g., ASM[16, 37]) estimate transmission maps but rely on assumptions that often break in real fog. Learning-based methods (e.g., LDR [46], DEA-Net [6], EDI [28]) restore visibility via data-driven mappings but may introduce artifacts or distort motion, impairing action recognition. Most foggy datasets are synthetic (e.g., Foggy Cityscapes [32], NH-HAZE [8]), low-cost but unrealistic. Real datasets like Foggy Zurich [31] and EDHaze [48] are more realistic but static. In contrast, our FogAct captures diverse real-fog actions, motivating future exploration of defogging techniques in foggy environments.

3. FogAct Dataset

3.1. Data Collection and Annotation

To capture real-world foggy actions alongside their fog-free counterparts, we designed a stereo video acquisition system (Fig. 3). Both cameras share identical settings (focal length and lens) and are horizontally aligned with minimal separation. A professional fogging machine (DJPOWERE-1500W) generates high-quality fog before the right camera, producing realistic foggy action videos. The cameras are synchronized: the left records clean actions, and the right records foggy ones.

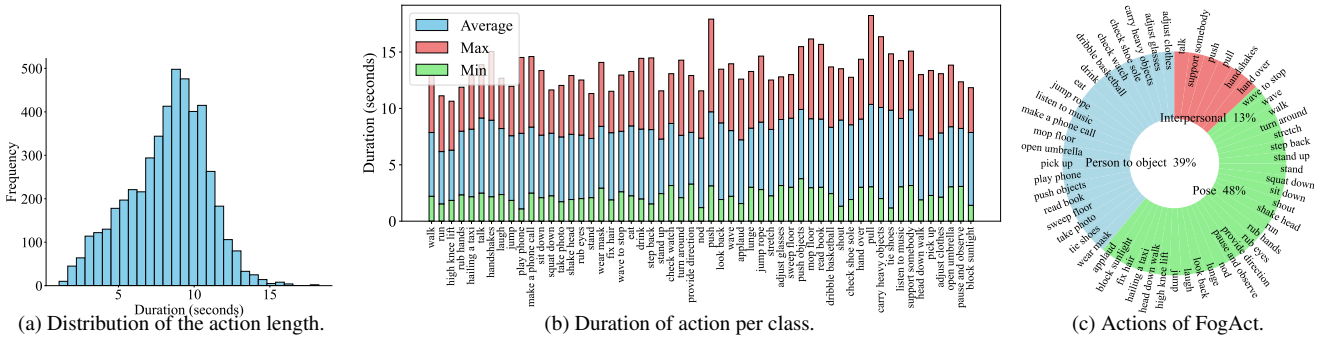
Table 2. KL divergence between FogAct, three simulated datasets, and several real-world fog datasets, including VIREDA [11], BeDDE [50], RHVD [7], SOTS [20], and Foggy Zurich [31].

	VIREDA	BeDDE	RHVD	SOTS	Zurich
HMDB-51	0.20	0.24	0.78	0.30	2.79
UCF-101	0.19	0.22	0.44	0.25	3.38
Kinetics-100	0.14	0.17	0.52	0.21	3.21
FogAct	**0.13**	**0.15**	**0.42**	**0.19**	**1.47**

The action data are collected from volunteers who are fully informed about the intended use of the dataset, ensuring compliance with ethical standards and legal requirements. Approval for data collection was obtained from the appropriate local ethics committee. The proposed FogAct dataset covers both indoor and outdoor scenes—such as construction sites, academic buildings, and classroom playgrounds—and comprises 55 distinct real-world foggy actions grouped into three categories, as shown in Fig. 4c. Each action video is recorded from four perspectives (front, back, left, and right) at 25 fps and 1920×1080 resolution. Videos are shot under light/dense fog across diverse scenes with varying camera positions. After data collection, all videos are segmented and annotated manually by two experienced experts. To ensure label reliability and consistency, a cross-validation process between annotators is conducted, and any inconsistencies are resolved through joint review and consensus.

3.2. Data Quality Analysis

The dataset contains 9724 videos, represented as triplets $\{(\mathbf{S}_\mathrm{f}, \mathbf{S}_\mathrm{c}, \mathbf{L})\}$, where $\mathbf{S}_\mathrm{f}$ and $\mathbf{S}_\mathrm{c}$ denote the foggy and cor-



(a) Distribution of the action length. (b) Duration of action per class. (c) Actions of FogAct.

Figure 4. Summary of FogAct statistics. (a) 91.8% of samples last 3–12 seconds, resembling a normal distribution. (b) Action durations across classes show minimal variance, ensuring temporal balance. (c) Actions are grouped into three categories based on interaction patterns, reflecting the dataset's diversity. Best viewed in color.

responding clean videos, respectively. $\mathbf{L}$ denotes the action category, annotated by two experts. Among them, 552 are dual-person action pairs and 4,310 are single-person action pairs, covering diverse human behaviors under real fog. Fig. 4a shows 91.8% of the actions have durations between 3 and 12 seconds, with an average length of 8.27 seconds. Fig. 4b illustrates duration distributions by category. For training and evaluation, we randomly split the dataset into 80% training and 20% testing, ensuring robust model performance on unseen data.

KL Distribution. In Tab. 2, we compute the KL divergence between our FogAct dataset and four real-world fog datasets: VIREDA [11], BeDDE [50], RHVD [7], and SOTS [20]. We also compare the three simulated fog datasets: HMDB-51, UCF-101, and Kinetics-100, which are generated through ASM simulation, with the real-world fog datasets. To simplify the computation, we randomly sample two videos per action category from all datasets. The results indicate that our dataset exhibits a lower KL divergence with all real fog datasets, suggesting that FogAct better aligns with real-world fog distributions.

