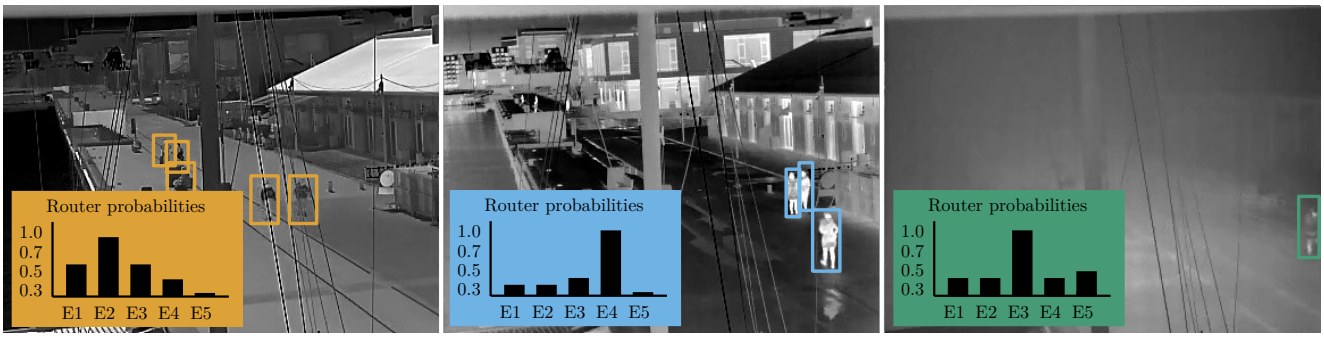
Robust Thermal Image Object Detection via Appearance-Guided Mixture of Experts

Andreas Aakerberg[1,2], Kamal Nasrollahi[1,2],Thomas B. Moeslund[1]
[1]Aalborg University, Denmark, [2]Milestone Systems, Denmark
{anaa,kn,tbm}@create.aau.dk, {andreas.aakerberg,kna}@milestone.dk



Low object/background contrast | High object/background contrast | Haze

Figure 1. As thermal conditions drift, our appearance-guided router activates a small, specialized subset of experts, yielding more true detections and fewer false positives compared to single models or ensembles.

Abstract

Thermal object detection must remain reliable as object and background appearance drifts across time of day, weather, and season. We tackle this challenge with an appearance-guided Mixture of Experts (MoE) that learns to route each image to a subset of specialized backbones. A self-supervised appearance encoder produces embeddings that drive a lightweight router; experts are pretrained on clusters of these embeddings to encourage specialization, and all experts share a single detection head to avoid the linear growth in parameters typical of ensembles. At inference, we adopt a tuning-free, compute-aware policy that activates the fewest experts whose cumulative routing probability exceeds a fixed threshold. Training is stabilized with complementary batch- and sample-level load-balancing losses that prevent expert collapse and promote diverse routing. On LTDv2 (natural long-term drift) and FLIR ADAS (simulated drift), our MoE achieves the highest peak accuracy and superior month-to-month ranking consistency, demonstrating that appearance-guided routing provides more reliable performance across diverse thermal conditions than monolithic scaling. The result is a practical and scalable detector that remains accurate under distribution shift and adapts its compute at test time. Code available at: https://github.com/AndreasAakerberg/agme

1. Introduction

In the deep learning literature, a single end-to-end neural network is typically trained to generalize from training to unseen test data. Yet, in real-world deployments, data distributions often shift over time, leading to degraded performance [29]. This phenomenon is referred to as data drift when the input distribution $p(X)$ changes, or concept drift when the relationship between input and output $p(y|X)$ shifts. Addressing drift is essential in safety-critical applications such as surveillance, where missed detections or false alarms can have serious consequences.

Thermal imaging [22] is widely used in surveillance and urban monitoring due to its robustness to low-light conditions and its inherent ability to preserve privacy [14]. However, unlike RGB images, thermal images lack fine texture and structural details, which pushes models to rely more heavily on shape cues and temperature contrasts between objects and their backgrounds. Furthermore, these contrasts are unstable, because thermal appearance varies with time of day, ambient temperature, solar radiation, and weather conditions [23]. For example, an object that appears hot against a cool morning backdrop may become indistinguishable from a sun-warmed surface later in the day, as seen in Fig. 1. Seasonal changes, such as the transition from winter to summer, further compound this variability. Moreover, because most publicly deployed thermal cameras for monitoring are relative (non-radiometric), their frequent

auto-calibration amplifies fluctuations in image appearance. Consequently, thermal object detection represents a prime setting where models must be able to handle drift in order to maintain robust performance.

Recent studies highlight this complex relationship between thermal image appearance, environmental factors, and detector accuracy [18, 19]. While weather variables, such as temperature and humidity, correlate with detector performance in large-scale thermal datasets [24] (and based on the data seen in Tab. 1, where a linear combination achieves a Pearson correlation coefficient of $r = 0.735$ with monthly mAP@.5:.95 (eight monthly points, Jan–Aug)), prior attempts to incorporate weather metadata directly into detectors have shown limited success [19]. A key reason is likely the thermal inertia of natural materials: for instance, a brick wall may remain warm long after sunset, decoupling the thermal scene from current weather measurements. Consequently, relying solely on metadata is unlikely to yield more robust detections.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
Moisture (%)	91.3	85.4	83.6	66.8	74.1	73.2	80.6	78.4
Sun (Hours)	48.0	97.5	145.9	265.6	285.1	256.3	184.5	239.6
Avg. Temp (°C)	4.1	-0.1	3.7	5.3	9.4	16.3	14.2	17.3
YOLOv8-m (mAP@.5:.95)	0.394	0.325	0.307	0.291	0.276	0.374	0.372	0.367

Table 1. Monthly weather and baseline detector performance on LTDv2 [24]. The baseline's mAP@.5:.95 co-varies with seasonal conditions, highlighting the need to handle appearance drift. Weather from [1]; detector from [17].

