Addressing Data Scarcity in Depth-Based Human Action Recognition via Zero-Shot Depth Estimation

Rebeka Angyal, Pedro Hermosilla, Martin Kampel, Irene Ballester
TU Wien, Austria
e11933837@student.tuwien.ac.at

Abstract

Human Action Recognition (HAR) identifies human activities from visual or sensor data, with applications in healthcare and robotics. While RGB video is the standard input modality for HAR, depth data offers advantages such as robustness to lighting conditions and improved privacy. However, the lack of large, publicly available depth datasets limits model performance and generalization. We demonstrate that synthetic depth, derived from RGB videos using zero-shot depth estimation, can effectively supplement real depth data for training HAR models. Unlike prior work that either uses depth alongside RGB or evaluates solely on synthetic data, we show that augmenting limited real depth data (20-25%) with synthetic depth enables effective transfer to real sensor-captured depth in depth-only HAR. This approach recovers performance lost to data scarcity and, in some cases, even surpasses models trained on full real depth datasets by up to 5 percentage points with robust video architectures. Our strategy provides a scalable and practical solution, enabling the use of readily available RGB datasets for depth-based HAR in real-world applications.

1. Introduction

Human Action Recognition (HAR) classifies human activities from visual or sensor data and finds applications in diverse fields, including robotics, surveillance, and healthcare [22, 31, 58]. While RGB is the dominant modality, depth data provides distinct advantages for specific applications. Depth sensors capture geometric structure rather than texture, making depth-based methods more robust to lighting conditions or color variations that affect RGB [26]. Depth data also supports privacy-sensitive applications, such as assistive care [3, 4, 14, 18, 37], by abstracting visual details that could identify individuals [32].

Despite these advantages, progress in depth-based HAR remains constrained by the limited scale and diversity of available depth datasets. State-of-the-art RGB video models benefit from massive-scale pretraining—for instance, V-JEPA 2 uses 22 million video clips totaling over 1 million hours [2]—while the largest publicly available depth dataset, NTU RGB+D 120 [28], contains only 114K video samples (averaging a few seconds each) recorded in controlled settings. This orders-of-magnitude difference in training data constrains the performance and generalization of depth-based models.

A common strategy to mitigate data scarcity is the generation of synthetic training data [20, 48, 49]. Recent advances in monocular depth estimation provide a promising approach for generating synthetic depth data, as these methods can produce depth maps from RGB images and videos [13, 21, 55, 56]. Unlike supervised depth estimation, which relies on scarce paired RGB-depth data, zero-shot depth estimation models can be applied to any large, diverse RGB video collection. This enables large-scale RGB benchmarks like Kinetics-700 [9] to be transformed into synthetic depth datasets for HAR training.

Prior work has explored estimated depth for HAR in two ways: as a complementary modality to RGB-based HAR [38, 40, 59], and in synthetic depth-only approaches, where depth maps are estimated from RGB videos and evaluated only on RGB-based HAR benchmarks, without testing on real depth data [10]. However, to our knowledge, no existing study has systematically evaluated whether off-the-shelf synthetic depth improves depth-only HAR on real sensor data. A key challenge in this setting is the domain gap between synthetic and real depth, arising from differences in sensor noise characteristics, quantization, depth range limitations, and resolution.

In this work, we address depth data scarcity by leveraging zero-shot depth estimation to generate synthetic depth from RGB action datasets for training depth-only HAR models evaluated on real sensor-capture depth. We systematically evaluate the impact of synthetic depth on recognition accuracy and analyze the effects of domain shift between synthetic and real depth. Our results show that synthetic depth improves depth-based HAR in low-data set-

tings and, in some cases, even outperforms models trained on full real-depth datasets.

The main contributions of this paper are as follows:

- **Overcoming data scarcity in depth-only HAR via zero-shot synthetic depth.** We demonstrate that augmenting limited real depth data with synthetic depth from zero-shot models improves depth-only HAR on real sensor data, in some cases surpassing full real-data baselines.
- **Systematic evaluation of synthetic depth usage.** We evaluate three video understanding methods adapted for depth-based HAR, and different training strategies across four datasets to characterize when and how synthetic depth improves recognition and generalization.
- **Quantifying and mitigating domain shift.** We quantify the synthetic-to-real depth domain gap using similarity metrics and evaluate lightweight domain adaptation techniques to mitigate its impact on HAR accuracy.

2. Related Work

Human Action Recognition on depth data. While early depth-based HAR relied on handcrafted descriptors [33, 57], the field has shifted to deep learning to capture temporal dynamics via 3D CNNs [29], ConvLSTMs [41], and Transformers [53]. Estimated depth is employed as a complementary modality to enhance RGB-based recognition in Depth2Action [59], HAR-Depth [40], and DEAR [38]. Other approaches generate synthetic depth using custom pipelines to augment training data [44]. In contrast, our work investigates the use of off-the-shelf zero-shot depth estimation to train depth-only models for deployment on real sensor-captured depth data, providing scalability without the need for dataset-specific generation pipelines.

To address the scarcity of large-scale depth datasets, we transfer representations from RGB-pretrained video understanding models, enabling depth-based HAR models to leverage general motion knowledge learned from RGB videos. We focus on unimodal visual architectures for computational efficiency and leverage state-of-the-art self-supervised video models to extract robust spatiotemporal features. We employ VideoMAE [47] and its successor VideoMAE V2 [50], which learn motion by reconstructing masked spatiotemporal tubes, as well as V-JEPA 2 [2], which predicts latent representations in a joint-embedding space. These models provide the necessary spatiotemporal foundation for our depth-based HAR task.

Monocular Depth Estimation. Monocular depth estimation methods divide into generative approaches [19, 21, 55] that leverage learned priors and probabilistic modeling, and discriminative models that predict depth directly via end-to-end learning [13, 56]. Zero-shot depth estimation approaches are designed to generalize across diverse domains without requiring dataset-specific fine-tuning. Recent image-based zero-shot methods employ vision-transformer backbones [6, 7, 56], diffusion-based generative models [21], and multi-dataset training strategies [5, 17, 36]. Video approaches extend these architectures to maintain temporal consistency [19, 55].