Evaluation of Vision-Language Models. We assess fog realism using VLM-based metrics. With Q-Align [42], FogAct obtains a realism score of 0.32, markedly higher than the synthetic dataset's 0.13. GPT-4o yields similar trends, scoring 3.45 vs. 2.47, respectively. These results verify the superior realism of FogAct (see supplemental for prompts).

Human Evaluation. To further assess the quality of FogAct, we conduct a human study with 21 participants (all graduate-level or above). Each participant is given 30 randomly sampled image pairs, each containing one FogAct image and one ASM-generated fog image. The study is blind, and participants rate which image has higher quality on a 1–5 scale (1 = lowest, 5 = highest). An example is shown in Fig. 7 of the supplementary material.

4. FogNet

Overview. We extract fog-invariant embeddings via the fog-invariant feature extractor, which integrates fog-aware selection, mutual enhancement, and cross-stream alignment. Given a foggy video sequence $\mathbf{S}_\mathrm{f} = \{\mathbf{I}_\mathrm{f}\}_{i=1}^{T}$ and its clean counterpart $\mathbf{S}_\mathrm{c} = \{\mathbf{I}_\mathrm{c}\}_{i=1}^{T}$, each with T uniformly sampled frames, we encode them into visual embeddings, $\mathbf{v}_\mathrm{f}$ and $\mathbf{v}_\mathrm{c}$, using a visual encoder. The fog-aware selection component employs global self-attention to identify semantically meaningful embeddings $\mathbf{v}_\mathrm{f}^\mathrm{A}$ and $\mathbf{v}_\mathrm{c}^\mathrm{A}$ from both streams. Next, the mutual enhancement component refines two embeddings via bidirectional cross-attention to obtain $\mathbf{v}_\mathrm{f}^\mathrm{D}$ and $\mathbf{v}_\mathrm{c}^\mathrm{D}$. Finally, leveraging the inherent temporal consistency between foggy and clean videos, we perform frame-level alignment between $\mathbf{v}_\mathrm{f}^\mathrm{D}$ and $\mathbf{v}_\mathrm{c}^\mathrm{D}$. To classify foggy videos, C action prompts $\mathbf{T}_{\mathrm{c}=1}^{C}$ are embedded as $\{\mathbf{t}_\mathrm{c}\}_{\mathrm{c}=1}^{C}$ via a textual encoder. The classification is performed by selecting the class $\hat{\mathrm{c}}$ whose textual embedding exhibits the highest cosine similarity with the fog-invariant embedding, *i.e.*, $\hat{\mathrm{c}} = \arg\max \mathrm{sim}(\mathbf{v}_\mathrm{f}^\mathrm{D}, \mathbf{t}_\mathrm{c})$. Fig. 5 illustrates our framework.

4.1. Fog-Aware Selection

Besides irrelevant background cues shared by clean and foggy embeddings, $\mathbf{v}_\mathrm{f}$ also contains fog-induced degradation. Directly aligning $\mathbf{v}_\mathrm{f}$ with text embeddings introduces noise and weakens classification performance.

To mitigate this, we apply a global self-attention module to jointly process $\mathbf{v}_\mathrm{c}$ and $\mathbf{v}_\mathrm{f}$, emphasizing features that are stable in clean conditions and discriminative under fog. This suppresses fog artifacts and strengthens semantic representations. The process is given by

$$\mathbf{v}_\mathrm{att} = \sum_{i=1}^{T} \frac{\exp\left(\mathrm{sim}\left(\mathbf{v}^\mathrm{cat}, \mathbf{v}^\mathrm{cat}{}_i\right)\right)}{\sum_{j=1}^{T} \exp\left(\mathrm{sim}\left(\mathbf{v}^\mathrm{cat}, \mathbf{v}^\mathrm{cat}{}_j\right)\right)} \mathbf{v}_i^\mathrm{cat}, \qquad (1)$$

where $\mathbf{v}^\mathrm{cat} = [\mathbf{v}_\mathrm{c}; \mathbf{v}_\mathrm{f}]$ denotes concatenated clean and foggy

