Egocentric Video Task Translation

Zihui Xue[1,2*] Yale Song[2] Kristen Grauman[1,2] Lorenzo Torresani[2]
[1]The University of Texas at Austin [2]FAIR, Meta AI

Abstract

Different video understanding tasks are typically treated in isolation, and even with distinct types of curated data (e.g., classifying sports in one dataset, tracking animals in another). However, in wearable cameras, the immersive egocentric perspective of a person engaging with the world around them presents an interconnected web of video understanding tasks—hand-object manipulations, navigation in the space, or human-human interactions—that unfold continuously, driven by the person's goals. We argue that this calls for a much more unified approach. We propose EgoTask Translation (EgoT2), which takes a collection of models optimized on separate tasks and learns to translate their outputs for improved performance on any or all of them at once. Unlike traditional transfer or multi-task learning, EgoT2's "flipped design" entails separate task-specific backbones and a task translator shared across all tasks, which captures synergies between even heterogeneous tasks and mitigates task competition. Demonstrating our model on a wide array of video tasks from Ego4D, we show its advantages over existing transfer paradigms and achieve top-ranked results on four of the Ego4D 2022 benchmark challenges.[1]

1. Introduction

In recent years, the introduction of large-scale video datasets (*e.g.*, Kinetics [6, 33] and Something-Something [22]) have enabled the application of powerful deep learning models to video understanding and have led to dramatic advances. These third-person datasets, however, have overwhelmingly focused on the single task of action recognition in trimmed clips [12, 36, 47, 64]. Unlike curated third-person videos, our daily life involves frequent and heterogeneous interactions with other humans, objects, and environments in the wild. First-person videos from wearable cameras capture the observer's perspective and attention as a continuous stream. As such,

[*]Work done during an internship at FAIR, Meta AI.
[1]Project webpage: https://vision.cs.utexas.edu/projects/egot2/.

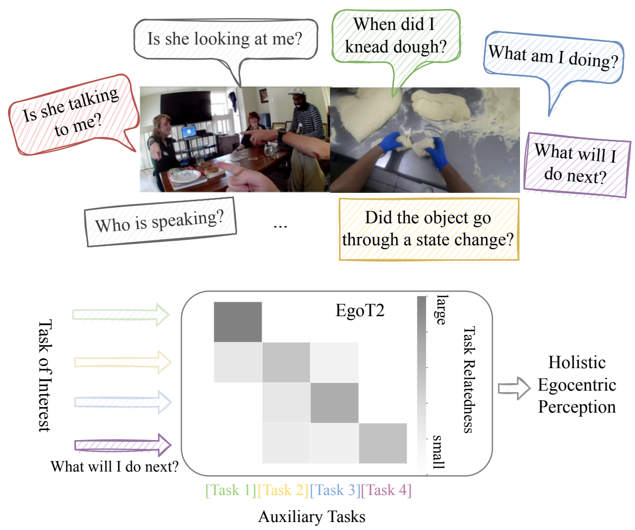

Figure 1. Given a set of diverse egocentric video tasks, the proposed EgoT2 leverages synergies among the tasks to improve each individual task performance. The attention maps produced by EgoT2 offer good interpretability on inherent task relations.

they are better equipped to reveal these multi-faceted, spontaneous interactions. Indeed egocentric datasets, such as EPIC-Kitchens [9] and Ego4D [23], provide suites of tasks associated with varied interactions. However, while these benchmarks have promoted a broader and more heterogeneous view of video understanding, they risk perpetuating the fragmented development of models specialized for each individual task.

In this work, we argue that the egocentric perspective offers an opportunity for *holistic perception* that can beneficially leverage synergies among video tasks to solve all problems in a unified manner. See Figure 1.

Imagine a cooking scenario where the camera wearer actively interacts with objects and other people in an environment while preparing dinner. These interactions relate to each other: a hand grasping a knife suggests the upcoming action of cutting; the view of a tomato on a cutting board suggests that the object is likely to undergo a state transition from whole to chopped; the conversation may further reveal the camera wearer's ongoing and planned actions.

Apart from the natural relation among these tasks, egocentric video's *partial observability* (*i.e*., the camera wearer is largely out of the field of view) further motivates us to seek synergistic, comprehensive video understanding to leverage complementary cues among multiple tasks.

Our goal presents several technical challenges for conventional transfer learning (TL) [65] and multi-task learning (MTL) [63]. First, MTL requires training sets where each sample includes annotations for all tasks [15, 24, 48, 53, 55, 62], which is often impractical. Second, egocentric video tasks are heterogeneous in nature, requiring different modalities (audio, visual, motion), diverse labels (*e.g*., temporal, spatial or semantic), and different temporal granularities (*e.g*., action anticipation requires long-term observations, but object state recognition operates at a few sparsely sampled frames)—all of which makes a unified model design problematic and fosters specialization. Finally, while existing work advocates the use of a shared encoder across tasks to learn general representations [3, 18, 26, 32, 39, 44, 45, 51], the diverse span of egocentric tasks poses a hazard to parameter sharing which can lead to negative transfer [21, 24, 38, 53].

To address the above limitations, we propose EgoTask Translation (EgoT2), a unified learning framework to address a diverse set of egocentric video tasks together. EgoT2 is flexible and general in that it can handle separate datasets for the different tasks; it takes video heterogeneity into account; and it mitigates negative transfer when tasks are not strongly related. To be specific, EgoT2 consists of specialized models developed for individual tasks and a *task translator* that explicitly models inter-task and inter-frame relations. We propose two distinct designs: (1) task-specific EgoT2 (EgoT2-s) optimizes a given primary task with the assistance of auxiliary tasks (Figure 2(c)) while (2) task-general EgoT2 (EgoT2-g) supports task translation for multiple tasks at the same time (Figure 2(d)).

Compared with a unified backbone across tasks [62], adopting task-specific backbones preserves peculiarities of each task (*e.g*. different temporal granularities) and mitigates negative transfer since each backbone is optimized on one task. Furthermore, unlike traditional parameter sharing [51], the proposed task translator learns to "translate" all task features into predictions for the target task by selectively activating useful features and discarding irrelevant ones. The task translator also facilitates interpretability by explicitly revealing which temporal segments and which subsets of tasks contribute to improving a given task.

We evaluate EgoT2 on a diverse set of 7 egocentric perception tasks from the world's largest egocentric video benchmark, Ego4D [23]. Its heterogeneous tasks extend beyond mere action recognition to speaker/listener identification, keyframe localization, object state change classification, long-term action anticipation, and others, and provide a perfect fit for our study. Our results reveal inherent task synergies, demonstrate consistent performance improvement across tasks, and offer good interpretability in task translation. Among all four Ego4D challenges involved in our task setup, EgoT2 outperforms all submissions to three Ego4D-CVPR'22 challenges and achieves state-of-the-art performance in one Ego4D-ECCV'22 challenge.