Since our input for HAR is video sequences, we require temporally consistent depth estimation. We use Video Depth Anything [13], which adapts the Depth Anything V2 architecture for video and produces metric depth estimates that align with the physical scale of real sensor data.

Domain shift and adaptation. A fundamental challenge in using synthetic data is domain shift, where source and target domains follow different distributions [23, 35]. In this work, domain shift arises from distributional differences between synthetic depth maps from zero-shot monocular models and real sensor measurements. Domain adaptation methods address this through learning domain-invariant representations, aligning distributions, or adapting models to the target domain. Classical approaches include transfer component analysis [34] and subspace alignment [16], while deep learning methods use discrepancy minimization [25, 30, 46] and adversarial training [8].

While deep domain adaptation methods have shown promise, they require architectural modifications or computationally expensive auxiliary training objectives that limit their practical applicability. In this work, we investigate whether lightweight input-level adaptation strategies are sufficient to bridge the synthetic-to-real gap. We align the value ranges and distributions of synthetic and real depth through thresholding and normalization. This input-level approach establishes a baseline for reducing domain shift without architectural changes or additional training overhead, making it compatible with any pretrained model.

3. Methodology

Our framework addresses the scarcity of depth data in HAR through a three-stage pipeline (illustrated in Figure 1): (1) generating synthetic depth from RGB videos, (2) adapting pretrained video understanding models to depth inputs, including standardized preprocessing of both synthetic and real data, and (3) applying lightweight domain adaptation to bridge the synthetic-to-real gap.

3.1. Synthetic Depth Generation

We generate synthetic depth maps from RGB videos using Video Depth Anything [13], a video-based zero-shot monocular depth estimator (ViT-L, 381.8M parameters). Unlike image-based approaches that process frames independently, Video Depth Anything operates on temporal windows, leveraging motion cues and temporal consistency. The model outputs metric depth, providing per-pixel distances in millimeters and eliminating the need for dataset-

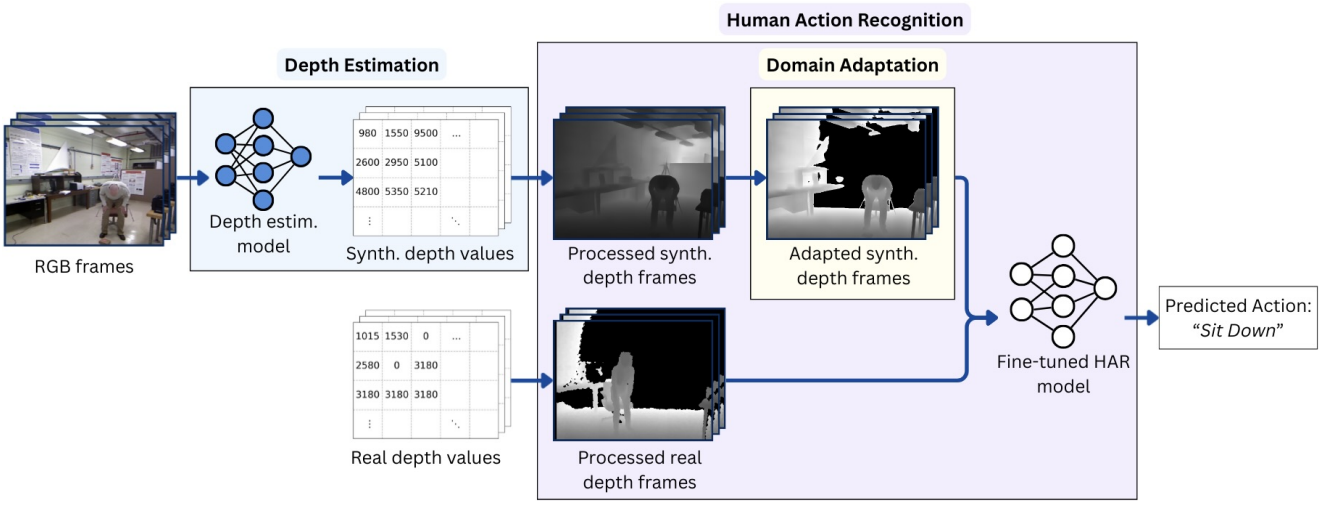


Figure 1. **Proposed framework.** (1) Depth Estimation: synthetic depth maps are generated from RGB video. (2) HAR Model: synthetic and real depth are converted into a standardized format, and a pretrained video understanding model is fine-tuned on the depth data to predict the action class. (3) Domain Adaptation (optional): the domain gap between synthetic and real depth is reduced using thresholding. (Sample images used in the figure are from the UTKinect dataset [54].)

specific scale calibration. Video Depth Anything extends the image-based Depth Anything V2 [56] with a spatial-temporal head and temporal gradient matching loss to enforce consistency across frames [13], producing smoother sequences than frame-by-frame processing.

3.2. Human Action Recognition

Prior to training, real and synthetic depth data are passed through a unified preprocessing pipeline. Raw depth values are converted into standardized frames, where values are clipped to the 0.5th and 99.5th percentiles and linearly normalized to [0, 255]. Single-channel maps are replicated across three channels, following [12, 21, 45], to enable direct use of pretrained weights without architectural modifications. We employ three pretrained video understanding models: VideoMAE [47], VideoMAE V2 [50], and V-JEPA 2 [2]. VideoMAE learns representations by reconstructing masked spatiotemporal tubes in video sequences. VideoMAE V2 improves upon this by introducing a dual masking strategy, masking both the encoder and the decoder. V-JEPA 2 predicts latent features in a joint-embedding space. All three use vision transformer backbones pretrained via self-supervised learning on large-scale RGB video datasets, capturing general motion dynamics.

We adapt these models to depth inputs via end-to-end fine-tuning, following recent work showing that RGB-pretrained vision models can effectively process depth data through fine-tuning [38]. The pretrained encoder is initialized with RGB weights, and the final classification layer is replaced with a randomly initialized head corresponding to the number of target action classes. Two fine-tuning strategies are compared: *mixed training*, where models are trained jointly on combined synthetic and real data, and *sequential fine-tuning*, where models are first trained on synthetic depth and then fine-tuned on real data. Their effectiveness is analyzed in an ablation study reported in the supplementary material (Section 5).