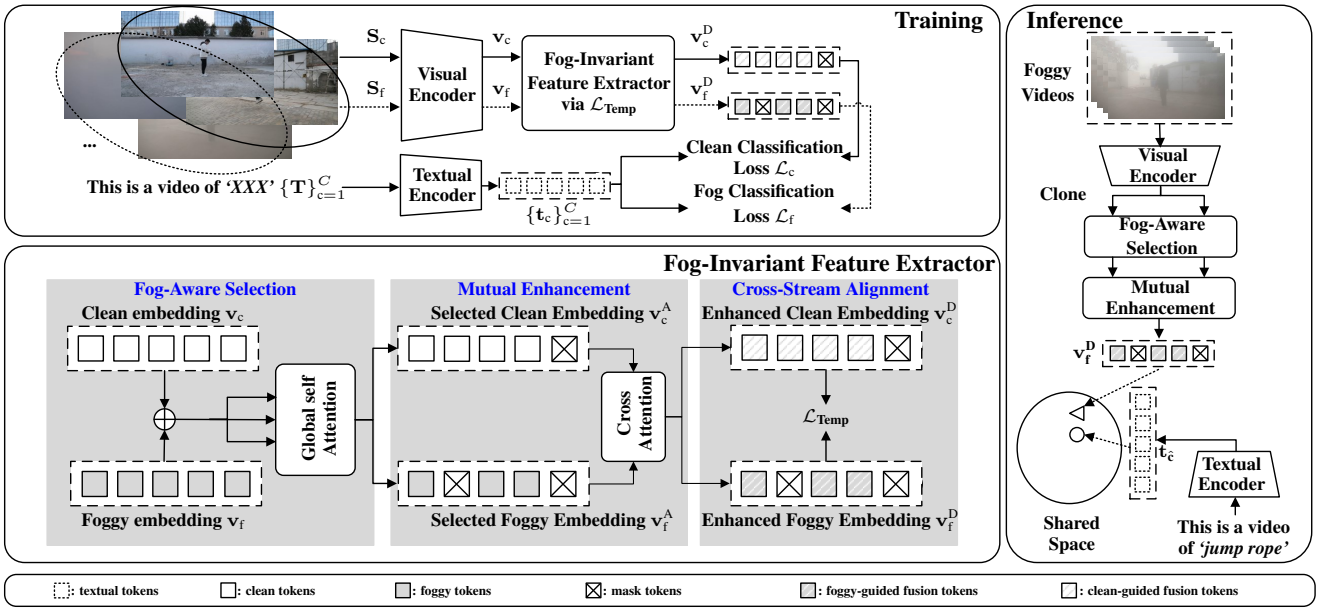


Figure 5. An overview of our framework. In the joint training stage, we jointly learn label supervision and fog-invariant representations across clean and foggy videos. Fog-invariant feature extractor involves three key components: i) Fog-Aware Selection. $\mathbf{v}_{\text{c}}$ and $\mathbf{v}_{\text{f}}$ are fused via global self-attention to highlight semantically relevant regions. ii) Mutual Enhancement. Bidirectional cross-attention enables semantic interaction between $\mathbf{v}_{\text{c}}^{\text{A}}$ and $\mathbf{v}_{\text{f}}^{\text{A}}$ while retaining domain-specific traits. iii) Cross-Stream Alignment. $\mathbf{v}_{\text{c}}^{\text{D}}$ and $\mathbf{v}_{\text{f}}^{\text{D}}$ are aligned at the frame level, with the latter used for recognition. During inference, only foggy videos are used as input and serve as the query, key, and value for the Fog-Aware Selection, followed by Mutual Enhancement to obtain $\mathbf{v}_{\text{f}}^{\text{D}}$. Prediction $\hat{\text{c}}$ is made by the nearest text embedding $\mathbf{t}_{\hat{\text{c}}}$ to $\mathbf{v}_{\text{f}}^{\text{D}}$.

embeddings, and $\mathbf{v}_i^{\text{cat}}$ represents the i-th element of $\mathbf{v}^{\text{cat}}$. We then split $\mathbf{v}_{\text{att}}$ back into separate clean $\mathbf{v}_{\text{c}}^{\text{A}}$ and foggy $\mathbf{v}_{\text{f}}^{\text{A}}$ as

$$\mathbf{v}_{\text{c}}^{\text{A}}, \mathbf{v}_{\text{f}}^{\text{A}} = \text{chunk}\left(\mathbf{v}_{\text{att}}, 2\right). \tag{2}$$

4.2. Mutual Enhancement

While the selected $\mathbf{v}_{\text{f}}^{\text{A}}$ and $\mathbf{v}_{\text{c}}^{\text{A}}$ capture important and discriminative features, this process may miss some key information. The clean embedding helps guide the optimization of the foggy embedding to some extent, and vice versa.

To enhance fog-invariant representations, bidirectional cross-modal attention mutually refines $\mathbf{v}_{\text{f}}^{\text{A}}$ and $\mathbf{v}_{\text{c}}^{\text{A}}$, with the clean embedding querying the foggy one for refinement.

$$\mathbf{Q}_{\text{c}} = \mathbf{v}_{\text{c}}^{\text{A}}\mathbf{W}^{\text{q}}, \quad \mathbf{K}_{\text{f}} = \mathbf{v}_{\text{f}}^{\text{A}}\mathbf{W}^{\text{k}}, \quad \mathbf{V}_{\text{f}} = \mathbf{v}_{\text{f}}^{\text{A}}\mathbf{W}^{\text{v}}, \tag{3}$$

$$\mathbf{v}_{\text{f}}^{\text{D}} = \text{Attention}(\mathbf{Q}_{\text{c}}, \mathbf{K}_{\text{f}}, \mathbf{V}_{\text{f}}) + \mathbf{v}_{\text{f}}^{\text{A}}, \tag{4}$$

where $\mathbf{W}^{\text{q}}$, $\mathbf{W}^{\text{k}}$, and $\mathbf{W}^{\text{v}}$ project the query $\mathbf{Q}_{\text{c}}$, key $\mathbf{K}_{\text{f}}$, and value $\mathbf{V}_{\text{f}}$. We then reverse the roles, using foggy embeddings as the query and clean embeddings as key and value,

$$\mathbf{v}_{\text{c}}^{\text{D}} = \text{Attention}(\mathbf{Q}_{\text{f}}, \mathbf{K}_{\text{c}}, \mathbf{V}_{\text{c}}) + \mathbf{v}_{\text{c}}^{\text{A}}. \tag{5}$$