A natural response is to increase the model capacity or ensemble multiple detectors [10]. Unfortunately, this does not guarantee stability under drift, *e.g.* CNNs have been shown to outperform larger Transformer models on thermal datasets [19], and ensembles incur linear inference cost [10]. Instead, we argue that robustness requires models that can adapt to domain variability and drift. To this end, we explore Mixture of Experts (MoE) architectures [30]. MoE splits a task across multiple specialized expert models, with a gating network that routes each input to the most relevant expert. Unlike traditional ensembles that aggregate all models' outputs, MoE leverages sparse activation, engaging only a subset of experts per input. This paradigm has recently enabled the scaling of large language models to trillions of parameters without proportional inference costs, as the router implicitly learns to partition the input distribution.

However, our motivation extends beyond scaling. We propose MoE as a mechanism to decouple expert specialization from the detection head. By routing based on self-supervised appearance embeddings rather than class logits, we force experts to specialize in environmental domains to improve robustness in thermal object detection under data drift. Specifically, we introduce an MoE model in which an encoder produces appearance-based embeddings to support routing, encouraging experts to specialize in domain-specific features. By explicitly linking routing to image appearance, our approach enables more resilient detection under varying environmental conditions. Furthermore, at inference time, we activate the fewest experts whose cumulative routing mass exceeds a fixed threshold, providing a simple, tuning-free accuracy–compute trade-off. The main contributions of this paper are:

- **Self-supervised appearance encoder:** A lightweight contrastive encoder learns image-level appearance embeddings tailored to thermal imagery.
- **Appearance drift analysis:** We show that compact, self-supervised appearance embeddings align with seasonal/environmental factors and explain month-to-month performance variation in thermal detection, providing a principled basis for routing and expert specialization.
- **Appearance-guided MoE detector:** A learned router operates on appearance embeddings to select a subset of experts sharing a single detection head. Adaptive selection of experts at inference enables dynamic compute allocation, achieving higher peak accuracy than larger monolithic models while maintaining lower inference cost on simpler examples.
- **Robustness under drift:** On LTDv2 (natural long-term drift) and FLIR ADAS (simulated drift), our approach outperforms strong baselines and ensembles and achieves superior monthly rank-consistency.

2. Related Work

The mixture of experts (MoE) paradigm [6, 20, 30] has gained significant attention in recent years, particularly within large language models, for its ability to scale capacity without increasing inference cost. A MoE model consists of a pool of expert sub-networks, of which only a subset is activated per input through a learned gating mechanism. To prevent expert collapse and maintain balanced utilization, auxiliary load-balancing losses are commonly used, although MoE training remains challenging.

In computer vision, early research in MoE approaches [11] focused on much smaller architectures than today's large MoE-based LLMs [9]. DeepMoE [28] applied conditional computation at the level of convolutional channels or kernels, activating parameters adaptively per instance. Other shallow MoE designs combined CNN experts with a router, trained either separately or jointly [3, 15], sometimes incorporating cost-aware routing [2]. In [4] a router is used to select, and add new experts as tasks change, as an attempt to lifelong learning without catastrophic forgetting. More recently, V-MoE [25] extended sparse expert activation to Vision Transformers, achieving competitive results with dense baselines.

At large scales, models such as the Sparsely Gated MoE [26] and Switch Transformer [12] demonstrated the effectiveness of conditional routing for efficient scaling.

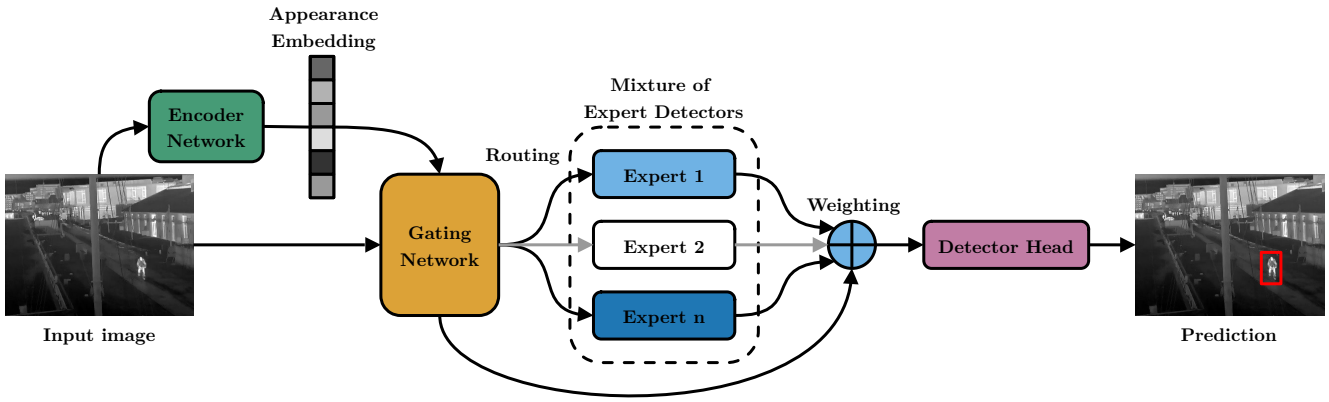


Figure 2. **MoE architecture overview.** Each input image is first mapped into an embedding by the appearance encoder, which drives the gating network to produce expert probabilities. A subset of experts is then selected (either fixed K or adaptively chosen), and their multi-scale features are fused into a weighted representation before passing through the shared detection head.

Subsequent work proposed alternative balancing strategies, including optimal transport formulations [21] and stochastic expert activation [31]. The study most closely related to ours is DriftMoE [5], which addresses concept drift in tabular data streams by using Hoeffding trees as experts. However, its design is inherently incompatible with high-dimensional visual inputs, making direct comparison inapplicable to object detection.

Overall, existing MoE research has focused primarily on efficiency and scaling. In contrast, robustness under drift in vision tasks remains mostly underexplored. To this end, we introduce a MoE-based model tailored to thermal image object detection, where significant drift occurs due to environmental changes. Our approach emphasizes robustness and consistency under shifting conditions, while retaining the efficiency benefits of sparse expert activation. Unlike V-MoE and Switch/DeepMoE, primarily motivated by scaling efficiency, our design targets robustness under drift in thermal detection by coupling routing to learned appearance.