Models are trained using cross-entropy loss with linear learning rate scheduling and standard spatial augmentation (random scaling and cropping). Hyperparameters are tuned per dataset scale; full details are in the supplementary material (Section 1.6).

3.3. Domain Adaptation

To mitigate the domain shift between synthetic and real depth data, we employ lightweight input-level adaptation techniques that are computationally efficient and compatible with the pretrained models, without modifying their architectures. These methods establish a baseline to evaluate how distributional discrepancies impact HAR performance.

Adaptation methods. Two thresholding strategies are applied to synthetic depth maps: *clipping*, where values outside a given range are set to the nearest bound, and *cutting*, where out-of-range values are set to zero. These strategies address the value range mismatch between synthetic depth estimators, which produce metric depth across wide ranges, and hardware-constrained real depth sensors. After thresholding, depth values are normalized to the specified range. Three range configurations are considered:

- **Align to real depth:** Uses the 0.5th and 99.5th percentiles of real depth frames to match the distribution.
- **Align to ratio of real depth:** Uses synthetic boundaries

proportionally according to the real-data range.

- **Align to Kinect sensor range:** Uses standard Kinect depth range of 500–4500 mm [15], independent of dataset-specific statistics.

These configurations assess whether alignment should be data-driven (percentiles), proportionally scaled (ratio), or based on known sensor specifications (Kinect range).

Measuring domain alignment. To evaluate the alignment between synthetic and real depth, we compute quantitative metrics comparing the two types of frames, investigating whether improvements in HAR performance correlate with quantified reductions in domain shift.

We use pixel-wise error metrics standard in monocular depth estimation [7, 13, 55]: Absolute Relative Difference (AbsRel) and Threshold Accuracy (δ_τ), where N is the number of pixels and d_i^{synth}, d_i^{real} represent synthetic and real depth values:

$$\text{AbsRel} = \frac{1}{N}\sum_{i=1}^{N}\frac{|d_i^{\text{real}} - d_i^{\text{synth}}|}{d_i^{\text{real}}}, \tag{1}$$

$$\delta_\tau = \frac{1}{N}\sum_{i=1}^{N}\left(\max\left(\frac{d_i^{\text{synth}}}{d_i^{\text{real}}}, \frac{d_i^{\text{real}}}{d_i^{\text{synth}}}\right) < \tau\right). \tag{2}$$

We also consider absolute and perceptual similarity measures. RMSE provides a pixel-wise estimate of value magnitude errors, while SSIM [51] evaluates structural alignment and perceptual similarity. These metrics complement the relative error measures by focusing on absolute distance accuracy and visual structure, respectively:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(d_i^{\text{synth}} - d_i^{\text{real}})^2}, \tag{3}$$

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \tag{4}$$

We calculate the 1-Wasserstein distance (Earth Mover's Distance) [39] to compare synthetic and real depth distributions [1]:

$$W_1(p, q) = \int_{-\infty}^{\infty} |F_p(x) - F_q(x)|dx, \tag{5}$$

where F_p and F_q denote the cumulative distribution functions of the depth distributions p and q. Metric aggregation details are provided in the supplementary material (Section 1.3).

4. Evaluation

We evaluate the proposed framework on four depth-based HAR datasets using standard classification metrics, specifically accuracy and F1 score. The evaluation is structured to answer three core questions: (i) how effective synthetic depth is when used within the same dataset, (ii) whether lightweight input-level adaptation can reduce the synthetic-to-real domain gap, and (iii) whether synthetic depth improves cross-dataset generalization.

4.1. Experimental Configurations

To investigate these questions, we design controlled experimental configurations grouped into three categories: within-dataset evaluation, domain adaptation, and cross-dataset generalization. Across all configurations, a fixed test set of real sensor-acquired depth is used per dataset to ensure comparability. Unless stated otherwise, training refers to fine-tuning pretrained video understanding models rather than training from scratch. Table 1 provides an overview of all configurations.

Within-dataset. This experiment evaluates the effectiveness of synthetic depth when generated and tested on the same dataset. This setting establishes reference performance levels, allowing for quantification of the domain gap and assessment of whether synthetic data can compensate for limited real depth. We use $\alpha \in [0, 1]$ to denote the fraction of real depth in training.

We first establish a *full real-depth reference* configuration ($\alpha = 1$), where models are trained and tested on the full real depth dataset. This represents the ideal scenario with unrestricted access to sensor data. The low-data *baseline* configuration ($\alpha \approx 0.20$–0.25, depending on dataset) is defined by training on a small subset of real depth, reflecting realistic data-scarce conditions. Furthermore, to measure the synthetic-to-real *domain gap* configuration ($\alpha = 0$), the models are trained only on synthetic depth and evaluated on real depth. Finally, we evaluate the *mixed training* configuration, where a small fraction of real depth ($\alpha \approx 0.20$–0.25) is combined with synthetic depth ($1 - \alpha$). In this setting, synthetic data serves as augmentation. This configuration tests whether synthetic depth can improve performance beyond the low-data real baseline.

Domain adaptation. The second experiment evaluates whether lightweight input-level adaptation can reduce the domain gap between synthetic and real depth. We focus on thresholding-based strategies, which adjust the value distribution of synthetic depth without modifying model architectures or training objectives.

We assess these strategies in two configurations. First, we evaluate the *domain gap with thresholding* configuration by training models on thresholded synthetic depth and eval-

Table 1. **Experimental configurations.** We define $\alpha \in [0, 1]$ as the fraction of the real depth dataset used for training. In mixed configurations, synthetic data corresponds to the complementary portion $(1 - \alpha)$. Typical values for α in low-data configurations are $\alpha \approx 0.20$–0.25, depending on the dataset.