4.3. Cross-Stream Alignment

Considering fog does not distort action semantics, temporal dynamics should be preserved for fog-invariant embeddings. To achieve this, we employ cross-stream alignment to align the enhanced clean embedding $\mathbf{v}_{\text{c}}^{\text{D}}$ and the foggy embedding $\mathbf{v}_{\text{f}}^{\text{D}}$, which extracts fog-invariant embeddings by maintaining temporal consistency across varying conditions. Concretely, we construct a consistency matrix s^{c} to align foggy and clean video embeddings at the frame-level

$$s^{\text{c}} = \text{sim}\left(\mathbf{v}_{\text{f}}^{\text{D}}, \mathbf{v}_{\text{c}}^{\text{D}}\right). \tag{6}$$

Each element $s_{i,j}^{\text{c}}$ represents the correlation between the i-th frame of the foggy sequence and the j-th frame of the clean sequence. To learn temporal alignments between clean and foggy videos, we optimize with a contrastive loss,

$$\mathcal{L}_{\text{Temp}} = \frac{1}{T}\sum_{i\in T}\log\frac{\exp(s_{i,i}^{\text{c}})}{\sum_{j=1}^{T}\exp(s_{i,j}^{\text{c}})}. \tag{7}$$

4.4. Loss

We optimize the network to maximize the similarity $s_{\text{b,c}^*}$ between each foggy video embedding and its ground-truth text embedding, while suppressing similarities to all other classes $\{s_{\text{b,c}}\}_{\text{c}=1,\text{c}\neq\text{c}^*}^{C}$. Following [26], we adopt the InfoNCE loss.

$$\mathcal{L}_{\text{f}}^{\text{T2V}} = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{|\mathbf{k}_{\text{b}}|}\sum_{k\in\mathbf{k}_{\text{b}}}\log\frac{\exp(s_{k,\text{c}^*})}{\sum_{j=1}^{B}\exp(s_{j,\text{c}^*})} \tag{8}$$

Table 3. Comparison results on the FogAct dataset. We report the GFLOPs and parameters in the inference phase. 'Views' indicates # temporal clip × # spatial crop. 'LLaVA1.5-VI' indicates LLaVA1.5-VideoChatGPT-Instruct. The magnitude is Million (10^6) for parameters (Param). 'Pre-training' indicates training data: for the backbone in one-stage, and the defogger in two-stage methods. For methods with two stages, a defogging method is first applied to remove the fog, followed by the use of the SOTA action recognition method: OST [4] for classification. We achieve the highest Top-1 and Top-5 accuracy by employing an 8-frame RGB input evaluated with a single view. The best numbers are highlighted in ***bold***.

Method	Venue	Input	Pre-training	Top-1(%)	Top-5(%)	Views	GFLOPs	Param
Methods with one stage								
VideoMAE ViT-B/16 [36]	NeurIPS'22	8×224^2	Kinetics-400	11.1	31.1	1×1	90	87
DINO-v2 [27]	arXiv'23	8×224^2	LVD-142M	45.4	74.2	1×1	-	86.6
LLaMa-VID [21]	ECCV'24	8×224^2	LLaVA1.5-VI	26.2	-	-	-	7000
X-CLIP ViT-B/16 [25]	ECCV'22	8×224^2	Kinetics-400	67.4	92.8	1×1	145	131.5
ActionCLIP ViT-B/16 [38]	TNNLS'23	8×224^2	WIT-400M	75.0	95.7	1×1	141	141.7
AIM ViT-B/16 [47]	ICLR'23	8×224^2	WIT-400M	80.3	98.0	1×1	202	96.4
ATM ViT-B/16 [43]	ICCV'23	8×224^2	WIT-400M	73.2	94.5	1×1	95	87.8
Vita-CLIP ViT-B/16 [41]	CVPR'23	8×224^2	WIT-400M	18.8	43.8	1×1	97	161.8
ViFi-CLIP ViT-B/16 [30]	CVPR'23	8×224^2	WIT-400M	78.4	97.7	1×1	141	124.7
BIKE ViT-B/16 [44]	CVPR'23	8×224^2	WIT-400M	13.4	19.5	1×1	-	124.1
M^2-CLIP ViT-B/16 [39]	AAAI'24	8×224^2	WIT-400M	56.5	88.8	1×1	214	165
TC-CLIP ViT-B/16 [18]	ECCV'24	8×224^2	WIT-400M	71.5	95.0	1×1	-	127.5
SFAR ViT-B/16 [23]	NeurIPS'25	8×224^2	WIT-400M	73.6	95.4	1×1	90.2	126.1
OST ViT-B/16 [4]	CVPR'24	8×224^2	WIT-400M	83.2	98.9	1×1	141	149.6
Methods with two stages								
UCL [40] + OST	TIP'24	8×224^2	Unsupervised	58.0	86.9	1×1	-	169.1
PTTD [5] + OST	ECCV'24	8×224^2	DIV&Flickr	85.2	98.2	1×1	-	152.2
LDR [46] + OST	CVPR'24	8×224^2	All-weather	83.3	98.6	1×1	776	163.6
PTTD [5] + OST	ECCV'24	8×224^2	FogAct	85.4	98.8	1×1	-	152.2
Ours ViT-B/16	2025	8×224^2	WIT-400M	**88.7**	**99.4**	**1×1**	143	146.9

$$\mathcal{L}_{\mathrm{f}}^{\mathrm{V2T}} = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{|\mathbf{k}_{\mathrm{b}}|}\sum_{k\in\mathbf{k}_{\mathrm{b}}}\log\frac{\exp(s_{k,c^*})}{\sum_{\mathrm{c}=1}^{C}\exp(s_{k,\mathrm{c}})} \tag{9}$$

$$\mathcal{L}_{\mathrm{f}} = \mathcal{L}_{\mathrm{f}}^{\mathrm{T2V}} + \mathcal{L}_{\mathrm{f}}^{\mathrm{V2T}} \tag{10}$$

where B is the batch size, $\mathbf{kb}$ denotes indices of videos sharing the same class as the b-th video, and $|\mathbf{kb}|$ is their count.