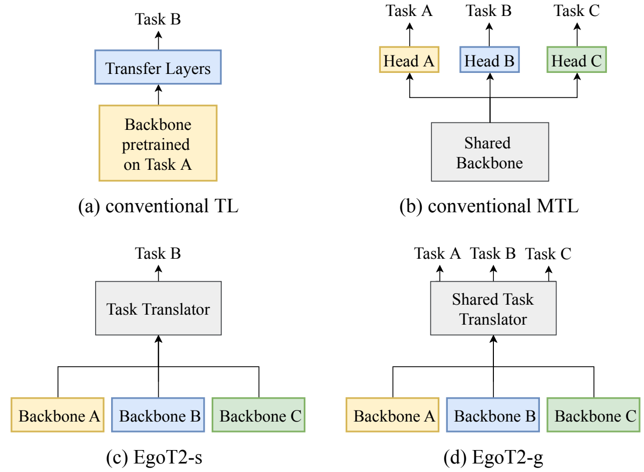

Figure 2. (a) Conventional TL uses a backbone pretrained on the source task followed by a head transferring supervision to the target task; (b) Traditional MTL consists of a shared backbone and several task-specific heads; (c) EgoT2-s adopts task-specific backbones and optimizes the task translator for a given primary task; (d) EgoT2-g jointly optimizes the task translator for all tasks.

2. Related Work

Transfer Learning. TL [65] aims at transferring knowledge from a source domain to improve the performance in a target domain. The most widely adopted approach is to pretrain a model on a source task then finetune on the target task, as shown in Figure 2(a). Following this paradigm, many video classification models [1, 5, 42, 59] are initialized from models pretrained on ImageNet [11]. In addition, many works propose to transfer knowledge from a large-scale video dataset (*e.g*., Kinetics [6, 33]) to benefit action recognition in smaller-scale datasets [54] such as UCF-101 [52] and HMDB-51 [37] or to improve other video tasks, such as spatiotemporal action localization [2, 17, 19, 27, 49] and video anomaly detection [25, 41]. While this technique is ubiquitous in video understanding, prior approaches only consider the transfer from one single source task (dataset) and are thus unable to model the relations among multiple video tasks simultaneously.

Taskonomy [62] presents task transfer with a thorough analysis on the structure of multiple visual tasks. Many works [15, 48, 53, 61] continue along this direction and explore visual task relations, yet they limit the discussion to static images and generally require a unified design across all tasks. In contrast, we consider a diverse set of egocentric video tasks, which are addressed with a heterogeneous set of task-specific video architectures (*e.g.*, accommodating different time, space, or multimodality). Clearly, forcing the same network architecture across all tasks can be suboptimal for each individual task. This motivates our proposed EgoT2-s (Figure 2(c)), where we preserve the heterogeneous backbones developed for each task and build a task translator on top of the task-specific models.

Multi-task Learning. In MTL [63], a single model is trained to address multiple tasks simultaneously in order to capture synergistic supervision across tasks. As depicted in Figure 2(b), hard parameter sharing [51] (*i.e.*, sharing a backbone among tasks and keeping one separate head for each task) is the most commonly used technique within this genre. Although MTL has shown to be beneficial of video analysis [3, 18, 26, 32, 39, 44, 45], there is ongoing debate about the best strategies to determine what parameters to share across which tasks [7, 24, 31, 53, 55]. As pointed out in [34], when MTL is achieved by means of a single common backbone, the performance tends to decrease when the number of tasks grows beyond a certain point. Furthermore, many works [21, 24, 38, 53] observe that over-sharing a network across unrelated tasks causes negative transfer and hinders individual task performance. While soft parameter sharing [14, 60] mitigates this by retaining distinct copies of parameters, it still requires adopting the same identical architecture and "similar" weight values across all tasks.

In the video domain, several works utilize synergies between related tasks (*e.g.*, action recognition with gaze prediction [18, 26, 39] or body pose estimation [44]). However, when selected tasks are not strongly related, prior approaches that split the learning capacity of a shared backbone over multiple tasks can suffer from task competition and inferior performance. In the image domain, with the great advancement of transformers [58], training with multiple datasets together for a generalist model is gaining popularity. Recent work [8, 20, 29, 30, 35, 43] investigates a unified transformer architecture across a diverse set of tasks. Our variant EgoT2-g (Figure 2(d)) is motivated by the desiderata of shared knowledge encapsulated by MTL and of a generalist model. Unlike previous learning paradigms, we adopt a "flipped design" involving separate task-specific backbones and a task translator shared across all tasks. This effectively mitigates task competition and achieves task translation for all tasks simultaneously.

3. Approach

We are given K video tasks, $\mathcal{T}_k$ for $k = 1, \cdots, K$. We note that our approach does not require a common training set with annotations for all tasks. Let the dataset for task $\mathcal{T}_k$ be $\mathcal{D}^{\mathcal{T}_k} = \{(\mathbf{x}_i^{\mathcal{T}_k}, y_i^{\mathcal{T}_k})\}_{i=1}^{N_k}$, where $(\mathbf{x}_i^{\mathcal{T}_k}, y_i^{\mathcal{T}_k})$ denotes the i-th pair of (input video, output label) and N_k represents the number of given examples. Note that "labels" $y_i^{\mathcal{T}_k}$ can be a variety of output types, and are not limited to category labels. For simplicity we omit the subscript i hereafter.

We consider two formulations with distinct advantages: (1) task-specific translation, where we partition the tasks into one primary task $\mathcal{T}_p$ and $K-1$ auxiliary tasks, and optimize the objective to improve performance on $\mathcal{T}_p$ with the assistance of the auxiliary tasks (EgoT2-s, Sec. 3.1); (2) task-general translation, where we treat all K tasks equally, and the goal is to maximize the collective performance of all the tasks (EgoT2-g, Sec. 3.2). As demonstrated in our experiments, objective (1) leads to the strongest performance on the primary task, while objective (2) offers the benefit of a single unified model addressing all tasks at once.

3.1. Task-Specific Translation: EgoT2-s

The training of EgoT2-s is split over two stages.

Stage I: Individual-Task Training. We train a separate model f_k on each individual task dataset $\mathcal{D}^{\mathcal{T}_k}$, obtaining K task-specific models $\{f_k\}_{k=1}^{K}$. We do not place any restrictions on the task-specific model designs, nor do we require a unified design (*i.e.*, identical encoder-decoder architecture) across tasks. Therefore, any available model checkpoint developed for task $\mathcal{T}_k$ can be adopted as f_k within our framework, offering maximum flexibility.