Exp.	ID	Configuration	Training Data	Purpose
Within-dataset	1	Full real-depth reference	Real depth ($\alpha = 1$)	Establish target performance.
	2	Real baseline	Small subset of real depth ($\alpha < 1$)	Define baseline under data scarcity.
	3	Domain gap	Synth. depth ($\alpha = 0$)	Quantify the domain gap via direct transfer.
	4	Mixed training	Small subset of real (α) + large synth. ($1 - \alpha$).	Test if synthetic data as augmentation improves baseline performance.
Domain adaptation	5	Domain gap with thresholding	Thresholded synth. ($\alpha = 0$)	Measure if thresholding improves direct transfer performance.
	6	Mixed training with thresholding	Subset of real (α) + thresholded synth. ($1 - \alpha$).	Evaluate if thresholding improves performance in a mixed-data setting.
Cross-dataset	7	Generalization	Source synth. ($1-\alpha$) + target subset of real (α)	Test cross-dataset transferability.
	8	Generalization with thresholding	Source thresholded synth. ($1 - \alpha$) + target subset of real (α)	Test if thresholding improves cross-dataset transferability.

uating on real data to assess improvements in direct transfer. Second, thresholded synthetic depth is combined with a small subset of real depth in a mixed training setting, allowing us to examine whether value alignment further enhances the effectiveness of synthetic data as augmentation in the *mixed training with thresholding* configuration.

Cross-dataset generalization. We investigate whether synthetic depth generated from a source dataset can improve performance on a target dataset with limited real depth, reflecting practical scenarios where real depth data is scarce.

We evaluate the *cross-dataset generalization* configuration in a mixed training setting using synthetic depth from the source dataset and a small subset of real depth from the target dataset. To further assess the robustness of this transfer, we additionally apply thresholding-based adaptation to the synthetic data, examining whether value alignment improves generalization across datasets in the *cross-dataset with thresholding* configuration.

4.2. Datasets

We evaluate our framework on four RGB-D public datasets: the small-scale UTD-MHAD [11] and UTKinect [54], and the large-scale NTU RGB+D [42] and PKU-MMD [27]. While our method requires only RGB to generate synthetic depth, evaluation requires RGB-D datasets where real depth serves as ground truth. This enables us to establish a real-depth reference (*config. 1*), directly quantify the synthetic-real domain gap (*config. 3*), and evaluate whether synthetic data can compensate for limited real depth data (*config. 4*).

Small-scale datasets are filtered to seven overlapping action classes, and large-scale datasets to 41 classes, enabling consistent evaluation across all settings including cross-dataset experiments. We use frontal-views from NTU RGB+D and middle camera views from PKU-MMD for consistent viewpoints. For all datasets, we follow official cross-subject evaluation protocols with subject-level independence between training and testing. Validation sets are reserved from the training split for model selection (details are in supplementary material Section 1.1).

To evaluate performance under data scarcity, we train on stratified subsets of the real data. Results are reported as mean and standard deviation over three random splits for subset-based configurations (*e.g.*, *real baseline and mixed training*). Configurations trained on the full dataset, such as the *full real-depth reference* or *domain gap* (full synthetic depth dataset), do not involve random sampling and are evaluated with a single run. In *mixed training* experiments, we implement a strict exclusion policy: any video selected for the real subset is excluded from the synthetic pool, ensuring the model never sees the same action instance in both modalities during training. Implementation details are provided in the supplementary material (Section 1).

5. Experiments

5.1. Within-Dataset Performance

Table 2 summarizes the results of integrating synthetic depth into the training pipeline. Detailed results for individual synthetic-to-real training configurations and F1 scores are reported in the supplementary material (Section 2).

Overcoming scarcity. As expected, training on a small subset of real data (*real baseline*) results in a performance drop compared to the *full real-depth reference*. Augmenting this low-data baseline with synthetic depth (*mixed train-*

Table 2. **Within-dataset accuracy.** Comparison of VideoMAE, VideoMAE V2, and V-JEPA 2 across datasets. *Full real-depth reference*: trained on all real data ($\alpha = 1$). *Real baseline*: small subset of real data ($\alpha \approx 0.20$–0.25). *Domain gap*: trained on synthetic only ($\alpha = 0$), tested on real. *Mixed training*: subset of real combined with synthetic ($\alpha + (1 - \alpha)$). Mixed training recovers performance lost to data scarcity, with V-JEPA 2 exceeding the full reference on three datasets.

Model	Configuration	UTKinect	UTD-MHAD	PKU-MMD	NTU RGB+D
VideoMAE	1. Full real-depth reference ($\alpha = 1$)	0.90	0.93	0.48	0.70
	2. Real baseline ($\alpha < 1$)	0.79 ± 0.08	0.74 ± 0.10	0.30 ± 0.01	0.58 ± 0.01
	3. Domain gap ($\alpha = 0$)	0.59	0.31	0.11	0.29
	4. Mixed training ($\alpha + (1 - \alpha)$)	0.87 ± 0.02	0.89 ± 0.03	0.43 ± 0.02	0.69 ± 0.02
VideoMAE V2	1. Full real-depth reference ($\alpha = 1$)	0.93	0.99	0.57	0.75
	2. Real baseline ($\alpha < 1$)	0.89 ± 0.02	0.92 ± 0.07	0.44 ± 0.02	0.70 ± 0.01
	3. Domain gap ($\alpha = 0$)	0.74	0.74	0.50	0.63
	4. Mixed training ($\alpha + (1 - \alpha)$)	0.94 ± 0.02	0.96 ± 0.01	0.57 ± 0.02	0.74 ± 0.04
V-JEPA 2	1. Full real-depth reference ($\alpha = 1$)	0.96	0.95	0.62	0.81
	2. Real baseline ($\alpha < 1$)	0.78 ± 0.08	0.87 ± 0.12	0.45 ± 0.03	0.80 ± 0.05
	3. Domain gap ($\alpha = 0$)	0.91	0.89	0.59	0.78
	4. Mixed training ($\alpha + (1 - \alpha)$)	0.90 ± 0.04	0.96 ± 0.07	0.67 ± 0.03	0.82 ± 0.09

ing) consistently recovers performance. For VideoMAE, accuracy improvements range from +0.08 to +0.15 across datasets. VideoMAE V2 shows similar patterns, with mixed training matching or exceeding the full reference (*e.g.*, 0.94 vs 0.93 on UTKinect). V-JEPA 2 is particularly effective: mixed training outperforms the full real-depth reference on UTD-MHAD, PKU-MMD, and NTU RGB+D. Real depth data exhibits noise, which is qualitatively visible in Figures 2a and 2b. Mixing real and synthetic depth combines realistic sensor artifacts with geometrically cleaner representations of the same scenes. For V-JEPA 2 specifically, the ability to exceed the full real-depth baseline on three datasets suggests that these modalities provide complementary training signals.