Similarly, we define the loss for aligning text embeddings with clean video features as $\mathcal{L}_{\mathrm{c}} = \mathcal{L}_{\mathrm{c}}^{\mathrm{T2V}} + \mathcal{L}_{\mathrm{c}}^{\mathrm{V2T}}$, computed analogously to $\mathcal{L}_{\mathrm{f}}^{\mathrm{T2V}}$ and $\mathcal{L}_{\mathrm{f}}^{\mathrm{V2T}}$ but using clean instead of foggy embeddings. The total loss is formulated as:

$$\mathcal{L}_{\mathrm{all}} = \mathcal{L}_{\mathrm{f}} + \lambda\mathcal{L}_{\mathrm{c}} + \beta\mathcal{L}_{\mathrm{Temp}}, \tag{11}$$

where $\lambda = 0.4$, $\beta = 0.1$ are hyperparameters.

5. Experiment

5.1. Experiment Setup

Datasets. We conduct experiments on four benchmarks: the collected FogAct dataset and three simulated datasets (apply ASM [24] to UCF-101 [34], HMDB-51 [19], and Kinetics-100 [17] following [35]). In addition, two experts manually annotate the real-world foggy action videos.

Table 4. Comparison on the synthesized datasets, UCF-101 [34], HMDB-51 [19], and Kinetics-100 [17].

Method	UCF	HMDB	Kinetics	Average
ActionCLIP [38]	84.3	57.1	70.6	70.7
AIM [47]	90.0	64.7	79.2	78.0
ATM [43]	90.2	63.4	77.2	76.9
M^2-CLIP [39]	89.0	62.0	78.0	76.3
OST [4]	92.4	69.1	75.4	79.0
Ours	**93.2**	**71.1**	**85.2**	**83.2**

Implementation Details. We initialize our network with CLIP [29] pre-trained on WIT-400M. Then, we fine-tune the model for 30 epochs using the AdamW optimizer with a batch size of 128. Learning rate is 5×10^{-5} with cosine annealing and 5-epoch warm-up.

Compared Methods. We compare two paradigms: **one-stage** and **two-stage**. One-stage methods directly recognize actions in foggy videos, while two-stage methods first defog the videos and then apply OST [4]. Unless otherwise specified, ablations are conducted on FogAct with Action-

Table 5. Comparative experiments are conducted on UCF-101 , HMDB-51 , Kinetics-100 (denoted as 'UCF', 'HMDB', and 'K100', respectively), and the FogAct dataset. We report accuracy (%) for a single 8-frame clip with a spatial resolution of 224×224, unless otherwise specified. 'Pers.' indicates perspectives. 'Sim.' indicates simulated.

FAS	ME	CSA	Top-1(%)	Top-5(%)
Baseline			75.0	95.7
✓	✗	✗	86.9	98.9
✗	✓	✗	79.4	97.2
✗	✗	✓	84.1	98.5
✓	✓	✗	88.1	99.4
✓	✓	✓	**88.7**	**99.4**

(a) The effectiveness of our model components.

Model	Train	Test	Top-1(%)	Top-5(%)
ActionCLIP	clean	clean	86.7	98.9
	foggy	foggy	75.0	95.7
	clean	foggy	59.8	85.9
	clean+foggy	foggy	75.7	96.3
FogNet	clean+sim.	foggy	71.5	93.6
	clean+foggy	foggy	**88.7**	**99.4**

(b) Comparison for different input settings on FogAct.

Pers.	Top-1	Top-5
Single	52.7	79.4
Four	**88.7**	**99.4**

(c) Results of models trained on single or four perspectives and evaluated on four.

Backbone	Top-1	+ FAS	+ ME	+ CSA
RN50	56.1	70.2	71.0	71.5
ViT-B/32	63.0	78.8	79.0	79.2
ViT-B/16	75.0	86.9	88.1	88.7

(d) Evaluation of using backbones with different components. The frame rate is set to 8.

Frames	Views	FogAct	HMDB	UCF	K100
4	1×1	85.2	69.6	92.5	82.9
4	4×1	86.6	70.4	93.3	85.1
8	1×1	88.7	**71.1**	93.2	85.1
8	4×1	**89.1**	70.9	**93.4**	**85.4**

(e) Effects of different inference schemes.

Intensity	Views	Top-1(%)	Top-5(%)
Light	1×1	66.0	87.2
Dense	1×1	67.5	92.2
Multi	1×1	**88.7**	**99.4**

(f) Results of different fog levels.