Stage II: Task-Specific Translation. We train a task translator that takes features produced by task-specific models as input and outputs predictions for the primary task. Formally, let $\mathbf{h}_k \in \mathbb{R}^{T_k \times D_k}$ be features produced by the k-th task-specific model f_k, where T_k is the temporal dimension and D_k is the per-frame feature dimension for model f_k. Following the feature extraction step, we design a projection layer $\mathbf{P}_k \in \mathbb{R}^{D_k \times D}$ for each f_k to map task-specific features to a shared latent feature space. This yields a temporal sequence of task-specific tokens $\tilde{\mathbf{h}}_k \in \mathbb{R}^{T_k \times D}$.

We process this collection of task-specific temporal sequences using a transformer encoder [58] of L layers to capture both *inter-frame* and *inter-task* dependencies. We denote the propagation rule of layer l by $\mathbf{z}^{l+1} = Encoder^l(\mathbf{z}^l)$. Finally, we adopt a decoder head $Decoder^{\mathcal{T}_p}$ to obtain predictions for the primary task $\mathcal{T}_p$.

In all, this stage has four major steps: (1) feature extraction; (2) feature projection; (3) transformer fusion; and (4) feature decoding. The procedure is summarized below:

$$\mathbf{h}_k = f_k(\mathbf{x}^{\mathcal{T}_p}), \quad \forall k \in \{1, 2, \cdots, K\} \tag{1}$$

$$\tilde{\mathbf{h}}_k = \mathbf{P}_k \mathbf{h}_k, \quad \forall k \in \{1, 2, \cdots, K\} \tag{2}$$

$$\begin{aligned} \mathbf{z}^0 &= [\tilde{\mathbf{h}}_1, \tilde{\mathbf{h}}_2, \cdots, \tilde{\mathbf{h}}_K] \\ \mathbf{z}^{l+1} &= Encoder^l(\mathbf{z}^l), \forall l \in \{0, 1, \cdots, L-1\} \end{aligned} \tag{3}$$

$$y_{pred}^{\mathcal{T}_p} = Decoder^{\mathcal{T}_p}(\mathbf{z}^L) \tag{4}$$

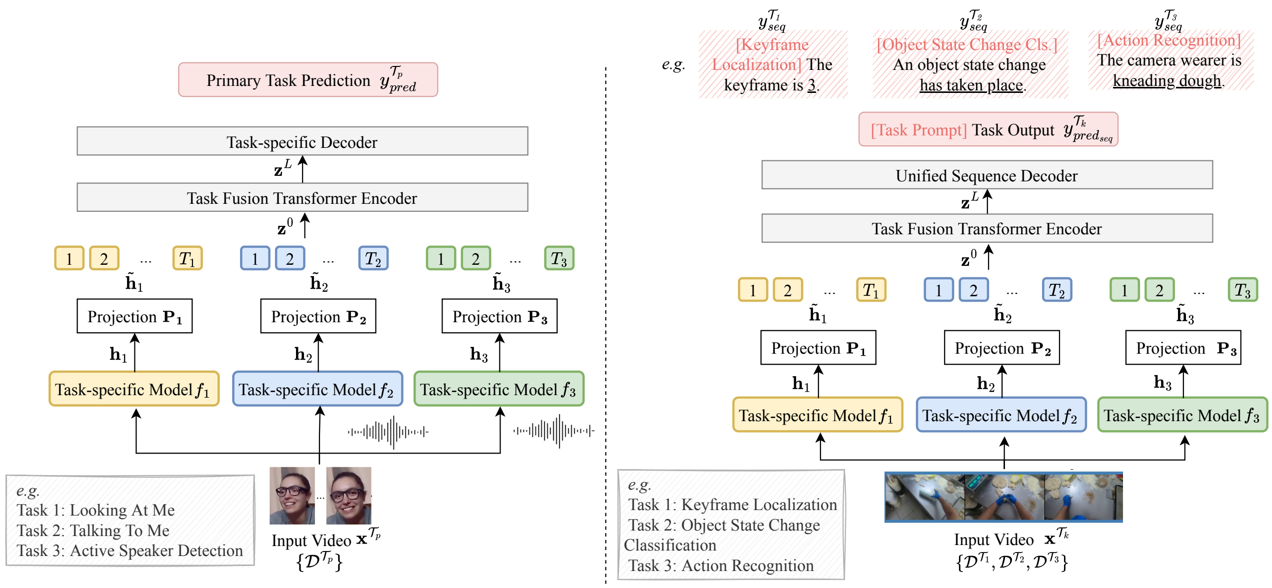

Figure 3. An illustration of EgoT2-s (left) and EgoT2-g (right) on three candidate tasks. The left figure illustrates EgoT2-s on three social interaction tasks, where the input to each model is unimodal (*i.e.*, video) or multimodal (*i.e.*, video and audio). The right figure shows the design of EgoT2-g on three example tasks that focus on different aspects of human-object interactions (*i.e.*, localization, object state change classification, and action recognition). EgoT2-s learns to "translate" auxiliary task features into predictions for the primary task and EgoT2-g conducts task translation conditioned on the task of interest.

where $y_{pred}^{\mathcal{T}_p}$ denotes the prediction given by EgoT2-s. During the second stage of training, we freeze the task-specific models and optimize the task translator with respect to the primary task dataset $\mathcal{D}^{\mathcal{T}_p}$.

Figure 3 (left) illustrates the design of EgoT2-s using three social interaction tasks from Ego4D [23] as an example. EgoT2-s allows heterogeneity in the task-specific models (*i.e.*, f_1 is unimodal while f_2 and f_3 are multimodal; also the three task-specific models can be associated with different frame rates and temporal durations) and utilizes a transformer encoder to model inter-frame and inter-task relations. The resulting EgoT2-s learns to adaptively utilize auxiliary task features for the primary task prediction.

3.2. Task-General Translation: EgoT2-g

EgoT2-s optimizes performance for a single primary task. Therefore, in the event all K tasks must be addressed, it requires K separate training runs and K distinct translators. This motivates us to extend EgoT2-s to perform task translation for all K tasks at once. In EgoT2-g, the task translator processes features from all K tasks and learns to "translate" features conditioned on the task of interest.