Backbone robustness. The three architectures show different resilience to the synthetic-to-real domain gap. Training only on synthetic data (*domain gap*) causes sharp drops for VideoMAE (*e.g.*, 0.93 to 0.31 on UTD-MHAD), moderate drops (up to 0.25) for VideoMAE V2, and minimal drops (at most 0.06) for V-JEPA 2, which remains near the full real-depth reference. This robustness hierarchy (V-JEPA 2 > VideoMAE V2 > VideoMAE) aligns with V-JEPA 2's stronger motion understanding [2], making it the most reliable backbone in scenarios where real depth data is limited.

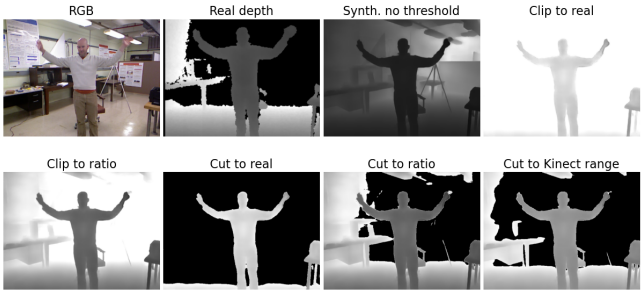
5.2. Domain Adaptation

We investigate whether lightweight input-level adaptation (thresholding) reduces the domain gap by analyzing similarity metrics (Table 3) and HAR accuracy (Table 4). Results for direct transfer with thresholding (config. 5) are provided in the supplementary material (Section 3).

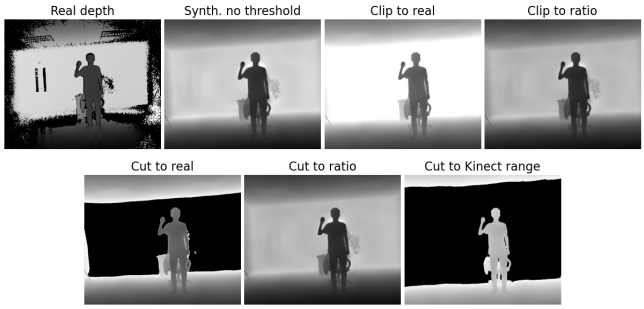
Small-scale datasets. UTKinect and UTD-MHAD feature

Table 3. **Domain alignment metrics.** Quantitative comparison of all thresholding strategies across datasets. For small-scale datasets (UTKinect, UTD-MHAD), "Cutting" strategies substantially improve alignment. For large-scale datasets (PKU-MMD, NTU RGB+D), simple "Clipping" or the Baseline often yield better similarity due to the sensor noise in the real data.

Strategy	RMSE ↓	AbsRel ↓	Wass. ↓	$\delta_{0.25}$ ↑	SSIM ↑
UTKinect					
Baseline	122.47	59.15	45.45	1.34	0.15
Clip to Real	196.16	136.17	172.73	14.10	0.23
Clip to Ratio	180.18	127.98	129.84	35.63	0.26
Cut to Real	118.17	**3.07**	45.24	52.99	0.49
Cut to Ratio	**89.11**	30.28	19.73	**72.11**	0.57
Cut to Kinect	92.18	17.75	**12.07**	67.32	**0.59**
UTD-MHAD					
Baseline	173.68	157.02	154.69	0.88	0.02
Clip to Real	241.99	229.77	238.56	0.11	0.02
Clip to Ratio	230.60	213.91	215.73	3.35	0.02
Cut to Real	62.29	**15.48**	**16.84**	**84.36**	**0.83**
Cut to Ratio	**57.42**	19.19	17.34	79.51	0.75
Cut to Kinect	141.61	88.79	94.21	51.82	0.45
PKU-MMD					
Baseline	77.16	36.28	38.41	**43.63**	**0.44**
Clip to Real	108.42	55.84	76.79	37.27	0.40
Clip to Ratio	**73.75**	32.34	41.99	39.23	**0.44**
Cut to Real	150.98	47.66	**25.71**	15.37	0.18
Cut to Ratio	73.76	**32.33**	41.99	39.23	**0.44**
Cut to Kinect	173.47	34.81	34.40	18.59	0.11
NTU RGB+D					
Baseline	61.07	26.35	26.73	**50.92**	**0.45**
Clip to Real	83.10	36.75	60.41	23.10	0.43
Clip to Ratio	**53.28**	20.28	**21.52**	20.93	0.44
Cut to Real	85.16	34.53	52.45	22.62	0.42
Cut to Ratio	53.29	**20.27**	**21.52**	20.93	0.44
Cut to Kinect	141.51	31.09	51.28	12.85	0.16



(a) UTKinect (small-scale): Applying the "Cutting" strategy (i.e., setting values exceeding a certain threshold to zero) to the synthetic background aligns it with the clean, zero-value background of the real sensor data.



(b) PKU-MMD (large-scale): Real depth contains sensor noise, and applying the "Cutting" strategy to the synthetic background increases the domain gap.

Figure 2. **Visual impact of thresholding strategies.** Small-scale datasets (a) exhibit range-limited depth (black background). Large-scale datasets (b) show fragmented sensor noise. Thresholding succeeds only when sensor characteristics are simple enough to replicate.

Table 4. **HAR accuracy with input-level adaptation.** Comparison of the baseline (no threshold) vs. the best thresholding strategy for each dataset (Mixed Training). Best strategy for small-scale datasets: *Cut to Ratio*; for large-scale datasets: *Clip to Ratio*.

