Extreme Amodal Face Detection

Changlin Song[1]   Yunzhong Hou[1]   Michael Randall Barnes[2]   Rahul Shome[1]   Dylan Campbell[1]

[1]Australian National University   [2]University of Oslo

{changlin.song, yunzhong.hou, rahul.shome, dylan.campbell}@anu.edu.au

michael.barnes@ifikk.uio.no

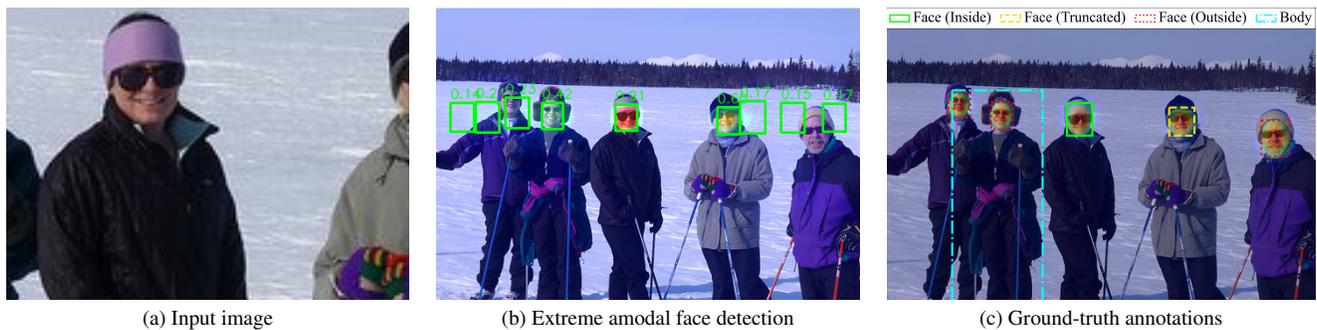(a) Input image   (b) Extreme amodal face detection   (c) Ground-truth annotations

Figure 1. Extreme amodal face detection. This task predicts, given an input image, the likelihood of faces at all locations within an expanded field-of-view frame. Specifically, a face presence heatmap and bounding boxes are estimated both inside and outside the image. In the example pictured, there is direct visual evidence of three faces (one in-frame, one partially in-frame, and one with a partially-observed correlate—the person's body), and two more faces without direct evidence but with a non-zero conditional probability.

Abstract

Extreme amodal detection is the task of inferring the 2D location of objects that are not fully visible in the input image but are visible within an expanded field-of-view. This differs from amodal detection, where the object is partially visible within the input image, but is occluded. In this paper, we consider the sub-problem of face detection, since this class provides motivating applications involving safety and privacy, but do not tailor our method specifically to this class. Existing approaches rely on image sequences so that missing detections may be interpolated from surrounding frames or make use of generative models to sample possible completions. In contrast, we consider the single-image task and propose a more efficient, sample-free approach that makes use of the contextual cues from the image to infer the presence of unseen faces. We design a heatmap-based extreme amodal object detector that addresses the problem of efficiently predicting a lot (the out-of-frame region) from a little (the image) with a selective coarse-to-fine decoder. Our method establishes strong results for this new task, even outperforming less efficient generative approaches. Code, data, and models are available at https://charliesong1999.github.io/exaft_web/.

1. Introduction

Object detection has been a central problem in computer vision for decades [30, 31], with significant advances in closed-set detection of predefined categories [31] and open-set detection that generalizes beyond fixed taxonomies [30]. However, existing detectors are fundamentally constrained to objects visible within the input frame. This restricts their applicability in scenarios that require extrapolating beyond what is directly observable.

We take a step toward this broader goal by introducing the task of *extreme amodal detection*, where the objective is to detect and localize objects that may lie partially or entirely outside the visible field-of-view. While our design is applicable to the general task, in this paper we focus on the sub-problem of extreme amodal *face* detection, which is especially well-motivated due to its relevance to safety-critical (e.g., anticipating pedestrians), accessibility-related (e.g., assisting those with visual impairments [4]), and privacy-sensitive applications. As shown in Figure 1, we categorize extreme amodal faces into (1) *truncated faces*, which are partially within the field-of-view; and (2) *outside faces*, where the face is completely outside the field-of-view. The latter is subdivided into two cases: (2a) *with evidence*, where direct visual evidence, such as a visible body,

is observed; and (2b) *without evidence*, where the model must rely on indirect contextual cues.

The impact on privacy is especially relevant, and worth elaborating as it explains our focus on human faces in particular. In brief, extreme amodal face detection can improve privacy by enabling computer vision systems to actively avoid capturing sensitive information, i.e., human faces. Cameras in public spaces pose inherent privacy risks, and cameras that move in public spaces (e.g., on self-driving cars, drones, or other semi-autonomous robotic systems) exacerbate those risks. Existing solutions often aim to secure data during post-processing, for example, by detecting and blurring faces. This is not a robust strategy, however, as raw data is susceptible to theft [6, 7, 19], corporate misuse [14, 20, 25], or legally enforced retrieval [3, 24]. More fundamentally, this strategy overlooks data collection as a site of intervention, and that the best privacy-preserving strategy is often to not collect sensitive data at all. Extreme amodal face detection can serve this end. If deployed successfully, it can limit the need for actual surveillance, preserving privacy without sacrificing utility.