3. Method

Our proposed method differs from ensembles, and generic MoEs in three ways: (i) we route to experts using learned appearance embeddings rather than raw pixels or class logits; (ii) we pre-train experts for specialization by clustering appearance embeddings (PCA→KMeans) and training on nearest samples; and (iii) we share a single detection head across experts, decoupling capacity from output-layer cost.

To illustrate our design, Fig. 2 provides a schematic overview of our proposed MoE architecture, and in the following, we explain each component in detail.

Appearance Representation Learning. To model the appearance of thermal images in a self-supervised manner, we follow MoCo-v2 [8] with a momentum queue and L2-normalized embeddings; we use a two-layer MLP projection head as in [7, 8]. Our encoder Enc is a six-layer CNN with 3×3 convolutions, batch normalization, and LeakyReLU activations, with occasional stride-2 layers for downsampling. This encoder has approximately 1.3M learnable parameters. Each thermal image is randomly cropped twice, yielding paired patches $\{(\mathbf{p}_i^0, \mathbf{p}_i^1)\}_{i=1}^B$, where crops from the same image form positives and crops from different images form negatives. The crops are embedded by $Enc(\cdot)$ and projected through a two-layer MLP projection head (removed at test time) following [7, 8]. We L2-normalize all embeddings so the dot product equals cosine similarity. Let N_- denote the number of negatives drawn from the momentum queue and/or in-batch. We adopt the standard InfoNCE:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp\big(\langle q, k_+\rangle / t\big)}{\exp\big(\langle q, k_+\rangle / t\big) + \sum_{i=1}^{N_-} \exp\big(\langle q, k_i^-\rangle / t\big)}, \tag{1}$$

where $q = Enc(\mathbf{p}^0)$, $k_+ = Enc(\mathbf{p}^1)$, $\{k_i^-\}_{i=1}^{N_-}$ are negatives from a momentum queue (and/or in-batch), $\langle \cdot, \cdot \rangle$ is the dot product (equal to cosine similarity due to L2-normalization), and t is the temperature. The loss is averaged over the B pairs in a mini-batch. This objective pulls together embeddings of patches with identical thermal characteristics while pushing apart those from different images.

In Sec. 4, we demonstrate that our appearance encoder learns meaningful representations that are beneficial for the subsequent routing.

Expert Backbones with Shared Detection Head. Our MoE architecture consists of E expert backbones $\{f_j\}_{j=1}^E$ and a *single* shared detection head H. Given an image x, expert j produces multi-scale features $f_j(x)$. We fuse the selected experts' features into a weighted sum $\mathbf{h}$ and feed $\mathbf{h}$ to H. Sharing H avoids scaling head parameters with E

while still enabling expert specialization. In this work, we experiment with YOLOv8-m style expert backbones and a YOLOv8 detection head, but the routing and head-sharing mechanism is model-agnostic.

Routing (Gating) Mechanism. To dynamically select the most suitable expert(s) for a given input image, a trainable router R (4-layer MLP, ~ 0.52M parameters) maps the appearance embedding $\mathbf{r} = Enc(x) \in \mathbb{R}^d$ to logits $\mathbf{z} \in \mathbb{R}^E$. Softmax then yields expert-selection probabilities:

$$p(j \,|\, \mathbf{r}) \;=\; \frac{\exp(z_j)}{\sum_{\ell=1}^{E} \exp(z_\ell)}, \quad j = 1, \ldots, E.$$

For a fixed number of active experts K, let $S = TopK(p, K) \subset \{1, \ldots, E\}$ denote the indices of the top-K experts (by probability). We renormalize over S via

$$\tilde{p}_j \;=\; \frac{p(j \,|\, \mathbf{r})}{\sum_{i \in S} p(i \,|\, \mathbf{r})} \quad \text{for } j \in S,$$

and fuse expert features as

$$\mathbf{h} \;=\; \sum_{j \in S} \tilde{p}_j \, f_j(x), \qquad \hat{y} \;=\; H(\mathbf{h}).$$

When $K = 1$, the router assigns each image to a single expert; for $K > 1$, it forms a sparse weighted mixture. Gradients flow through the probabilities $p(j \,|\, \mathbf{r})$; the TopK selection S is treated as a non-differentiable set operation. While differentiable routing alternatives such as Gumbel-Softmax could in principle be applied, we found the simpler hard Top-K strategy sufficient in practice.

Adaptive Number of Experts at Inference. At inference time we also experiment with adaptively choosing K. Let $j_{(1)}, j_{(2)}, \ldots$ be expert indices sorted by $p(j \,|\, \mathbf{r})$ in descending order. We pick the smallest K such that the cumulative mass exceeds a threshold $\tau \in (0, 1]$:

$$K \;=\; \min\Big\{k : \sum_{i=1}^{k} p\big(j_{(i)} \,|\, \mathbf{r}\big) \;\geq\; \tau\Big\},$$

$$S \;=\; \{j_{(1)}, \ldots, j_{(K)}\}.$$

We then renormalize over S and proceed as described above. This strategy can preserve nearly all routing confidence while avoiding unnecessary activation of low-probability experts. In all experiments, we set $\tau = 0.90$ without additional tuning. For completeness, Tab. 6 reports an ablation over alternative thresholds.

3.1. Training Procedure

We first pre-train the encoder network on the respective training datasets. To encourage diversity across experts, we pre-train each expert on a subset of the training data. We obtain the subsets by performing K-means on the encoder's 10-D PCA embeddings, and set the number of clusters equal to the number of experts (*e.g.*, 3 or 5), and for each cluster pre-train the corresponding expert on the 1,000 images nearest to its centroid, which we found to provide a good balance between diversity and specialization in preliminary experiments. Next, we freeze the experts and train the router together with the detection head, initializing the head by averaging the weights obtained from the expert pre-training runs. Finally, we unfreeze the experts and train the router + experts + detection head end-to-end, while keeping the appearance encoder frozen.

Because sparsely activated MoEs are prone to load imbalance and expert collapse, we add two complementary router regularizers: a global loss that promotes uniform expert usage across the batch, and a local top-K loss that spreads probability mass among the selected experts.