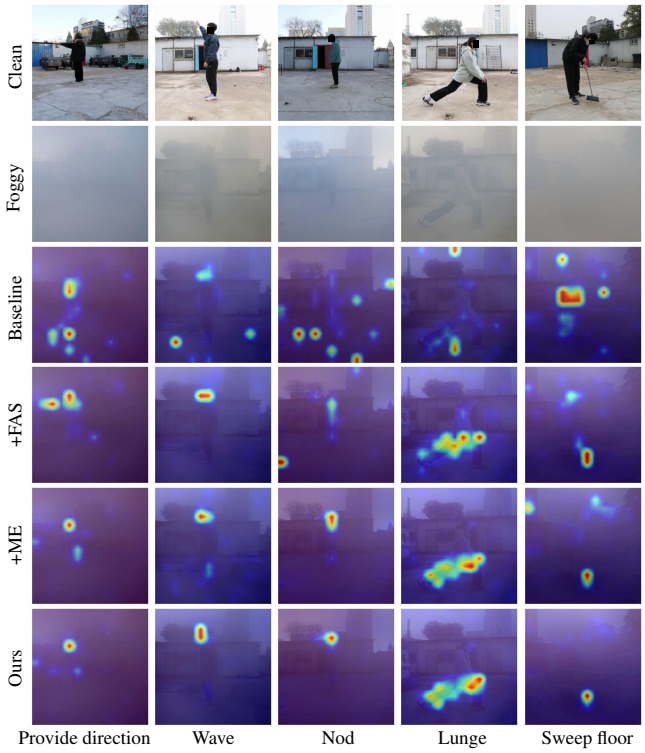


Figure 6. Heatmap comparisons with the baseline on FogAct.

CLIP [38] as the default baseline.

5.2. Experimental Results

The recognition results comparing our method with others are shown in Tab. 3. **One-stage methods**. (i) Image reconstruction-based methods perform suboptimally, with even the best-performing approach in this category, DINO-based frameworks [27], achieving only a 45% Top-1 accuracy. This is because they primarily focus on restoring visual details but struggle with action-relevant semantic features. (ii) Video understanding method LLaMa-VID [21] achieves only 26.2% accuracy. Despite leveraging large language models, it lacks effective temporal modeling under fog, failing to extract fog-invariant features from low-quality, blurred videos. (iii) Action recognition methods relying solely on visual signals [18, 30, 47] or vision-language-based strategies [4, 38, 39] perform well in standard settings. However, they struggle under fog. This highlights the difficulty of transferring pre-trained models to degraded environments, where low visibility impairs feature extraction and action recognition.

The **two-stage approaches** combine the SOTA defogging works with SOTA action recognition approach. Regardless of using paired [5] or unpaired [40] defogging strategies, they focus on restoring visual details yet struggle to recover action-relevant semantics. We even introduce the SOTA all-in-one approach [46], which claims better generalization as it is trained on all-weather datasets. However, the two-stage framework still struggles to improve classification performance of the defogged video embeddings. The best two-stage setup, combining the SOTA defogging method PTTD (fine-tuned on FogAct) and the SOTA action recognition model OST, achieves only 85.4% Top-1 accuracy. In comparison, our one-stage approach bypasses the image-defogging step and achieves the best results, with 88.7% Top-1 accuracy and 99.4% Top-5 accuracy.

Additionally, to further demonstrate the effectiveness of our framework, we conduct experiments on the simulated datasets HMDB-51, UCF-101, and Kinetics-100, synthe-

sized using common practices. As shown in Tab. 4, our method outperforms OST [4] by 9.8% on the Kinetics-100 dataset, with an average accuracy improvement of 4.2% across all three simulated datasets.

5.3. Ablation Study and Analysis

Model Architecture. We study the effectiveness of our three components in Tab. 5a, including Fog-Aware Selection (FAS), Mutual Enhancement (ME), and Cross-Stream Alignment (CSA). We observe that the three components all enhance action recognition accuracy in foggy conditions. The FAS structure outperforms the baseline by 13%, while ME and CSA contribute increases of 6% and 11%, respectively. When combined, using all components achieves a Top-1 accuracy of 88.71% and a Top-5 accuracy of 99.40%.

Attention Visualization. As shown in Fig. 6, our model effectively highlights action-relevant regions under fog by progressively refining attention. This improved focus over the baseline supports the effectiveness of our components, consistent with 'Model Architecture'.

Necessity of FogNet and FogAct. Tab. 5b shows that models trained on clean data degrade under fog due to low visibility, and defogging methods help only slightly while ignoring action semantics. Clean-trained models generalize poorly under fog due to domain shifts. In contrast, FogNet learns fog-invariant features, improving Top-1 accuracy by 13% under fog. Evaluations on ASM-simulated versus real-world foggy data further show that FogAct is more challenging and better reflects practical scenarios.

Effectiveness of Multiple Perspectives. We compare single- and multi-perspective training on FogAct, evaluating both on all views as shown in Tab. 5c. Using all four perspectives increases Top-1 by 36.0% and Top-5 by 20.0%. To control for data volume, we replicate the single-view data to match the size of the multi-view set.

Different Backbones. Tab. 5d presents an evaluation of the applicability of our method using different backbones. We observe that the effectiveness of the proposed modules remains consistent across different backbone architectures.

Analysis of Inference. Inference performance is strongly influenced by frame rate and views (# temporal clip $\times$ # spatial crop), as shown in Tab. 5e. Increasing frames from 4 to 8 yields consistent accuracy gains across all datasets, highlighting the importance of temporal richness. Furthermore, using 4 temporal clips significantly boosts Top-1 accuracy over a single clip on FogAct, UCF-101, and Kinetics-100.

Fog Intensity Analysis. Tab. 5f shows training with multiple fog intensities outperforms single-intensity settings. Models trained on dense fog generalize better than those on light fog. Even when controlling data volume via replication, the multi-intensity model achieves the best performance, surpassing dense fog by 21.2% in Top-1 accuracy.