Model	TH	UTKinect	UTD-MHAD	PKU-MMD	NTU RGB+D
VideoMAE	None	0.87 ± 0.02	0.89 ± 0.03	$\mathbf{0.43 \pm 0.02}$	0.60 ± 0.01
	Best	$\mathbf{0.89 \pm 0.03}$	$\mathbf{0.95 \pm 0.01}$	0.42 ± 0.01	0.60 ± 0.03
VideoMAE V2	None	$\mathbf{0.94 \pm 0.02}$	0.96 ± 0.01	$\mathbf{0.57 \pm 0.02}$	0.68 ± 0.02
	Best	0.92 ± 0.02	0.96 ± 0.04	0.55 ± 0.01	$\mathbf{0.70 \pm 0.02}$
V-JEPA 2	None	0.90 ± 0.04	0.96 ± 0.07	$\mathbf{0.67 \pm 0.03}$	0.77 ± 0.09
	Best	$\mathbf{0.93 \pm 0.04}$	$\mathbf{0.99 \pm 0.01}$	0.66 ± 0.01	$\mathbf{0.79 \pm 0.05}$

Table 5. **Missing depth region structure by sensor type.** Kinect v2 exhibits far more fragmented patterns than v1: more regions per frame (disconnected components of missing pixels), with smaller largest region fractions. Statistics are computed over all frames.

Dataset	Sensor	Regions/Frame	Largest Frac.
UTKinect	Kinect v1	28.2 ± 5.8	0.852 ± 0.128
UTD-MHAD	Kinect v1	2.8 ± 1.4	0.999 ± 0.001
PKU-MMD	Kinect v2	926.1 ± 81.6	0.635 ± 0.216
NTU RGB+D	Kinect v2	1620.4 ± 370.8	0.835 ± 0.100

zero-value backgrounds due to the sensor's limited range. Synthetic data, however, contains depth values for the entire scene. Thresholding strategies that "cut" the synthetic background (*e.g.*, *Cut to Real/Kinect Range*) improve structural alignment, reducing the similarity metrics, *e.g.*, the Wasserstein distance on UTKinect decreases from 45.45 to 12.07 (Table 3). This improvement translates directly to HAR performance for VideoMAE and V-JEPA 2: on UTD-MHAD, VideoMAE improves from 0.89 to 0.95 and V-JEPA 2 from 0.96 to 0.99 with thresholding, and on UTKinect, V-JEPA 2 accuracy increases from 0.90 to 0.93. VideoMAE V2, however, shows no consistent benefit: thresholding yields a marginal drop on UTKinect (0.94 to 0.92) and no change on UTD-MHAD (0.96), suggesting that its stronger pretrained representations are already robust to the range mismatch. Figure 2a confirms that cutting thresholding replicates the sensor's range limitation, producing zero-value (or null) backgrounds similar to real depth data.

Large-scale datasets. In contrast, PKU-MMD and NTU RGB+D contain substantial sensor noise and clutter rather than clean backgrounds (Figure 2b). Applying "cut" strategies to the clean synthetic data artificially removes contextual information, increasing the domain gap (*e.g.*, RMSE on NTU increases from 61.07 to 85.16). As a result, thresholding yields no consistent improvement across models (Table 4): while V-JEPA 2 and VideoMAE V2 show marginal gains on NTU RGB+D, VideoMAE does not benefit, and no model improves on PKU-MMD. This suggests that input-level adaptation can partially correct range mismatches in some settings, but cannot reliably replicate complex sensor noise. These observations indicate that such domain gaps likely require more advanced domain adaptation methods and are rooted in differences between sensor generations.

Missing depth region structure. To understand why cutting thresholding succeeds on small-scale but fails on large-scale datasets, we examine the underlying depth data characteristics. Since cutting creates artificial missing regions by zeroing distant values, its success depends on matching real sensor patterns. For each frame, we extract connected components of missing pixels (pixel value $= 0$, 8-connectivity) and measure (i) the mean number of distinct missing regions per frame and (ii) the fraction of missing pixels belonging to the single largest region, where a high ratio indicates one coherent hole, a low ratio indicates scattered noise. Table 5 shows that small-scale datasets have few, coherent missing regions (3-28/frame), whereas large-scale datasets show highly fragmented patterns (926-1620/frame). This explains why cutting succeeds only on the former, as it can replicate a large missing region in the background hole, but not hundreds of scattered fragments.

This difference aligns with underlying sensor technol-

Table 6. **Cross-dataset generalization.** Detailed accuracy results across all source-target dataset pairs and configurations. *Target Real Baseline:* limited target real data ($\alpha \approx 0.20$–0.25). *Source Synth. Only:* source synthetic data only ($\alpha = 0$). *Cross-Dataset Mixed:* source synthetic and α real target data. *Target Reference:* full target real depth ($\alpha = 1$). Mixing source synthetic with limited target real consistently improves over target real alone and closes the gap toward the full target reference, demonstrating cross-dataset transferability.

Source → Target	Model	Target Real Baseline	Source Synth. Only	Cross-Dataset Mixed	Target Reference
UTD-MHAD → UTKinect	VideoMAE	0.79 ± 0.08	0.40	0.84 ± 0.04	0.90
	VideoMAE V2	0.89 ± 0.02	0.63	0.89 ± 0.01	0.93
	V-JEPA 2	0.78 ± 0.08	0.41	0.88 ± 0.01	0.96
UTKinect → UTD-MHAD	VideoMAE	0.74 ± 0.10	0.25	0.85 ± 0.02	0.93
	VideoMAE V2	0.92 ± 0.07	0.32	0.91 ± 0.03	0.99
	V-JEPA 2	0.87 ± 0.12	0.49	0.87 ± 0.10	0.95
NTU RGB+D → PKU-MMD	VideoMAE	0.30 ± 0.01	0.22	0.36 ± 0.02	0.48
	VideoMAE V2	0.44 ± 0.02	0.29	0.45 ± 0.02	0.57
	V-JEPA 2	0.45 ± 0.03	0.43	0.51 ± 0.04	0.62
PKU-MMD → NTU RGB+D	VideoMAE	0.58 ± 0.01	0.07	0.61 ± 0.01	0.70
	VideoMAE V2	0.70 ± 0.01	0.35	0.72 ± 0.02	0.75
	V-JEPA 2	0.80 ± 0.05	0.39	0.79 ± 0.03	0.81

ogy. Small-scale datasets use Kinect v1 (structured light), where missing pixels arise from range limitations [24, 43], creating coherent background holes that cutting replicates by removing far values. Large-scale datasets use Kinect v2 (Time-of-Flight), where missing pixels stem from geometry-dependent factors—multipath interference and edge artifacts [52]—producing fragmented patterns that cutting cannot replicate.