The first stage of EgoT2-g is identical to EgoT2-s. For the second stage, we propose two main modifications. First, we replace the task-specific decoder in EgoT2-s with a "generalist" decoder that outputs predictions conditioned on the task of interest. Natural language provides us with a flexible scheme to specify all tasks as a sequence of symbols. Inspired by [8], we tokenize all task outputs and replace the original task-specific decoder with a sequence decoder [50] for a unified interface. Specifically, we first transform the original label $y^{\mathcal{T}_k}$ to a target output sequence $\mathbf{y}_{seq}^{\mathcal{T}_k} \in \mathbb{R}^M$, where M is the target sequence length. For the task translator to produce task-dependent outputs, we prepend a task prompt token $\mathbf{y}_{prompt}$ to the target output, *i.e.*, $\mathbf{y}_{seq_1}^{\mathcal{T}_k} = \mathbf{y}_{prompt}$. We then let the sequence decoder generate a sentence answering the requested task. Figure 3 (right) illustrates how we express task outputs as sequences of discrete tokens and attach task prompts.

With the transformed output, we treat the problem as a language modeling task and train the task translator to predict subsequent tokens (one token at a time) conditioned on the input video and its preceding tokens. The training objective is $\mathcal{L}^{\mathcal{T}_k} = \sum_{j=1}^{M} \mathbf{w}_j \log P(\mathbf{y}_{seq_j}^{\mathcal{T}_k} | \mathbf{x}^{\mathcal{T}_k}, \mathbf{y}_{seq_{1:j-1}}^{\mathcal{T}_k})$. Note that the maximum likelihood loss is weighted to mask the loss corresponding to the task prompt token: $\mathbf{w}_j$ is set to 0 for $j = 1$, and to 1 for any other j. During inference, the task prompt is prepended, and the task translator predicts the remaining output tokens. We use argmax sampling (*i.e.*, take the token with the largest likelihood) to sample tokens from the model likelihood and transform the output tokens back to the original label space. Detokenization is easy as we simply reverse the tokenization process.

The second modification lies in the training strategy. While EgoT2-s adopts the primary task dataset for training, EgoT2-g requires joint training on all K task datasets. Similar to the training strategy in [8, 20], we sample one batch from each task, compute the task loss, aggregate the K gradients, and perform model updates in one training iteration. The final training objective is $\mathcal{L} = \sum_{k=1}^{K} \mathcal{L}^{\mathcal{T}_k}$.

Figure 3 contrasts the design of EgoT2-s and EgoT2-g. They both provide a flexible framework that can incorporate multiple heterogeneous task-specific models (*e.g.*, the three example tasks we give here focus on different aspects of human-object interactions). With a design and an optimization that are specialized to a single primary task, EgoT2-s is expected to lead to superior individual task performance while EgoT2-g brings the efficiency and compactness benefits of a single translator addressing all tasks.

4. Experiments

4.1. Experimental Setup

Dataset and Tasks. We evaluate on Ego4D [23], the world's largest egocentric dataset with 3,670 hours of videos spanning hundreds of scenarios (*e.g.*, household, outdoor, leisure). It offers five benchmarks: episodic memory (EM), hands and objects (HO), audio-visual diarization (AV), social interactions (Social) and forecasting. For our study, we select 7 tasks spanning 4 benchmarks, representing a variety of tasks in egocentric perception, as illustrated in Figure 4. The 7 tasks fall into two broad clusters: (a) human-object interactions and (b) human-human interactions. Table 1 summarizes our task setup. For each cluster, we use tasks from the same benchmark as well as tasks across benchmarks, in an attempt to reveal connections among seemingly unrelated tasks. The 7 candidate tasks are heterogeneous in nature as they are defined on videos of varying duration, adopt different video models as backbones, and process unimodal (*i.e.*, video) or multimodal (*i.e.*, video and audio) input, offering a diverse task setup for our study. See Appendix A.2.1 for more details.

Models and Baselines. For each task, we adopt for consistency the baseline models introduced with the Ego4D dataset[2] as the task-specific **(TS) models** in EgoT2. For task-specific translation (Sec. 4.2), we train one task translator for each primary task and use all the other tasks in the same cluster (either human-object interactions or human-human interactions) as auxiliary tasks. We compare EgoT2-s with two representative transfer learning approaches: (1) **Transfer** [62] denotes finetuning a transfer function on top of features produced by the auxiliary task models (Figure 2(a)). (2) **Late Fusion [45] (LF)** concatenates auxiliary task features along with primary task features, and finetunes a few layers that receive the concatenated features as input for the final prediction. Furthermore, to gauge possible improvements over TS by increasing capacity, we consider a

[2]We use model checkpoints provided on the Ego4D website: https://github.com/EGO4D.

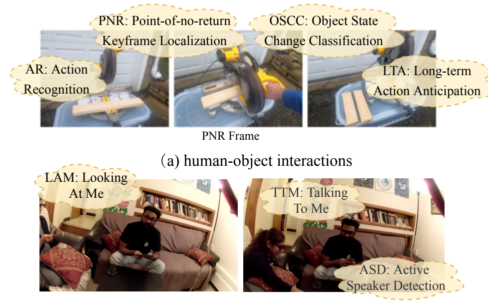

Figure 4. Task Setup. We select a broad set of egocentric video tasks that focus on (a) human-object interactions and (b) human-human interactions from Ego4D benchmarks.

	Task	Benchmark	Mod.	Duration (seconds)	Model backbone
(a)	PNR	HO	V	8.0	I3D RN-50 [6]
	OSCC	HO	V	8.0	I3D RN-50 [6]
	AR	Forecasting	V	8.0	SlowFast [19]
	LTA	Forecasting	V	16.0	SlowFast [19]
(b)	LAM	Social	V	0.2	3D RN-18 [57]
	TTM	Social	A&V	2.7	3D RN-18 [57]
	ASD	AV	A&V	3.7	TalkNet [56]

Table 1. Task Descriptions. 'Mod.' is short for modality; 'A' and 'V' denote audio and video, respectively.

Finetuning [13] baseline, which finetunes a few layers on top of the features produced by the primary task model. In order to make a fair comparison, the first-stage training of these baselines is identical to that of EgoT2, and the number of parameters in the second stage of training is set to match that of EgoT2-s as closely as possible.

For task-general translation (Sec. 4.3), the task translator is jointly optimized for all tasks within a cluster[3], thus we have one task translator for human-object interactions that attends to all tasks simultaneously and one translator that performs three human-human interaction tasks at the same time. For comparison with EgoT2-g, we implement the most widely adopted **multi-task** learning approach, hard parameter sharing [51] (Figure 2(b)).