Confusion Matrix. Fig. 7 presents the confusion matrix

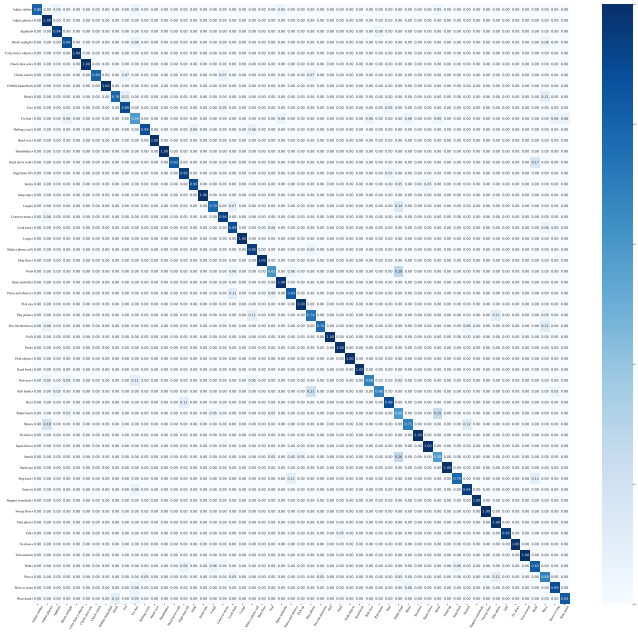


Figure 7. Confusion matrix of our FogNet on the FogAct dataset.

on FogAct. FogNet delivers high accuracy on actions with distinct motion patterns, such as *Talk*, *Adjust Glasses*, and *Hand Over*. Performance drops on subtle or short-duration actions like *Stand*, *Fix Hair*, and *Nod*, which exhibit high visual similarity and are easily obscured under dense fog.

6. Conclusion

In this paper, we present FogAct, the first real-world dataset for foggy action recognition with foggy–clean video pairs, covering 10 scenes and 55 action categories, providing sufficient training data and benchmarking. We further propose FogNet, an end-to-end framework that extracts fog-invariant features through three components: fog-aware selection, mutual enhancement, and cross-stream alignment. Leveraging large-scale pre-trained models, FogNet demonstrates strong generalization on real-world data. Extensive experiments on four challenging benchmarks (FogAct, UCF-101, HMDB-51, and Kinetics-100) validate the effectiveness of both FogAct and FogNet.

7. Acknowledgement

This work is supported by the National Natural Science Foundation of China (62302045), the Fundamental Research Funds for the Central Universities, and the BIT Special-Zone. This work is supported (in part) by the Opening Project of the State Key Laboratory of General Artificial Intelligence, BIGAI/Peking University, Beijing, China. (Project NO. SKLAGI20250P05).

References

[1] Qi Bi, Shaodi You, and Theo Gevers. Learning generalized segmentation for foggy-scenes by bi-directional wavelet guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 801–809, 2024. 2

[2] Sachin Chaudhary and Subrahmanyam Murala. Tsnet: deep network for human action recognition in hazy videos. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3981–3986. IEEE, 2018. 1, 2

[3] Sachin Chaudhary and Subrahmanyam Murala. Depth-based end-to-end deep network for human action recognition. *IET Computer Vision*, 13(1):15–22, 2019. 2

[4] Tongjia Chen, Hongshan Yu, Zhengeng Yang, Zechuan Li, Wei Sun, and Chen Chen. Ost: Refining text knowledge with optimal spatio-temporal descriptor for general video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18888–18898, 2024. 2, 6, 7, 8

[5] Zixuan Chen, Zewei He, Ziqian Lu, Xuecheng Sun, and Zhe-Ming Lu. Prompt-based test-time real image dehazing: a novel pipeline. In *European Conference on Computer Vision*, pages 432–449. Springer, 2024. 1, 6, 7

[6] Zixuan Chen, Zewei He, and Zhe-Ming Lu. Dea-net: Single image dehazing based on detail-enhanced convolution and content-guided attention. *IEEE Transactions on Image Processing*, 2024. 3

[7] Ying Chu, Guoxing Luo, and Fan Chen. A real haze video database for haze level evaluation. In *2021 13th International Conference on Quality of Multimedia Experience (QoMEX)*, pages 69–72. IEEE, 2021. 3, 4

[8] Ancuti C.O. et al. Nh-haze: An image dehazing benchmark with non-homogeneous hazy and haze-free images. In *CVPR Workshops*, 2020. 3

[9] Jingtao Dong, Hao Zhuang, Hao Yang, and Liyuan Pan. Rgb-event fusion for robust lane detection. In *BMVC*, 2025. 1

[10] Keval Doshi and Yasin Yilmaz. Multi-task learning for video surveillance with limited data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3889–3899, 2022. 1

[11] Alexandra Duminil, Jean-Philippe Tarel, and Roland Brémond. A new real-world video dataset for the comparison of defogging algorithms. *arXiv preprint arXiv:2310.01020*, 2023. 3, 4

[12] Hao Fang, Ajian Liu, Jun Wan, Sergio Escalera, Hugo Jair Escalante, and Zhen Lei. Surveillance face presentation attack detection challenge. in 2023 ieee. In *CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 6361–6371. 1

[13] Himanshu Gupta, Oleksandr Kotlyar, Henrik Andreasson, and Achim J Lilienthal. Robust object detection in challenging weather conditions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7523–7532, 2024. 2

[14] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J Black. Populating 3d scenes by learning human-scene interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14708–14718, 2021. 1

[15] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17853–17862, 2023. 1

[16] Md Tanvir Islam, Nasir Rahim, Saeed Anwar, Muhammad Saqib, Sambit Bakshi, and Khan Muhammad. Hazespace2m: A dataset for haze aware single image dehazing. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9155–9164, 2024. 3

[17] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 6

[18] Minji Kim, Dongyoon Han, Taekyung Kim, and Bohyung Han. Leveraging temporal contextualization for video action recognition. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024. 2, 6, 7