Correlation between metrics and performance. The relationship between domain alignment metrics and HAR accuracy is dataset-dependent. On small-scale datasets (UTKinect, UTD-MHAD), improved similarity metrics align with HAR improvements, indicating they effectively capture the domain gap when sensor characteristics are simple. On large-scale datasets (PKU-MMD, NTU RGB+D), this relationship breaks down: models perform best with unmodified synthetic data despite worse metric scores after thresholding. This indicates that standard depth estimation metrics do not reliably predict HAR performance in the presence of complex sensor noise.

5.3. Cross-Dataset Generalization

We assess cross-dataset generalization by training on synthetic data from a *source* dataset combined with a small subset of real data (α) from a *target* dataset. Table 6 shows that this strategy outperforms the target-only baseline in most scenarios. Across most cross-dataset settings, augmenting a small subset of real target data with a large volume of synthetic source data is an effective strategy, confirming that the spatiotemporal patterns learned from the synthetic source data are transferable to the target domain.

For small-scale datasets (UTKinect and UTD-MHAD), cross-dataset mixed training consistently improves performance for VideoMAE and V-JEPA 2. VideoMAE V2, however, shows a more modest effect: cross-dataset mixing yields only marginal gains over the baseline (0.89 vs 0.89 on UTKinect, 0.91 vs 0.92 on UTD-MHAD), suggesting that its stronger pretrained representations already extract sufficient structure from the limited real target data alone.

For large-scale datasets, cross-dataset mixed training consistently outperforms the real baseline across all three models. V-JEPA 2 best bridges the synthetic-to-real gap in the source-only setting: on NTU RGB+D → PKU-MMD, its synthetic-only accuracy (0.43) nearly matches the real baseline (0.45), whereas VideoMAE and VideoMAE V2 fall further behind (0.22 and 0.29, respectively). Thresholding strategies in cross-dataset training (config. 8) show inconsistent effects on small-scale datasets; detailed results are in the supplementary material (Section 4).

6. Conclusion

We addressed depth data scarcity by leveraging zero-shot depth estimation to generate synthetic depth from RGB videos for augmenting limited real depth training data. Integrating synthetic depth consistently improved performance in data-scarce settings, with V-JEPA 2 demonstrating the strongest resilience to the synthetic-to-real domain gap. Lightweight input-level adaptation succeeded for Kinect v1 datasets (UTKinect, UTD-MHAD) with coherent missing depth regions, but failed for Kinect v2 datasets (PKU-MMD, NTU RGB+D) with fragmented noise, indicating that sensor-specific characteristics require tailored strategies. Cross-dataset experiments confirmed that synthetic depth transfers effectively across domains, enabling knowledge transfer from abundant RGB sources to sparse depth domains. Overall, our framework offers a scalable solution for depth-based HAR and provides a foundation for future exploration in broader depth-based recognition tasks.

Acknowledgments
This project was partially supported by the STeRN project under the FFG-KIRAS program (grant agreement number 999926276).

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 214–223. PMLR, 2017. 4

[2] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. 1, 2, 3, 6

[3] Irene Ballester, Tamar Mujirishvili, and Martin Kampel. Rita: a privacy-aware toileting assistance designed for people with dementia. In *International Conference on Pervasive Computing Technologies for Healthcare*, pages 318–330. Springer, 2021. 1

[4] Irene Ballester, Markus Gall, Thomas Münzer, and Martin Kampel. Vision-based toilet assistant for people with dementia in real-life situations. In *2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, pages 141–147. IEEE, 2024. 1

[5] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 2

[6] Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3. 1–a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023. 2

[7] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *International Conference on Learning Representations*, 2025. 2, 4

[8] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, page 343–351, Red Hook, NY, USA, 2016. Curran Associates Inc. 2

[9] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 1

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

[11] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 168–172, 2015. 5

[12] Qian Chen, Ze Liu, Yi Zhang, Keren Fu, Qijun Zhao, and Hongwei Du. Rgb-d salient object detection via 3d convolutional neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1063–1071, 2021. 3

[13] Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22831–22840, 2025. 1, 2, 3, 4

[14] Edward Chou, Matthew Tan, Cherry Zou, Michelle Guo, Albert Haque, Arnold Milstein, and Li Fei-Fei. Privacy-preserving action recognition for smart hospitals using low-resolution depth images. *arXiv preprint arXiv:1811.09950*, 2018. 1

[15] Moustafa Fayad, Mohamed-Yacine Hachani, Kamal Ghoumid, Ahmed Mostefaoui, Samir Chouali, Fabien Picaud, Guillaume Herlem, Isabelle Lajoie, and Réda Yahiaoui. Fall detection approaches for monitoring elderly healthcare using kinect technology: A survey. *Applied Sciences*, 13(18), 2023. 4

[16] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2960–2967, 2013. 2

[17] Vitor Guizilini, Igor Vasiljevic, Dian Chen, Rareș Ambruș, and Adrien Gaidon. Towards zero-shot scale-aware monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9233–9243, 2023. 2

[18] Albert Haque, Michelle Guo, Alexandre Alahi, Serena Yeung, Zelun Luo, Alisha Rege, Jeffrey Jopling, Lance Downing, William Beninati, Amit Singh, et al. Towards vision-based smart hospitals: a system for tracking and monitoring hand hygiene compliance. In *Machine Learning for Healthcare Conference*, pages 75–87. PMLR, 2017. 1

[19] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2005–2015, 2025. 2

[20] Nikita Jaipuria, Xianling Zhang, Rohan Bhasin, Mayar Arafa, Punarjay Chakravarty, Shubham Shrivastava, Sagar Manglani, and Vidya N. Murali. Deflating dataset bias using synthetic data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020. 1

[21] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth

estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9492–9502, 2024. 1, 2, 3

[22] Muhammad Attique Khan, Kashif Javed, Sajid Ali Khan, Tanzila Saba, Usman Habib, Junaid Ali Khan, and Aaqif Afzaal Abbasi. Human action recognition using fusion of multiview and deep features: an application to video surveillance. *Multimedia Tools and Applications*, 83(5): 14885–14911, 2024. 1