Implementation Details. There is one video preprocessing step before the feature extraction step in Equation (1), where we transform the original video input from $\mathbf{x}^{\mathcal{T}_p}$ to match the input format of the k-th task-specific model f_k. In particular, $\mathbf{x}^{\mathcal{T}_p}$ is first upsampled or downsampled to match the frame rates required by f_k. Next, if the temporal span of the auxiliary task is smaller than that of the primary task, we slide f_k in a moving window to extract a sequence of features, where the window length is the temporal span required by f_k, and stride size is a hyperparameter. Conversely, if f_k requires video inputs of a longer temporal span

[3]There is a significant domain gap between human-human and human-object interaction videos. See Appendix A.3 for cross-cluster EgoT2-g.

	$\mathcal{T}_p$ is PNR		$\mathcal{T}_p$ is OSCC		$\mathcal{T}_p$ is AR			$\mathcal{T}_p$ is LTA		
	# Params $\cdot 10^6$ Trainable (*All*)	Error (s) ↓	# Params $\cdot 10^6$ Trainable (*All*)	Acc. (%) ↑	# Params $\cdot 10^6$ Trainable (*All*)	Acc. (%) ↑ Verb	Noun	# Params $\cdot 10^6$ Trainable (*All*)	ED@20 ↓ Verb	Noun
TS model [23]	32.2 *(32.2)*	0.615	32.2 *(32.2)*	68.22	63.3 *(63.3)*	22.18	21.55	180 *(242)*	0.746	0.789
Finetuning [13]	8.4 *(40.6)*	0.611	8.4 *(40.6)*	67.93	4.9 *(66.8)*	21.64	22.84	48.6 *(266)*	0.744	0.787
Transfer [62] (PNR)	N/A	N/A	8.4 *(40.6)*	66.80	4.9 *(37.1)*	19.98	5.44	65.4 *(97.6)*	0.778	0.902
Transfer [62] (OSCC)	8.4 *(40.6)*	0.611	N/A	N/A	4.9 *(37.1)*	20.00	9.61	65.4 *(97.6)*	0.774	0.899
Transfer [62] (AR)	9.5 *(71.4)*	0.613	9.4 *(71.4)*	70.98	N/A	N/A	N/A	53.3 *(115)*	0.745	0.806
LF [45] (All Tasks)	9.6 *(135)*	0.610	9.6 *(135)*	72.10	5.2 *(131)*	21.11	19.24	83.6 *(427)*	0.744	0.788
EgoT2-s (All Tasks)	6.4 *(132)*	**0.610**	7.4 *(133)*	**72.69**	4.3 *(130)*	**23.04**	**23.28**	41.8 *(348)*	**0.731**	**0.769**

Table 2. Results of EgoT2-s as we vary the primary human-object interaction task $\mathcal{T}_p$. First row records performance of the task-specific (TS) model we obtain in the first-stage training; we compare EgoT2-s with other baseline methods in the second-stage training. We list the number of trainable parameters for each separate stage as well as the total (*i.e.*, trainable parameters plus parameters of frozen TS models) in parentheses. Following [23], the evaluation metric is temporal localization error (unit: seconds) for PNR, accuracy for OSCC and AR, and edit distance at future 20 time stamps (*i.e.*, ED@20) for LTA. For localization error and ED@20, lower is better. EgoT2-s reliably adapts the auxiliary tasks to suit the target task.

	$\mathcal{T}_p$ is TTM		$\mathcal{T}_p$ is ASD	
	# Params $\cdot 10^6$ Trainable (*All*)	mAP (%) ↑	# Params $\cdot 10^6$ Trainable (*All*)	mAP (%) ↑
TS model [23]	20.2 *(20.2)*	58.91	15.7 *(15.7)*	79.05
Finetuning [13]	0.8 *(20.8)*	59.67	1.1 *(16.8)*	78.62
Transfer [62] (LAM)	0.8 *(15.4)*	63.59	1.6 *(16.2)*	66.40
Transfer [62] (TTM)	N/A	N/A	1.6 *(21.6)*	71.06
Transfer [62] (ASD)	0.8 *(16.5)*	62.31	N/A	N/A
LF [45] (All Tasks)	1.2 *(51.5)*	64.29	1.6 *(51.9)*	77.54
EgoT2-s (All Tasks)	0.7 *(51.1)*	**66.54**	1.5 *(51.9)*	**79.38**

Table 3. Results of EgoT2-s as we vary the primary human-human interaction task $\mathcal{T}_p$. EgoT2-s consistently improves the TS model.

than $\mathbf{x}^{\mathcal{T}_p}$, we exclude task k from auxiliary task candidates to avoid providing potential advantages of a longer observation window to our framework as otherwise we need to expand video length of $\mathbf{x}^{\mathcal{T}_p}$ to match the requirement of f_k. Moreover, if the auxiliary task dataset is multimodal (*i.e.*, video and audio) and the primary task involves only video, we apply the unimodal video pathway of f_k to obtain features; if the primary task is multimodal, we provide all task-specific features that are computable from these modalities. See Appendix A.2.2 for more implementation details.

4.2. Evaluation of Task-Specific Translation

Results. We conduct experiments with EgoT2-s for each task being the primary task[4] and summarize the results for human-object interactions and human-human interactions in Table 2 and 3, respectively.

From the two tables, we observe uneven performance by the baseline methods. Transfer and Late Fusion sometimes outperform the dedicated TS model and sometimes underperform it. When tasks do not exhibit a strong transfer relation, reusing the backbone of the auxiliary task for the primary task leads to negative transfer and performance degradation. For instance, in Table 2, when $\mathcal{T}_p$ is AR, Transfer (OSCC) and Late Fusion both downgrade noun prediction accuracy, suggesting object state change is more dependent on verbs and unrelated to noun prediction tasks in AR.

On the contrary, our proposed EgoT2-s learns to adaptively utilize task-specific features and effectively mitigates negative transfer, demonstrating consistent improvement over the TS model for all 6 cases. For instance, in Table 3, when $\mathcal{T}_p$ is ASD, Late Fusion indicates there is a deleterious relation from LAM and TTM to ASD, as it suffers from an accuracy degradation of 1.51% over TS, yet EgoT2-s still obtains slightly better performance compared to TS (*i.e.*, 79.38% v.s. 79.05%). Moreover, when auxiliary tasks are beneficial for the primary task, EgoT2-s outperforms all baselines with fewer trainable parameters. For example, when $\mathcal{T}_p$ is TTM, it achieves a +7.63% mAP improvement over the original TS model by training a lightweight task translator with only 0.7M parameters on top of it (TS is kept frozen). These results across different primary and auxiliary task combinations help demonstrate the generalizability of EgoT2-s. See Appendix A.3 for experiments using a subset of auxiliary tasks rather than all tasks.