Batch–level Load Balancing Loss. Let $\bar{p}_j = \frac{1}{B} \sum_{b=1}^{B} p_{b,j}$ denote the average expert usage in a batch, and let $u_j = 1/E$ be the uniform target. Here $p_{b,j}$ denotes the *pre-TopK* softmax probabilities. We minimize

$$\mathcal{L}_{\text{global}} \;=\; \mathrm{KL}(u \,\|\, \bar{p}) \;=\; \sum_{j=1}^{E} u_j \log \frac{u_j}{\bar{p}_j},$$

which strongly penalizes collapsed experts ($\bar{p}_j \to 0$) ; since $u_j = 1/E$, minimizing $\mathcal{L}_{\text{global}}$ is equivalent (up to an additive constant) to minimizing $-\sum_j \log \bar{p}_j$.

Top–K Sample–level Balancing Loss. For each sample b, let S_b be the index set of the selected experts (either $|S_b|{=}K$ in the fixed setting or $|S_b|{=}K_b$ adaptively), and let $\tilde{p}_{b,j}$ be the renormalized probabilities over S_b. We minimize

$$\mathcal{L}_{\text{local}} \;=\; -\frac{1}{B} \sum_{b=1}^{B} \frac{1}{|S_b|} \sum_{j \in S_b} \log \tilde{p}_{b,j}.$$

This equals the cross-entropy to the uniform distribution over S_b and is computed on the TopK-renormalized probabilities $\tilde{p}_{b,j}$, which encourages balanced probability mass among the chosen experts. Together, $\mathcal{L}_{\text{global}}$ and $\mathcal{L}_{\text{local}}$ mitigate both global under-utilization of experts and per-sample collapse onto a single expert.

Implementation Details. Our MoE uses YOLOv8-m backbones with a shared YOLOv8 head. Input sizes 384×384 (LTDv2) and 512×512 pixels (FLIR ADAS), Adam optimizer (β_1=0.937, β_2=0.999), batch size 32, learning rate 10^{-3}, weight decay $5{\times}10^{-3}$, and augmentations (flip, $\pm 10\%$ translation, $\pm 25\%$ scale) are used to train both our, and the compared methods. All training is done on single

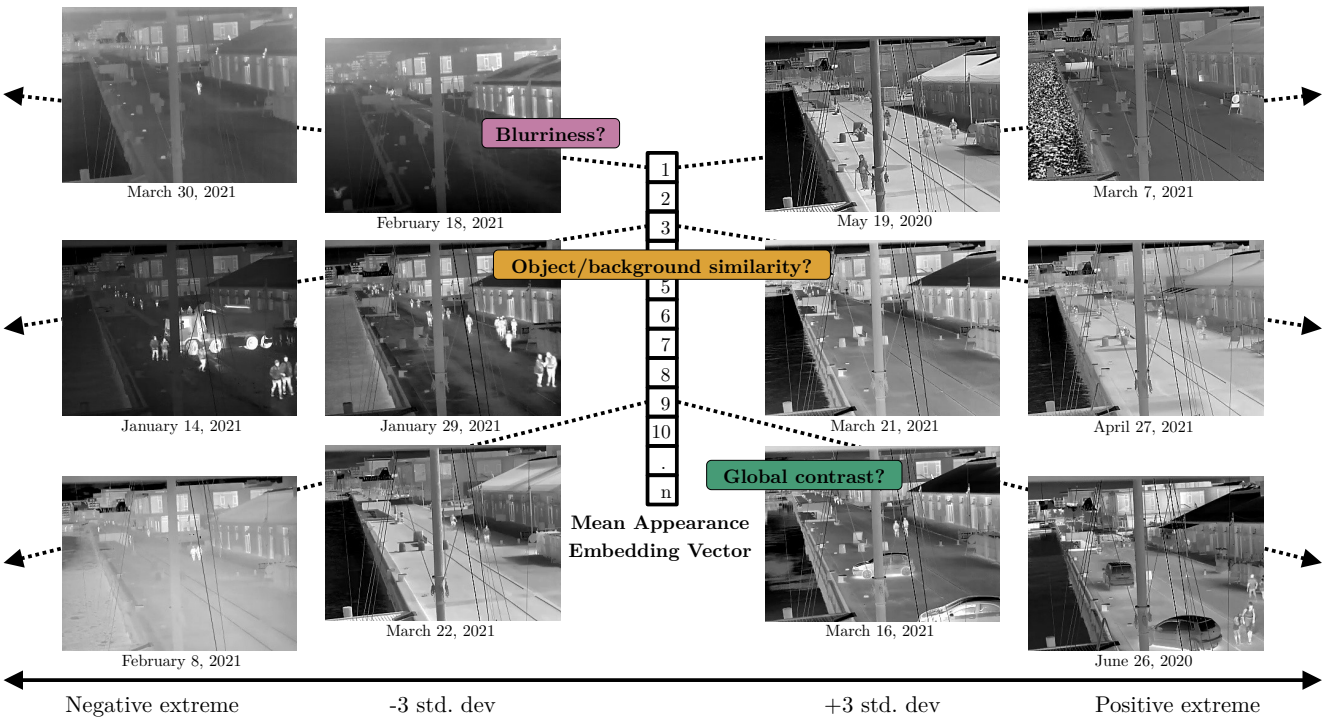


Figure 3. LTDv2 [24] exemplar images from the key principal components of the learned appearance embedding space.

A100 GPUs. The encoder is pre-trained for 100 epochs with 192×192 pixel patches and a queue of 65,536 negatives, and $t = 0.007$ to output an appearance vector of dim. $d = 256$. Experts are pre-trained for 20 epochs, the router and head for 1 epoch, and the full MoE model for 200 epochs. For LTDv2, we subsample 25% per epoch to reduce overfitting, similar to [18]. We weight the global and local router losses by 10.0 and 0.1, respectively, in addition to the standard YOLOv8 detection losses (cls (0.5), bbox (7.5), DFL (1.5)).