[19] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 2, 6

[20] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28(1):492–505, 2018. 3, 4

[21] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024. 6, 7

[22] Enqi Liu and Liyuan Pan. A lightweight multi-level relation network for few-shot action recognition. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024. 2

[23] Enqi Liu, Liyuan Pan, Yan Yang, Yiran Zhong, Zhijing Wu, Xinxiao Wu, and Liu Liu. Storyboard guided alignment for fine-grained video action recognition. *arXiv preprint arXiv:2410.14238*, 2024. 2, 6

[24] Srinivasa G. Narasimhan and Shree K. Nayar. Contrast restoration of weather degraded images. *IEEE transactions on pattern analysis and machine intelligence*, 25(6):713–724, 2003. 2, 6

[25] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *European conference on computer vision*, pages 1–18. Springer, 2022. 6

[26] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5

[27] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al.

Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 6, 7

[28] Liyuan Pan, Cedric Scheerlinck, Xin Yu, Richard Hartley, Miaomiao Liu, and Yuchao Dai. Bringing a blurry frame alive at high frame-rate with an event camera. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6820–6829, 2019. 3

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6

[30] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6545–6554, 2023. 2, 6, 7

[31] Christos Sakaridis, Dengxin Dai, Simon Hecker, and Luc Van Gool. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In *Proceedings of the european conference on computer vision (ECCV)*, pages 687–704, 2018. 3

[32] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126:973–992, 2018. 3

[33] Christos Sakaridis, Haoran Wang, Ke Li, René Zurbrügg, Arpit Jadon, Wim Abbeloos, Daniel Olmeda Reino, Luc Van Gool, and Dengxin Dai. Acdc: The adverse conditions dataset with correspondences for robust semantic driving scene perception. *arXiv e-prints*, pages arXiv–2104, 2021. 3

[34] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. A dataset of 101 human action classes from videos in the wild. *Center for Research in Computer Vision*, 2(11):1–7, 2012. 2, 6

[35] Sri Girinadh Tanneru and Snehasis Mukherjee. Action recognition in haze using an efficient fusion of spatial and temporal features. In *Computer Vision and Image Processing: 5th International Conference, CVIP 2020, Prayagraj, India, December 4-6, 2020, Revised Selected Papers, Part II 5*, pages 29–38. Springer, 2021. 1, 2, 6

[36] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 6

[37] Hayat Ullah, Khan Muhammad, Muhammad Irfan, Saeed Anwar, Muhammad Sajjad, Ali Shariq Imran, and Victor Hugo C de Albuquerque. Light-dehazenet: a novel lightweight cnn architecture for single image dehazing. *IEEE transactions on image processing*, 30:8968–8982, 2021. 3

[38] Mengmeng Wang, Jiazheng Xing, Jianbiao Mei, Yong Liu, and Yunliang Jiang. Actionclip: Adapting language-image pretrained models for video action recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 2, 6, 7

[39] Mengmeng Wang, Jiazheng Xing, Boyuan Jiang, Jun Chen, Jianbiao Mei, Xingxing Zuo, Guang Dai, Jingdong Wang, and Yong Liu. A multimodal, multi-task adapting framework for video action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5517–5525, 2024. 2, 6, 7

[40] Yongzhen Wang, Xuefeng Yan, Fu Lee Wang, Haoran Xie, Wenhan Yang, Xiao-Ping Zhang, Jing Qin, and Mingqiang Wei. Ucl-dehaze: Towards real-world image dehazing via unsupervised contrastive learning. *IEEE Transactions on Image Processing*, 2024. 6, 7

[41] Syed Talal Wasim, Muzammal Naseer, Salman Khan, Fahad Shahbaz Khan, and Mubarak Shah. Vita-clip: Video and text adaptive clip via multimodal prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23034–23044, 2023. 6

[42] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 4

[43] Wenhao Wu, Yuxin Song, Zhun Sun, Jingdong Wang, Chang Xu, and Wanli Ouyang. What can simple arithmetic operations do for temporal modeling? In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13712–13722, 2023. 2, 6

[44] Wenhao Wu, Xiaohan Wang, Haipeng Luo, Jingdong Wang, Yi Yang, and Wanli Ouyang. Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6620–6630, 2023. 2, 6

[45] Jiaqi Xu, Xiaowei Hu, Lei Zhu, Qi Dou, Jifeng Dai, Yu Qiao, and Pheng-Ann Heng. Video dehazing via a multi-range temporal alignment network with physical prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18053–18062, 2023. 3

[46] Hao Yang, Liyuan Pan, Yan Yang, and Wei Liang. Language-driven all-in-one adverse weather removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24902–24912, 2024. 3, 6, 7

[47] Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. Aim: Adapting image models for efficient video action recognition. *arXiv preprint arXiv:2302.03024*, 2023. 2, 6, 7

[48] Ruikun Zhang, Zhiyuan Yang, and Liyuan Pan. Dehaze-mamba: large multi-modal model guided single image dehazing via mamba. *Visual Intelligence*, 3(1):11, 2025. 3

[49] Xinyi Zhang, Hang Dong, Jinshan Pan, Chao Zhu, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Fei Wang. Learning to restore hazy video: A new real-world dataset and a new method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9239–9248, 2021. 3

[50] Shiyu Zhao, Lin Zhang, Shuaiyi Huang, Ying Shen, and Shengjie Zhao. Dehazing evaluation: Real-world benchmark datasets, criteria, and baselines. *IEEE Transactions on Image Processing*, 29:6947–6962, 2020. 3, 4