Ablation Study. In Table 4, we ablate three different design choices of EgoT2-s using TTM as the primary task: (a) We replace the LAM and ASD TS models in EgoT2-s with two TTM models with different parameters. This yields a task fusion transformer that is architecturally identical to EgoT2-s but takes only TTM tokens as input; (b) We pass features produced by TS models after temporal pooling as the input of our task fusion transformer; (c) We do not freeze TS models in our second-stage training. By comparing (a) with our default configuration (d), we see that EgoT2-s indeed benefits from the introduction of auxiliary tasks. Although equipped with three different TTM models

[4]Following the time-span guidelines in Sec. 4.1, LAM is not considered as the primary task and LTA is not adopted as an auxiliary task. Nonetheless, Appendix A.3 shows some special cases for completeness.

	# Params $\cdot 10^6$ Trainable (*All*)	Auxiliary Tasks	Temporal Information	Frozen TS model	mAP (%) ↑
(a)	0.7 (*60.8*)		✓	✓	63.40
(b)	0.7 (*51.1*)	✓		✓	65.47
(c)	51.1 (*51.1*)	✓	✓		66.00
(d)	0.7 (*51.1*)	✓	✓	✓	66.54

Table 4. Ablation study of EgoT2-s ($\mathcal{T}_p$ is TTM).

(a)	# Params Trainable	PNR ↓	OSCC ↑	AR Verb ↑	AR Noun ↑	LTA Verb ↑	LTA Noun ↑
TS model [23]	N/A	0.615	68.2	**22.18**	21.55	20.82	21.80
Multi-task [51]	32.2	0.617	66.0	N/A	N/A	N/A	N/A
EgoT2-g (P & O)	5.9	0.612	68.6	N/A	N/A	N/A	N/A
EgoT2-g (All)	34.5	**0.611**	**71.7**	21.93	**22.73**	**21.91**	**23.61**

(b)	# Params Trainable	LAM mAP (%) ↑	TTM mAP (%) ↑	ASD Acc.(%) ↑
TS model [23]	N/A	**77.79**	58.91	79.05
Multi-task [51]	20.2	60.53	61.91	N/A
EgoT2-g	1.4	77.63	**64.49**	**79.06**

Table 5. EgoT2-g for (a) human-object interaction and (b) human-human interaction tasks. The evaluation metric is error (seconds) for PNR (P) and accuracy (%) for OSCC (O), AR and LTA. We report the number of trainable parameters required for each method in the second-stage training (unit: million). Our model is flexible, accurate, and avoids negative transfer.

TTM Challenge	mAP ↑		
Random Guess [23]	0.50		
3D ResNet-18 Bi-LSTM [23]	0.54		
EgoT2-g (3D ResNet-18)	0.58		
EgoT2-s (3D ResNet-18)	**0.58**		
PNR Challenge	Error (s) ↓		
Always Center Frame [23]	1.01		
CNN LSTM [23]	0.76		
EgoVLP [40]	0.67		
Video Swin Transformer [16]	0.66		
SViT [4]	0.66		
EgoT2-s (I3D ResNet-50)	**0.66**		
OSCC Challenge	Acc. ↑		
Always Positive [23]	0.48		
I3D ResNet-50 [23]	0.68		
Video Swin Transformer [16]	0.68		
Divided ST Attention [28]	0.72		
EgoVLP [40]	0.74		
EgoT2-g (I3D ResNet-50)	0.70		
EgoT2-s (I3D ResNet-50)	0.71		
EgoT2-s (EgoVLP)	**0.75**		
LTA Challenge	ED@20 ↓		
	Verb	Noun	Action
SlowFast + Transformer [23]	0.74	0.78	0.94
Video + CLIP [10]	0.74	0.77	0.94
Hierarchical MLP Mixer [46]	0.74	**0.74**	**0.93**
EgoT2-s (SlowFast)	**0.72**	0.76	**0.93**

Table 6. Comparison of EgoT2 with SOTA approaches on four Ego4D challenges (test set). We list the TS model architecture of EgoT2 in parentheses. Our model improves the state of the art.

and a larger model size (the total number of parameters of three TTM models is larger than the sum of three TS models), variant (a) does not bring as much performance gain as EgoT2-s (d). Also, preserving the temporal information of task-specific tokens further boosts performance, as can be seen in the comparison of EgoT2-s (b) with EgoT2-s (d). Finally, not freezing TS (c) greatly increases the training cost yet brings no performance gain. These results validate the design of our proposed EgoT2-s.

4.3. Evaluation of Task-General Translation

Results. Table 5 provides results of EgoT2-g. Since the TTM and LAM baseline models use identical video backbones (*i.e*., 3D ResNet-18), the hard parameter sharing multi-task baseline [51] can jointly learn TTM and LAM. Yet this model design is unable to solve the ASD task without further modifications to the ASD backbone model. In contrast, our EgoT2-g provides a flexible solution that can incorporate a heterogeneous mix of pretrained models. Similarly, we apply the multi-task baseline to PNR and OSCC, as they use the same video backbone (*i.e*., I3D ResNet-50). Compared with dedicated TS models, our proposed EgoT2-g performs task translation for *all* tasks at the same time and achieves on parallel or better performance for all tasks. For instance, it achieves +5.58% mAP improvement for TTM and 3.5% accuracy gain for OSCC. Notably, on ASD, it retains the top-performance of the original TS models when the other two auxiliary tasks do not help. In contrast, we observe task competition for the multi-task baseline: the improvement for TTM (*i.e*., +3.0% mAP) is at the cost of significantly downgraded LAM performance (*i.e*., -17.26% mAP). Similarly, sharing an encoder for PNR and OSCC also leads to task competition and suboptimal performance for the multi-task baseline. For a side-by-side comparison, we also implement EgoT2-g that performs task translation for PNR and OSCC only and observe its advantages over the multi-task baseline in terms of both performance and trainable parameters. As EgoT2-g does not require re-training of the backbone, we can integrate any available model checkpoint developed for each individual task into our framework and train a lightweight task-general translator to further boost performance in the second stage.

Comparison with SOTA Approaches. To further demonstrate the efficacy of both EgoT2-s and EgoT2-g, we submit our model to the EvalAI server to compare it with winning submissions to Ego4D-CVPR'22 and Ego4D-ECCV'22 challenges on the withheld test set. Table 6 shows the results.[5] EgoT2-s achieves top performance for all 4 chal-

[5] ASD & AR are not applicable since they are not Ego4D challenges.

lenges. By only incorporating basic video backbones (*e.g.*, 3D ResNet-18 and SlowFast) as the task-specific model, EgoT2-s achieves similar or better performance than works that adopt more powerful, novel architectures such as Video Swin Transformer. Moreover, the benefits of our approach are orthogonal to such architecture improvements: *e.g.*, for the OSCC challenge, replacing the I3D ResNet-50 backbone with the one used in EgoVLP [40] can further elevate the accuracy of EgoVLP by 1%. This indicates the success of EgoT2 stems from its effective use of task synergies.