4. Experiments

Datasets and Setup. While RGB detectors can leverage massive datasets such as ImageNet and COCO to train powerful pretrained backbones, thermal datasets are typically smaller and often limited to single sessions or seasons. This restricts both the model's ability to learn robust, generalizable features and the evaluation of its robustness to changes in environmental conditions. To study long-term robustness, we use the LTDv2 dataset [24], the only suitable large-scale thermal drift dataset currently available. LTDv2 was collected from the same scene over nine months (Jan.-Sep.) and contains 1.06 million thermal frames with annotations for persons, bicycles, motorcycles, and vehicles. To further investigate generalization to other datasets, we also conduct experiments on the FLIR ADAS dataset [13], which contains 10,228 thermal images annotated for cars, persons, bicycles, and dogs. Since the FLIR ADAS dataset does not include significant drift, we simulate it by randomly applying one of the following degradations to each frame: JPEG compression (quality = [10,50]), Gaussian noise (σ = [10,25]), or leaving the image unchanged, each with equal probability. For both datasets, we adhere to the train and validation splits defined by the authors (test sets are non-public for both datasets).

Baselines. We compare our MoE detector against strong single-model baselines, YOLOv8-m/l [17], and YOLOv11-m/l [16], as well as ensembles of these models. We consider (i) learned feature-level ensembles, where all backbones are activated uniformly (i.e., without routing) for every input and a shared detection head merges their combined features, and (ii) test-time ensembles, where predictions from multiple default models are combined using Weighted Box Fusion (WBF) [27]. In all ensembles, each backbone is initialized with different weights to promote diversity.

Appearance Embeddings. We first seek to determine if the appearance embeddings produced by the encoder network capture information that correlates with appearance changes in the thermal images, and potentially also with detection performance. To simplify this, we reduce the dimensionality of the embeddings from 256 to 25 using PCA, which explains 95.4% of the variance.

First, we explore the principal axes of the reduced embedding space by retrieving exemplar images at varying distances along each principal component. Specifically, let $\bar{\mathbf{e}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{e}_i$ denote the global mean embedding and let $\mathbf{v}_k$ be the kth principal component with corresponding standard deviation σ_k in the PCA-projected data. We then form three synthetic embeddings: $\bar{\mathbf{e}}$, and $\bar{\mathbf{e}} \pm 3\,\sigma_k\,\mathbf{v}_k$.

For each synthetic point, we compute the cosine similarity between that point and all real image embeddings, and select the image whose embedding maximizes this similarity. Figure 3 shows examples at $\pm 3\sigma$, and the extremes, along different principal axes. As seen, the appearance embeddings encode distinct characteristics such as blur level, object/background separation, and global contrast.

To assess potential relationships between the appearance embeddings and the monthly detector performance (YOLOv8-m), we aggregate the embeddings per month and plot the first three principal components alongside the monthly mAP on LTDv2 [24] in Fig. 5. We observe that, dim0 and 2 are negatively correlated with the monthly mAP, while dim-1 is positively correlated. This pattern suggests that a linear combination $(dim1 - (dim0 + dim2))$ of these principal axes could serve as a proxy for distinguishing between "easy" and "difficult" appearance conditions in the thermal images.

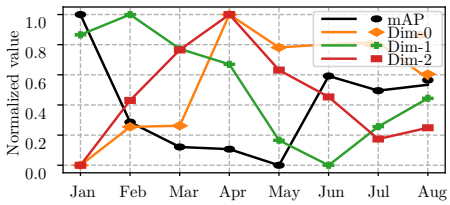


Figure 5. Top 3 embeddings vs. mAP@.5:.95 on LTDv2 [24].

Clustering and Specialization. We perform simple clustering of the embeddings by K-means to determine the possible number of experts needed to specialize to the different appearance conditions in the thermal images. While silhouette analysis indicates $k = 3$, we empirically find a larger k reveals additional sub-distributions in the dataset. Fig. 6, visualizes examples of different cluster center images. We observe that cluster 1 contains images where the image is bright and with low clarity, which is distinct from cluster 6 images, which are detailed and contrast-rich. Cluster 4 images show pronounced object/background separation, while cluster 8 contains images that appear to be corrupted by pixellation.

To investigate the relationship between cluster composition and detector performance, we plot in Fig. 7 the proportion of samples belonging to each cluster for every month, and we quantify the overall embedding diversity by computing the z-score of each month's mean embedding variance relative to the eight-month mean. Notably, January is dominated by clusters 3 and 4 and exhibits a z-score close to 0 (indicating appearance similar to the yearly mean), whereas May shows an almost uniform cluster distribution and the highest z-score (1.92), *i.e.*, maximal variety of appearances. Since the YOLOv8-m baseline detector achieves its highest mAP in January (0.394) and its lowest in May (0.276), these results suggest that increased diversity captured in the appearance embeddings corresponds to lower detection performance, likely because the single model must generalize over a broader range of visual conditions. These embedding analyses support the design of our MoE detector: different experts can specialize to distinct appearance conditions identified in the embedding space, while the router can dynamically select the appropriate subset.

Figure 6. Center images from different clusters in LTDv2[24]. The month of acquisition is indicated in parentheses.

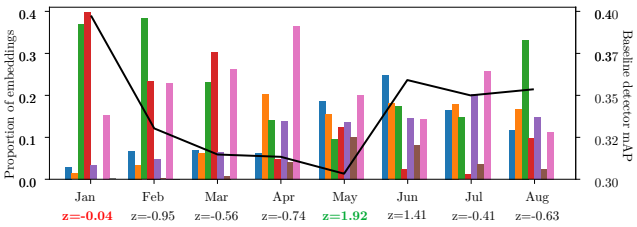


Figure 7. Dist. of cluster density, and z-score per month on LTDv2 [24]. Bar order corresponds to the cluster num. *i.e.* blue=cluster 1.

Object Detection Performance. Table 2 reports accuracy and temporal stability on LTDv2 and generalization on FLIR ADAS. We summarize mAP@.5:.95, mAP@.5, Precision, Recall, and month-to-month variability (by coefficient of variation (CV), σ, and range) to assess robustness under natural drift. Tab. 3 shows per-month rank for each method, while Fig. 4 provides qualitative detection results. To characterize the accuracy–efficiency trade-off, Tab. 4 lists parameter count, GFLOPs, and two efficiency scores (mAP@.5:.95/GFLOP and mAP@.5:.95 per million parameters) computed from LTDv2. Finally, Tab. 5 ablates the total experts E and active experts K, while Tab. 6 gives insights on the adaptive-K threshold τ.