Prior work provides only limited tools for this task. Tracking-based methods [9] leverage temporal continuity in video to recover partially unseen objects, but do not address the case of a single static frame. Another line of work relies on generative pipelines that outpaint the extended frame using diffusion-based models [2, 15], followed by conventional detectors. While straightforward, these approaches have several drawbacks: (a) they depend heavily on additional prompts (e.g., text or masks) whose quality can significantly affect the results; (b) diffusion models are computationally expensive and slow at inference time, making them ill-suited for time-critical detection scenarios; and (c) these pipelines are not end-to-end trainable, limiting their ability to adapt to new detection tasks. In contrast, humans can readily infer the existence and location of unseen objects based on prior knowledge, contextual cues, or reasoning from visible body parts.

The extreme amodal setting introduces three unique challenges. First, the extended region can, in principle, be arbitrarily larger than the input image. In our setup, we restrict this extension to $8\times$ the input size, which nonetheless requires the long-distance extrapolation of information. Second, naively querying the entire extended region is computationally prohibitive, requiring up to $8\times$ more tokens and wasting resources on regions that often contain no objects. Third, the underlying true conditional distribution cannot be accessed; we only have a single realization for any input image. This poses a challenge for evaluation, where we can only measure success indirectly using the ground-truth realization, as discussed in Sec. 6.

To address the first two issues, we propose a *coarse-to-fine selective decoder* that makes good use of limited in-

formation while remaining compute-efficient. Our decoder first queries the extended area at low resolution, dividing it into candidate regions. It then selectively refines only a subset of promising candidates to match the resolution of the input image. This design reduces the number of tokens by lowering the resolution at the initial stage and by refining only the most relevant candidates. As a result, our approach achieves both efficiency and strong detection performance. Our contributions are threefold. We

1. introduce extreme amodal face detection, the task of detecting and localizing faces partially or entirely outside the visible field-of-view;
2. construct a benchmark dataset derived from COCO [13] images, enabling systematic evaluation for faces inside the image, outside the image, and truncated by the image frame; and
3. design an efficient and effective extreme amodal detector with a novel coarse-to-fine selective decoder.

2. Related Work

Existing works related to our task can be broadly grouped into two categories: *tracking-based approaches*, which leverage temporal information across multiple frames, and *generative-based approaches*, which rely on large generative models conditioned on additional prompts such as masks or text.

Tracking-based methods. OccludTrack [21] introduced the problem of tracking objects even when fully invisible, either due to occlusion or containment. Their dataset was collected via simulation and manual labeling. Cotracker [10] extended point tracking by jointly tracking all points, demonstrating strong robustness in fully occluded and out-of-frame scenarios. TAO-amodal [9] expanded bounding boxes of pre-trained trackers beyond the visible frame by exploiting temporal consistency. ObjectRemember [16] lifted object points into 3D coordinates, storing them in memory to persist objects even when they leave the frame. While these methods can estimate truncated or outside objects, they inherently require temporal cues across multiple frames. In contrast, our work investigates how to detect such objects from a *single* static frame.

Generative-based methods. A second line of work uses generative models to complete or outpaint missing regions. In amodal completion, Pix2Gestalt [15] employed SAM [11] to obtain masks and fine-tuned a diffusion model for part-whole completion. PD-MC [26] used grounded-SAM [18] with text prompts to automatically generate masks, then progressively completed objects. OpenACC [1] further incorporated both masks and background context to reason about text prompts for flexible completion. These methods, however, primarily address the occlusion problem but do not necessarily generalize well to cases where objects

of interest are truncated or completely invisible. However, our method can infer completely invisible objects.

For outpainting, PQ-Diff [28] trained a diffusion model with positional queries for arbitrary-size extrapolation, though performance degrades in complex scenes. VIP [27] employed large multimodal models to provide semantic supervision during outpainting. Unseen [2] generated the unseen regions with additional text prompts before applying a detector. Despite their creativity, generative-based pipelines share key drawbacks: they rely on external prompts (mask or text), require large diffusion models that are computationally expensive and slow at inference, and are not end-to-end trainable. These limitations make them unsuitable for fast and efficient detection scenarios, such as detecting out-of-frame pedestrians in autonomous driving.