While EgoT2-g is a strong performer that surpasses or matches TS across all tasks, if we compare its resultswith those of EgoT2-s, we observe that EgoT2-s demonstrates superior performance. This is understandable given that EgoT2-s is individually optimized for each primary task and employs a specialized translator. On the other hand, EgoT2-g provides a favorable unified framework that performs task translation for all tasks simultaneously via the design of a task-general translator. Thus, EgoT2-s serves as the framework of choice for top performance while EgoT2-g provides added flexibility. See Appendix A.3 for a detailed comparison of the performance and efficiency of these two variants.

4.4. Visualization of Uncovered Task Relations

Our proposed EgoT2 explicitly models task relations via a task translator and offers good interpretability on task relations. For EgoT2-s, Figure 5 shows the attention weights of task tokens when the primary task is LTA and the auxiliary task is AR. Given two adjacent input video clips, the goal of LTA is to predict the next action (*e.g.*, put container and turn off nozzle for the two examples here). In the upper example, there is a scene change from the first clip (the temporal segment corresponding to put wheel) to the second clip (the clip corresponding to take container). The attention weights of AR tokens are small for the first clip and large for the second clip. Clearly, the future action to predict is more closely related to the second temporal segment due to similarities in the scene and objects. In the lower example, the AR tokens have large attention weights, as the video is temporally similar and the previous two actions are indicative of the next action. These results show how EgoT2-s accurately characterizes temporal and auxiliary task information to improve the primary task. More visualizations are in Appendix A.4.

Similarly, for EgoT2-g, we visualize its encoder-decoder attention weights from the last layer transformer in Figure 6. Given the same video clip as input, feature tokens are activated differently when EgoT2-g is given different task prompts, demonstrating that EgoT2-g learns to perform task translation conditioned on the task of interest. As it assigns small weights to task features that are not beneficial for the task of interest (*e.g.*, PNR features to noun prediction tasks), EgoT2-g discards non-relevant task features to

Results of EgoT2-g for PNR & LTA are unavailable (see Appendix A.2.2).

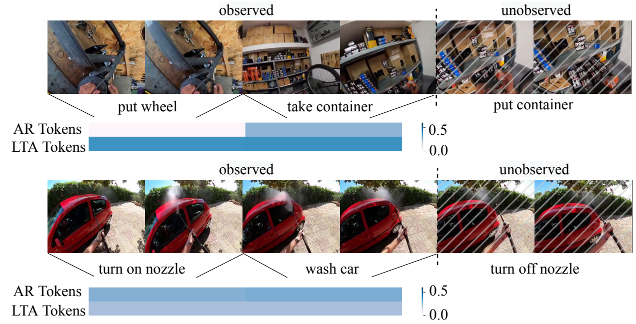

Figure 5. Attention weights of EgoT2-s when $\mathcal{T}_p$ is LTA. EgoT2-s learns to utilize tokens from relevant temporal segments and tasks. The attention weights of AR tokens are large when the current action is indicative of future action.

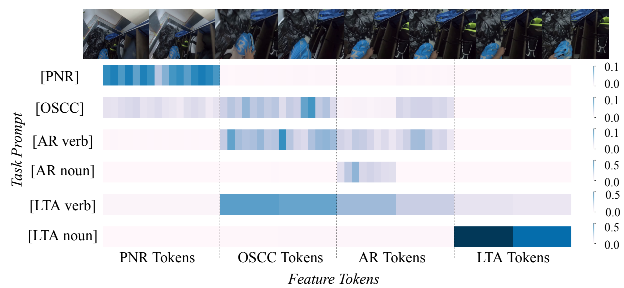

Figure 6. Attention weights of EgoT2-g. Given the same video and different task prompts, EgoT2-g assigns different weights to different task tokens. See text.

mitigate task competition. We also observe temporal differences of attention weights from same task features, indicating that EgoT2-g captures both inter-frame and inter-task dependencies to improve the task of interest. Finally, recall that in Figure 1, we derive task relations for 4 human-object interaction tasks via attention weights provided by EgoT2-g. The attention weights are temporally pooled and averaged over all validation data, revealing task relations from a global perspective. Results for human-human interaction tasks are presented in Appendix A.4. In all, EgoT2 provides good interpretability patterns on (1) which subset of tasks (2) which time segments lead to the final prediction.

5. Conclusion

As a step towards unified egocentric perception, we propose EgoT2, a general and flexible design for task translation. EgoT2 consists of heterogeneous video models optimized for each individual task and a transformer-based task translator that captures inter-frame and inter-task relations. We propose EgoT2-s to improve one primary task and EgoT2-g to conduct task translation for all tasks simultaneously. Results on 7 diverse egocentric video tasks reveal valuable task relations and validate the proposed design.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021. 2

[2] Anurag Arnab, Xuehan Xiong, Alexey Gritsenko, Rob Romijnders, Josip Djolonga, Mostafa Dehghani, Chen Sun, Mario Lučić, and Cordelia Schmid. Beyond transfer learning: Co-finetuning for action localisation. *arXiv preprint arXiv:2207.03807*, 2022. 2

[3] Fabien Baradel, Natalia Neverova, Christian Wolf, Julien Mille, and Greg Mori. Object level visual reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 105–121, 2018. 2, 3

[4] Elad Ben-Avraham, Roei Herzig, Karttikeya Mangalam, Amir Bar, Anna Rohrbach, Leonid Karlinsky, Trevor Darrell, and Amir Globerson. Structured video tokens@ ego4d pnr temporal localization challenge 2022. *arXiv preprint arXiv:2206.07689*, 2022. 7

[5] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. 2

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1, 2, 5

[7] Chao-Yeh Chen, Dinesh Jayaraman, Fei Sha, and Kristen Grauman. Divide, share, and conquer: Multi-task attribute learning with selective sharing. In *Visual attributes*, pages 49–85. Springer, 2017. 3

[8] Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J Fleet, and Geoffrey Hinton. A unified sequence interface for vision tasks. *arXiv preprint arXiv:2206.07669*, 2022. 3, 4, 5

[9] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4125–4141, 2020. 1

[10] Srijan Das and Michael S Ryoo. Video+ clip baseline for ego4d long-term action anticipation. *arXiv preprint arXiv:2207.00579*, 2022. 7

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[12] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jürgen Gall, Rainer Stiefelhagen, and Luc Van Gool. Large scale holistic video understanding. In *European Conference on Computer Vision*, pages 593–610. Springer, 2020. 1

[13] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655. PMLR, 2014. 5, 6

[14] Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 2: short papers)*, pages 845–850, 2015. 3