5. Discussion

Performance vs. strong baselines. Table 2 shows that our appearance-guided MoE attains the best results on LTDv2 [24]. Notably, our 3-expert adaptive model (mAP@.5:.95=0.3508) surpasses both the compact

YOLOv8-m | Ours, e=5,k=adaptive | YOLOv8-m | Ours, e=5,k=adaptive | YOLOv8-m | Ours, e=5,k=adaptive

Figure 4. Qualitative results on LTDv2 [24]. Our appearance-guided router activates specialized experts, yielding more correct detections at higher confidence and fewer false positives than the YOLOv8-m baseline.

	LTDv2							FLIR ADAS			
Method	mAP@.5:.95↑	mAP@.5↑	Precision ↑	Recall ↑	CV(%) ↓	σ ↓	Range ↓	mAP@.5:.95↑	mAP@.5↑	Precision ↑	Recall ↑
Baseline, *YOLOv8-m*	0.3319	0.5459	0.6702	0.5482	13.0367	0.0441	0.1180	0.2116	0.4190	0.4936	0.3851
Baseline, *YOLOv8-l*	0.3452	0.5645	0.6779	0.5725	18.8674	0.0678	0.2060	0.2192	0.4347	0.5123	0.3831
Baseline, *YOLOv11-m*	0.3323	0.5443	0.6657	0.5513	20.3381	0.0724	0.2275	0.2251	0.4106	0.7835	0.3370
Baseline, *YOLOv11-l*	0.3369	0.5530	0.6673	0.5511	20.2093	0.0725	0.2332	0.2263	0.4097	0.5253	0.3471
WBF Ensemble, *YOLOv8-m+l*	0.3414	0.5598	0.6522	0.5822	17.2900	0.0610	0.1845	0.2250	0.4399	0.4989	0.4203
WBF Ensemble, *YOLOv11-m+l*	0.3365	0.5516	0.6650	0.5699	17.2653	0.0608	0.1920	0.2318	0.4242	0.4708	0.3987
WBF Ensemble, *YOLOv8-l+v11-l*	0.3420	0.5609	0.6582	0.5788	19.9046	0.0717	0.2259	0.2211	0.4228	0.4665	0.4293
WBF Ensemble, *All baselines*	0.3397	0.5619	0.6454	0.5845	18.1679	0.0652	0.2021	0.2294	0.4388	0.4814	0.4118
Feat. Ensemble, *v8-m, b=2*	0.3382	0.5588	0.6854	0.5576	20.4963	0.0741	0.2380	0.2407	0.4736	0.7699	0.4369
Feat. Ensemble, *v8-m, b=3*	0.3499	0.5722	0.6907	0.5829	16.5080	0.0595	0.1870	0.2375	0.4531	0.5467	0.4043
Feat. Ensemble, *v8-m, b=4*	0.3489	0.5694	0.6884	0.5634	18.9261	0.0691	0.2094	0.2443	0.4729	0.5185	0.4342
Feat. Ensemble, *v8-m, b=5*	0.3491	0.5648	0.6970	0.5584	20.3923	0.0751	0.2287	0.2431	0.4786	0.7965	0.4209
Ours, *v8-m, e=3, k=2*	0.3491	0.5824	0.6897	0.5901	15.6972	0.0569	0.1782	0.2427	0.4744	0.5302	0.4283
Ours, *v8-m, e=3, k=adaptive*	0.3508	0.5800	0.6882	0.5869	16.2884	0.0601	0.1890	0.2455	0.4788	0.5336	0.4311
Ours, *v8-m, e=5, k=3*	0.3538	0.5810	0.6964	0.5828	15.0348	0.0548	0.1617	0.2468	0.4746	0.5136	0.4388
Ours, *v8-m, e=5, k=adaptive*	0.3592	0.5884	0.6993	0.5918	18.4678	0.0699	0.2171	0.2596	0.4962	0.5525	0.4441

Table 2. Evaluation on LTDv2 [24] (left) and FLIR ADAS [13] (right). b, e, k denote the number of base models in the ensemble, total experts, and active experts in the MoE, respectively. For LTDv2, CV(%) denotes the coefficient of variation, and σ the std. dev. of mAP@.5:.95 across all months. Note: While YOLOv8-m show lower CV due to consistently lower performance, our MoE achieves the highest peak performance and ranking stability (see Table 3).

Method	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Mean
Baseline, *YOLOv8-m*	16	16	12	16	11	14	16	16	14.62
Baseline, *YOLOv8-l*	8	7	14	14	9	1	12	7	9.00
Baseline, *YOLOv11-m*	5	14	11	15	8	16	15	14	12.25
Baseline, *YOLOv11-l*	4	9	7	13	12	15	13	10	10.38
WBF Ensemble, *YOLOv8-m+l*	15	8	16	12	16	3	10	12	11.50
WBF Ensemble, *YOLOv11-m+l*	12	10	10	11	15	12	14	13	12.12
WBF Ensemble, *YOLOv8-l+v11-l*	6	2	15	10	14	2	11	11	8.88
WBF Ensemble, *All baselines*	10	4	13	8	13	4	8	8	8.50
Feat. Ensemble, *v8-m, b=2*	2	11	6	7	10	10	9	15	8.75
Feat. Ensemble, *v8-m, b=3*	11	3	8	4	7	6	7	10	7.00
Feat. Ensemble, *v8-m, b=4*	7	12	9	6	4	8	5	5	7.00
Feat. Ensemble, *v8-m, b=5*	3	15	5	9	5	7	4	2	6.25
Ours, *v8-m, e=3, k=2*	13	13	4	2	6	13	6	6	7.88
Ours, *v8-m, e=3, k=adaptive*	9	6	2	1	3	11	2	3	4.63
Ours, *v8-m, e=5, k=3*	14	5	3	5	2	9	3	4	5.62
Ours, *v8-m, e=5, k=adaptive*	1	1	1	3	1	5	1	1	1.75

Table 3. Per-month rank on LTDv2 [24] (1 = best mAP@.5:.95).