[23] Wouter M Kouw and Marco Loog. An introduction to domain adaptation and transfer learning. *arXiv preprint arXiv:1812.11806*, 2018. 2

[24] Gregorij Kurillo, Evan Hemingway, Mu-Lin Cheng, and Louis Cheng. Evaluating the accuracy of the azure kinect and kinect v2. *Sensors*, 22(7):2469, 2022. 8

[25] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10277–10287, Los Alamitos, CA, USA, 2019. IEEE Computer Society. 2

[26] Chang Li, Qian Huang, Xing Li, and Qianhan Wu. Human action recognition based on multi-scale feature maps from depth video sequences. *Multimedia Tools and Applications*, 80(21):32111–32130, 2021. 1

[27] Chunhui Liu, Yueyu Hu, Yanghao Li, Sijie Song, and Jiaying Liu. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. *arXiv preprint arXiv:1703.07475*, 2017. 5

[28] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(10):2684–2701, 2020. 1

[29] Zhi Liu, Chenyang Zhang, and Yingli Tian. 3d-based deep convolutional neural network for action recognition with depth sequences. *Image and Vision Computing*, 55:93–100, 2016. 2

[30] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I. Jordan. Deep transfer learning with joint adaptation networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, page 2208–2217. JMLR.org, 2017. 2

[31] Guilherme Maeda, Marco Ewerton, Gerhard Neumann, Rudolf Lioutikov, and Jan Peters. Phase estimation for fast action recognition and trajectory generation in human-robot collaboration. *Int. J. Rob. Res.*, 36(13–14):1579–1594, 2017. 1

[32] Wiktor Mucha and Martin Kampel. Addressing privacy concerns in depth sensors. In *Computers Helping People with Special Needs: 18th International Conference, ICCHP-AAATE 2022, Lecco, Italy, July 11–15, 2022, Proceedings, Part II*, page 526–533, Berlin, Heidelberg, 2022. Springer-Verlag. 1

[33] Omar Oreifej and Zicheng Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 716–723, Los Alamitos, CA, USA, 2013. IEEE Computer Society. 2

[34] Sinno Jialin Pan, Ivor W. Tsang, James T. Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011. 2

[35] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE Signal Processing Magazine*, 32(3):53–69, 2015. 2

[36] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10106–10116, 2024. 2

[37] Christopher Pramerdorfer, Rainer Planinc, and Martin Kampel. Effective deep-learning-based depth data analysis on low-power hardware for supporting elderly care. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020. 1

[38] Sadegh Rahmaniboldaji, Filip Rybansky, Quoc Vuong, Frank Guerin, and Andrew Gilbert. Dear: Depth-enhanced action recognition. *arXiv preprint arXiv:2408.15679*, 2024. 1, 2, 3

[39] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000. 4

[40] Suraj Prakash Sahoo, Samit Ari, Kamalakanta Mahapatra, and Saraju P. Mohanty. Har-depth: A novel framework for human action recognition using sequential learning and depth estimated history images. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):813–825, 2021. 1, 2

[41] Adrián Sánchez-Caballero, David Fuentes-Jiménez, and Cristina Losada-Gutiérrez. Real-time human action recognition using raw depth video-based recurrent neural networks. *Multimedia Tools and Applications*, 82(11):16213–16235, 2023. 2

[42] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1010–1019, 2016. 5

[43] Wanbin Song, Anh Vu Le, Seokmin Yun, Seung-Won Jung, and Chee Sun Won. Depth completion for kinect v2 sensor. *Multimedia Tools and Applications*, 76(3):4357–4380, 2017. 8

[44] Christian Stippel, Thomas Heitzinger, Rafael Sterzinger, and Martin Kampel. Closing the gap in human behavior analysis: A pipeline for synthesizing trimodal data. In *2024 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, pages 793–798, 2024. 2

[45] Yiming Su, Haoran Gao, Mengyin Wang, and Fasheng Wang. Asymmetric cross-modality interaction network for

rgb-d salient object detection. *Expert Systems with Applications*, 275:127004, 2025. 3

[46] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision – ECCV 2016 Workshops*, pages 443–450, Cham, 2016. Springer International Publishing. 2

[47] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, pages 10078–10093. Curran Associates, Inc., 2022. 2, 3

[48] Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018. 1

[49] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[50] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14549–14560, 2023. 2, 3

[51] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 4

[52] Oliver Wasenmüller and Didier Stricker. Comparison of kinect v1 and v2 depth images in terms of accuracy and precision. In *Asian conference on computer vision*, pages 34–45. Springer, 2016. 8

[53] Hanbo Wu, Xin Ma, and Yibin Li. Transformer-based multiview spatiotemporal feature interactive fusion for human action recognition in depth videos. *Signal Processing: Image Communication*, 131:117244, 2025. 2

[54] Lu Xia, Chia-Chih Chen, and Jake K Aggarwal. View invariant human action recognition using histograms of 3d joints. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 20–27. IEEE, 2012. 3, 5

[55] Honghui Yang, Di Huang, Wei Yin, Chunhua Shen, Haifeng Liu, Xiaofei He, Binbin Lin, Wanli Ouyang, and Tong He. Depth any video with scalable synthetic data. *arXiv preprint arXiv:2410.10815*, 2024. 1, 2, 4

[56] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 1, 2, 3

[57] Xiaodong Yang, Chenyang Zhang, and YingLi Tian. Recognizing actions using depth motion maps-based histograms of oriented gradients. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 1057–1060, New York, NY, USA, 2012. Association for Computing Machinery. 2

[58] Xiaokang Zhou, Wei Liang, Kevin I-Kai Wang, Hao Wang, Laurence T. Yang, and Qun Jin. Deep-learning-enhanced human activity recognition for internet of healthcare things. *IEEE Internet of Things Journal*, 7(7):6429–6438, 2020. 1

[59] Yi Zhu and Shawn Newsam. Depth2action: Exploring embedded depth for large-scale action recognition. In *Computer Vision – ECCV 2016 Workshops*, pages 668–684, Cham, 2016. Springer International Publishing. 1, 2