[15] Kshitij Dwivedi and Gemma Roig. Representation similarity analysis for efficient task taxonomy & transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12387–12396, 2019. 2

[16] Maria Escobar, Laura Daza, Cristina González, Jordi Pont-Tuset, and Pablo Arbeláez. Video swin transformers for egocentric video understanding@ ego4d challenges 2022. *arXiv preprint arXiv:2207.11329*, 2022. 7

[17] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021. 2

[18] Alireza Fathi, Yin Li, and James M Rehg. Learning to recognize daily actions using gaze. In *European Conference on Computer Vision*, pages 314–327. Springer, 2012. 2, 3

[19] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 2, 5

[20] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16102–16112, 2022. 3, 5

[21] Ting Gong, Tyler Lee, Cory Stephenson, Venkata Renduchintala, Suchismita Padhy, Anthony Ndirango, Gokce Keskin, and Oguz H Elibol. A comparison of loss weighting strategies for multi task learning in deep neural networks. *IEEE Access*, 7:141627–141632, 2019. 2, 3

[22] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 1

[23] Kristen Grauman, Michael Wray, Adriano Fragomeni, Jonathan PN Munro, Will Price, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, et al. Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022. 1, 2, 4, 5, 6, 7, 14

[24] Pengsheng Guo, Chen-Yu Lee, and Daniel Ulbricht. Learning to branch for multi-task learning. In *International Conference on Machine Learning*, pages 3854–3863. PMLR, 2020. 2, 3

[25] Matheus Gutoski, Manassés Ribeiro, Leandro T Hattori, Marcelo Romero, André E Lazzaretti, and Heitor S Lopes. A comparative study of transfer learning approaches for video

anomaly detection. *International Journal of Pattern Recognition and Artificial Intelligence*, 35(05):2152003, 2021. 2

[26] Yifei Huang, Minjie Cai, Zhenqiang Li, Feng Lu, and Yoichi Sato. Mutual context network for jointly estimating egocentric gaze and action. *IEEE Transactions on Image Processing*, 29:7795–7806, 2020. 2, 3, 13

[27] Ahsan Iqbal, Alexander Richard, and Juergen Gall. Enhancing temporal action localization with transfer learning from action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[28] Md Mohaiminul Islam and Gedas Bertasius. Object state change classification in egocentric videos using the divided space-time attention mechanism. *arXiv preprint arXiv:2207.11814*, 2022. 7

[29] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J Henaff, Matthew Botvinick, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver IO: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations*, 2022. 3

[30] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 3

[31] Zhuoliang Kang, Kristen Grauman, and Fei Sha. Learning with whom to share in multi-task feature learning. In *ICML*, 2011. 3

[32] Georgios Kapidis, Ronald Poppe, Elsbeth van Dam, Lucas Noldus, and Remco Veltkamp. Multitask learning to improve egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2, 3

[33] 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. 1, 2

[34] Iasonas Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6129–6138, 2017. 3

[35] Alexander Kolesnikov, André Susano Pinto, Lucas Beyer, Xiaohua Zhai, Jeremiah Harmsen, and Neil Houlsby. Uvim: A unified modeling approach for vision with learned guiding codes. *arXiv preprint arXiv:2205.10337*, 2022. 3

[36] Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401, 2022. 1

[37] 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

[38] Isabelle Leang, Ganesh Sistu, Fabian Bürger, Andrei Bursuc, and Senthil Yogamani. Dynamic task weighting methods for multi-task networks in autonomous driving systems. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–8. IEEE, 2020. 2, 3

[39] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European conference on computer vision (ECCV)*, pages 619–635, 2018. 2, 3

[40] Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *arXiv preprint arXiv:2206.01670*, 2022. 7, 8, 15

[41] Kun Liu, Minzhi Zhu, Huiyuan Fu, Huadong Ma, and Tat-Seng Chua. Enhancing anomaly detection in surveillance videos with transfer learning from action recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4664–4668, 2020. 2

[42] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022. 2

[43] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022. 3

[44] Diogo C Luvizon, David Picard, and Hedi Tabia. Multi-task deep learning for real-time 3d human pose estimation and action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(8):2752–2764, 2020. 2, 3

[45] Minghuang Ma, Haoqi Fan, and Kris M Kitani. Going deeper into first-person activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1894–1903, 2016. 2, 3, 5, 6

[46] Esteve Valls Mascaro, Hyemin Ahn, and Dongheui Lee. Intention-conditioned long-term human egocentric action forecasting@ ego4d challenge 2022. *arXiv preprint arXiv:2207.12080*, 2022. 7

[47] Tansel Özyer, Duygu Selin Ak, and Reda Alhajj. Human action recognition approaches with video datasets—a survey. *Knowledge-Based Systems*, 222:106995, 2021. 1

[48] Arghya Pal and Vineeth N Balasubramanian. Zero-shot task transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2189–2198, 2019. 2

[49] Junting Pan, Siyu Chen, Mike Zheng Shou, Yu Liu, Jing Shao, and Hongsheng Li. Actor-context-actor relation network for spatio-temporal action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 464–474, 2021. 2

[50] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 4

[51] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017. 2, 3, 5, 7

[52] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2

[53] Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In *International Conference on Machine Learning*, pages 9120–9132. PMLR, 2020. 2, 3

[54] Jonathan Stroud, David Ross, Chen Sun, Jia Deng, and Rahul Sukthankar. D3d: Distilled 3d networks for video action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 625–634, 2020. 2

[55] Ximeng Sun, Rameswar Panda, Rogerio Feris, and Kate Saenko. Adashare: Learning what to share for efficient deep multi-task learning. *Advances in Neural Information Processing Systems*, 33:8728–8740, 2020. 2, 3

[56] Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3927–3935, 2021. 5

[57] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 5

[58] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[59] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 2

[60] Yongxin Yang and Timothy M Hospedales. Trace norm regularised deep multi-task learning. *arXiv preprint arXiv:1606.04038*, 2016. 3

[61] Amir R Zamir, Alexander Sax, Nikhil Cheerla, Rohan Suri, Zhangjie Cao, Jitendra Malik, and Leonidas J Guibas. Robust learning through cross-task consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11197–11206, 2020. 2

[62] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018. 2, 5, 6

[63] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 2021. 2, 3

[64] Yi Zhu, Xinyu Li, Chunhui Liu, Mohammadreza Zolfaghari, Yuanjun Xiong, Chongruo Wu, Zhi Zhang, Joseph Tighe, R Manmatha, and Mu Li. A comprehensive study of deep video action recognition. *arXiv preprint arXiv:2012.06567*, 2020. 1

[65] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020. 2