Model	Params (M)	GFLOPs	mAP/GFLOP	mAP/Mparam
Baseline, *YOLOv8-m*	25.9	28.4	0.0117	0.0128
Baseline, *YOLOv8-l*	43.7	59.4	0.0058	0.0079
Baseline, *YOLOv11-m*	20.1	22.0	0.0150	0.0164
Baseline, *YOLOv11-l*	25.3	40.7	0.0083	0.0133
WBF Ensemble, *YOLOv8-m+l*	69.6	87.8	0.0049	0.0039
WBF Ensemble, *YOLOv11-m+l*	45.4	62.7	0.0054	0.0074
WBF Ensemble, *YOLOv8-l+v11-l*	69.0	100.1	0.0034	0.0050
WBF Ensemble, *All baselines*	115	150.5	0.0023	0.0030
Feat. Ensemble, *v8-m, b=2*	48.0	50.9	0.0066	0.0070
Feat. Ensemble, *v8-m, b=3*	70.1	73.5	0.0048	0.0050
Feat. Ensemble, *v8-m, b=4*	92.1	96.1	0.0036	0.0038
Feat. Ensemble, *v8-m, b=5*	114.0	118.6	0.0029	0.0031
Ours, *v8-m, e=3, k=2*	49.8	94.6	0.0037	0.0070
Ours, *v8-m, e=3, k=adaptive*	27.7–71.9	72.0–117.1	0.0030–0.0049	0.0049–0.0127
Ours, *v8-m, e=5, k=3*	71.9	117.1	0.0030	0.0049
Ours, *v8-m, e=5, k=adaptive*	27.7–116.0	72.0–162.3	0.0022–0.0050	0.0031–0.0129

Table 4. Comparison of model complexity and efficiency. Params are in millions (M). GFLOPs are measured for a 384×384 single-channel input. Efficiency is reported using LTDv2 results from Tab. 2. For adaptive MoEs, ranges reflect the min/max cost depending on the number of experts activated.

YOLOv8-m (0.3319) and the substantially larger YOLOv8-l (0.3452) and YOLOv11-l (0.3369), indicating that simply scaling capacity is not sufficient under seasonal drift. Instead, conditional computation with specialized experts delivers superior returns. Moreover, while ensembling improves over single detectors, our 5-expert adaptive MoE (mAP@.5:.95=0.3592) outperforms even the strongest feature-level ensemble tested (b=3, mAP@.5:.95=0.3499).

Reliability and Rank Consistency. Regarding temporal robustness, we argue that standard variance metrics must be interpreted with care. The baseline YOLOv8-m exhibits

Active Experts K	Total Experts E			
	2	3	5	8
1	0.3425	0.3452	0.3501	0.3489
2	0.3416	0.3495	0.3447	0.3460
3	-	0.3436	**0.3538**	0.3481

Table 5. Ablation on the total and active number of experts. Entries are mAP@.5:.95 performance on the LTDv2 dataset [24].

	τ=0.3	τ=0.5	τ=0.7	τ=0.9
mAP@.5:.95	0.3231	0.3372	0.3487	0.3592

Table 6. Ablation on the cumulative–mass threshold τ for adaptive-K (LTDv2).

the lowest coefficient of variation (CV) (13.0%) primarily because it remains consistently lower-performing across all months. In contrast, our MoE significantly raises the performance floor in the most challenging conditions (summer months), ensuring high reliability despite the natural fluctuation of thermal difficulty.

Rank captures this robustness more faithfully. As shown in Table 3, our 5-expert adaptive MoE achieves the best mean rank (1.75) and ranks first in most months on LTDv2 (6/8), indicating strong rank stability. In comparison, ensembles oscillate across months, whereas our appearance-guided router consistently selects effective experts and preserves top performance under seasonal domain shift. Qualitative examples under changing environments appear in Fig. 4.

Generalization to Synthetic Drift. On the FLIR ADAS dataset [13] (simulated drift), our model maintains its lead. The 5-expert adaptive model achieves the best overall performance (mAP@.5:.95= 0.2596), balancing high precision and recall better than both the baselines and ensembles. This confirms that our approach generalizes beyond the seasonal transitions of LTDv2 to handle distinct distribution shifts effectively.

Together, these results demonstrate that our approach provides a practical alternative to large single models or ensembles. By explicitly linking routing to interpretable appearance cues, our appearance-guided MoE achieves superior accuracy and month-to-month rank stability, validating its effectiveness for real-world thermal surveillance.

6. Ablations

Table 4 compares the parameter count, computational cost, and efficiency of all methods.

MoE vs. Scaling. We observe that simply scaling up a monolithic detector yields diminishing returns. Moving from YOLOv8-m to YOLOv8-l increases computational cost by 109% (28.4 $\rightarrow$ 59.4 GFLOPs) for a modest 4% gain in mAP@.5:.95(0.3319 $\rightarrow$ 0.3452). Similarly, the YOLOv11-l, while efficient in parameters, fails to surpass the accuracy of the v8-l baseline on this challenging dataset. In contrast, our 5-expert adaptive MoE achieves a significantly higher peak accuracy (0.3592) than these large single models. This confirms that conditional computation through specialized experts can be more effective for handling drift than adding capacity to a generalist backbone.

MoE vs. Ensembles. Ensembles incur a linear cost penalty and the largest WBF ensemble requires 150.5 GFLOPs per image. Our Adaptive MoE provides a superior trade-off as it achieves higher accuracy than the best ensemble while allowing inference costs to drop as low as 27.7 GFLOPs for easy images. This flexibility allows the model to allocate budget where it is needed most, unlike ensembles which waste compute on simple scenes. Note that, in terms of pure mAP@.5:.95/GFLOP, the smallest YOLOv8-m baseline remains the most efficient option. Our MoE trades some efficiency for significantly higher peak accuracy and improved robustness under drift, and is therefore best suited when modest additional compute is acceptable.

Expert Configuration. Finally, Table 5 shows mAP@.5:.95 peaks at E=5, K=3. Notably, performance improves as the total expert pool E increases (from 3 to 5), validating that a wider variety of available experts allows for finer-grained specialization to the diverse thermal conditions found in LTDv2.

7. Conclusion

We introduced an appearance-guided Mixture of Experts (MoE) model for thermal object detection under long-term appearance drift. A self-supervised appearance encoder steers a router that activates only the necessary specialized backbones, while a shared detection head contains output-layer cost. Two simple load-balancing regularizers prevent expert collapse and encourage useful specialization. At inference, a probability-mass–based adaptive-K rule provides compute-aware execution. On LTDv2 (real seasonal drift) and FLIR ADAS, our models surpass strong single backbones and ensembles, delivering higher mAP and better precision/recall, while raising the performance floor in the hardest months and maintaining strong rank stability over time. These results indicate that conditional computation with appearance-guided experts is more effective for drift robustness than scaling monolithic capacity or naive ensembling. In summary, our appearance-guided MoE provides a practical approach to deployable thermal detectors that sustain high accuracy under long-term appearance drift, while keeping computational cost low via adaptive expert activation and a shared detection head.

References

[1] Dmi weather archive. https://www.dmi.dk/vejrarkiv. Accessed: 2025-07-03. 2

[2] Alhabib Abbas and Yiannis Andreopoulos. Biased mixtures of experts: Enabling computer vision inference under data transfer limitations. *IEEE Transactions on Image Processing*, 29:7656–7667, 2020. 2

[3] Karim Ahmed, Mohammad Haris Baig, and Lorenzo Torresani. Network of experts for large-scale image categorization. In *European Conference on Computer Vision*, pages 516–532. Springer, 2016. 2

[4] Rahaf Aljundi, Punarjay Chakravarty, and Tinne Tuytelaars. Expert gate: Lifelong learning with a network of experts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3366–3375, 2017. 2

[5] Miguel Aspis, Sebastián A Ordónez, Andrés L Suárez-Cetrulo, and Ricardo Simón Carbajo. Driftmoe: A mixture of experts approach to handle concept drifts. *arXiv preprint arXiv:2507.18464*, 2025. 3

[6] Ke Chen, Lei Xu, and Huisheng Chi. Improved learning algorithms for mixture of experts in multiclass classification. *Neural networks*, 12(9):1229–1252, 1999. 2

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[8] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3

[9] Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Yu Wu, et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*, 2024. 2

[10] Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. A survey on ensemble learning. *Frontiers of Computer Science*, 14(2):241–258, 2020. 2

[11] David Eigen, Marc'Aurelio Ranzato, and Ilya Sutskever. Learning factored representations in a deep mixture of experts. *arXiv preprint arXiv:1312.4314*, 2013. 2

[12] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. 2

[13] FLIR Systems, Inc. Flir thermal dataset for algorithm training. https://www.flir.com/oem/adas/adas-dataset-form/, 2019. Accessed: 2025-08-18. 5, 7, 8

[14] Rikke Gade and Thomas B Moeslund. Thermal cameras and applications: a survey. *Machine vision and applications*, 25 (1):245–262, 2014. 1

[15] Sam Gross, Marc'Aurelio Ranzato, and Arthur Szlam. Hard mixtures of experts for large scale weakly supervised vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6865–6873, 2017. 2

[16] Glenn Jocher and Jing Qiu. Ultralytics yolo11, 2024. 5

[17] Glenn Jocher and Ultralytics. YOLOv8: YOLO Plus – state-of-the-art real-time object detection. https://github.com/ultralytics/ultralytics, 2023. Version 8.0.0. 2, 5

[18] Anders Skaarup Johansen, Julio CS Jacques Junior, Kamal Nasrollahi, Sergio Escalera, and Thomas B Moeslund. Chalearn lap seasons in drift challenge: Dataset, design and results. In *European Conference on Computer Vision*, pages 755–769. Springer, 2022. 2, 5

[19] Anders Skaarup Johansen, Kamal Nasrollahi, Sergio Escalera, and Thomas B Moeslund. Who cares about the weather? inferring weather conditions for weather-aware object detection in thermal images. *Applied Sciences*, 13(18): 10295, 2023. 2

[20] Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2): 181–214, 1994. 2

[21] Mike Lewis, Shruti Bhosale, Tim Dettmers, Naman Goyal, and Luke Zettlemoyer. Base layers: Simplifying training of large, sparse models. In *International Conference on Machine Learning*, pages 6265–6274. PMLR, 2021. 3

[22] J Michael Lloyd. *Thermal imaging systems*. Springer Science & Business Media, 2013. 1

[23] Ivan Adriyanov Nikolov, Mark Philip Philipsen, Jinsong Liu, Jacob Velling Dueholm, Anders Skaarup Johansen, Kamal Nasrollahi, and Thomas B Moeslund. Seasons in drift: A long-term thermal imaging dataset for studying concept drift. In *Thirty-fifth Conference on Neural Information Processing Systems*. Neural Information Processing Systems Foundation, 2021. 1

[24] Marco Parola, Andreas Aakerberg, Anders S. Johansen, Ivan A. Nikolov, Mario G.C.A. Cimino, Kamal Nasrollahi, and Thomas B. Moeslund. Ltdv2: A large-scale long-term thermal drift dataset for robust multi-object detection in surveillance. 2025. 2, 5, 6, 7, 8

[25] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021. 2

[26] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017. 2

[27] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107: 104117, 2021. 5

[28] Xin Wang, Fisher Yu, Lisa Dunlap, Yi-An Ma, Ruth Wang, Azalia Mirhoseini, Trevor Darrell, and Joseph E Gonzalez. Deep mixture of experts via shallow embedding. In *Uncertainty in artificial intelligence*, pages 552–562. PMLR, 2020. 2

[29] Liheng Yuan, Heng Li, Beihao Xia, Cuiying Gao, Mingyue Liu, Wei Yuan, and Xinge You. Recent advances in concept drift adaptation methods for deep learning. In *IJCAI*, pages 5654–5661, 2022. 1

[30] Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty years of mixture of experts. *IEEE transactions*

on neural networks and learning systems, 23(8):1177–1193, 2012. 2

[31] Simiao Zuo, Xiaodong Liu, Jian Jiao, Young Jin Kim, Hany Hassan, Ruofei Zhang, Tuo Zhao, and Jianfeng Gao. Taming sparsely activated transformer with stochastic experts. *arXiv preprint arXiv:2110.04260*, 2021. 3