3. Extreme Amodal Detection

Given an image $\boldsymbol{x} \in \mathbb{R}^{H \times W \times 3}$, extreme amodal detection predicts the location of objects within a centrally-expanded region of size $KH \times KW$, where K denotes the expansion factor. To predict objects within this larger region, we consider two output types, commonly associated with the tasks of detection and localization. For the detection task, a set of N objects $o_i = (c_i, b_i)$ are predicted, where c_i denotes the object class and $b_i = (x_i, y_i, w_i, h_i)$ denotes the bounding box represented by center coordinate, width and height. For the localization task, a heatmap $\boldsymbol{h} \in [0,1]^{KH \times KW \times C}$ is predicted, where C denotes the number of classes, indicating the probability that an object of each class is located at that pixel. As motivated in the introduction, in this paper we consider a single class: human faces.

As shown in Figure 1, the difficulty of detecting extreme amodal faces varies, depending on whether there is direct visual evidence within the image of a face wholly or partially outside the image. We classify faces as
1. **Inside:** faces that are entirely within the image;
2. **Truncated:** faces that are partially within the image; and
3. **Outside:** faces that are entirely outside the image,
   (a) with direct visual evidence, such as a visible body in the image; and
   (b) without direct visual evidence, where indirect cues like eye gaze and semantic co-occurrences may need to be considered.

4. The EXAFace Dataset

In this section, we introduce the Extreme Amodal Face (EXAFace) dataset, derived from the MS COCO [13] object detection dataset. First, RetinaFace [5] was used to pseudolabel the many unlabeled faces in the COCO dataset, excluding those detections with a confidence below 0.9, resulting in $2.4\times$ more face labels. Next, the images were randomly cropped and the bounding boxes from the cropped

# $\times 10^3$ (%)	Inside	Truncated	Outside +	Outside -
Boxes (train)	116 (24%)	74 (15%)	66 (13%)	235 (48%)
Boxes (test)	5.0 (24%)	3.0 (14%)	2.0 (12%)	11 (50%)
Images (train)	30 (17%)	37 (20%)	32 (17%)	83 (46%)
Images (test)	1.0 (16%)	1.5 (20%)	1.0 (17%)	3.5 (47%)

Table 1. EXAFace dataset statistics. Sample counts ($\times 10^3$) and percentages are shown for bounding boxes and images. The data is divided into subsets of inside faces, truncated faces, outside faces with direct evidence (+), and outside faces without direct evidence (-). The category of an image is determined by its hardest face.

and uncropped regions were retained. For an image with height H and width W, the process is as follows.
1. Randomly sample crop height from $[0.3H, 0.6H]$ and aspect ratio from $[0.5, 2]$, yielding the crop size $H' \times W'$.
2. Randomly sample center x coordinate from $[0.5W', W - 0.5W']$ and y coordinate from $[0.5H', H - 0.5H']$.
3. Crop image using crop size and center.
4. Discard bounding boxes that are not fully contained within an expanded area $K^2 \times$ the size of the crop.
5. Update the bounding box center coordinates (x_b, y_b) to the expanded image coordinate frame: $(x_b - x + 0.5KW', y_b - y + 0.5KH')$.

This is repeated 4 times per image to generate diverse data. The dataset statistics are given in Tab. 1.

5. Extreme Amodal Face Detector

In this section, we outline our extreme amodal face detector, as shown in Figure 2. Our method involves feature extraction, a transformer encoder–decoder for sharing information between in-image tokens and out-of-image tokens, and two detection heads, one for in-image faces and one for out-of-image faces. First, a convolutional feature extractor f_{feat} computes a feature map $\boldsymbol{y}_{\text{in}}$ given the image. Then, a transformer encoder f_{enc} processes these features into a form useful for predicting out-of-image faces, given rotary positional encodings $\boldsymbol{p}_{\text{in}} = \phi(\mathcal{C}_{\text{in}})$ of the in-image coordinates $\mathcal{C}_{\text{in}}$ [8]. Next, our selective course-to-fine transformer decoder f_{dec} cross-attends to the in-image features, given the positional encodings $\boldsymbol{p} = \phi(\mathcal{C})$ of the expanded image coordinates $\mathcal{C}$. Finally, two detection heads g predict in- and out-of-image objects o and heatmaps $\boldsymbol{h}$. In summary, we have

$$\boldsymbol{y}_{\text{in}} = f_{\text{feat}}(\boldsymbol{x}) \tag{1}$$

$$\boldsymbol{z}_{\text{in}} = f_{\text{enc}}(\boldsymbol{y}_{\text{in}}, \boldsymbol{p}_{\text{in}}) \tag{2}$$

$$\boldsymbol{y}_{\text{out}} = f_{\text{dec}}(\boldsymbol{z}_{\text{in}}, \boldsymbol{p}) \tag{3}$$

$$(o_{\text{in}}, \boldsymbol{h}_{\text{in}}) = g_{\text{in}}(\boldsymbol{y}_{\text{in}}) \tag{4}$$

$$(o_{\text{out}}, \boldsymbol{h}_{\text{out}}) = g_{\text{out}}(\boldsymbol{y}_{\text{out}}). \tag{5}$$

(a) Extreme amodal detector flowchart
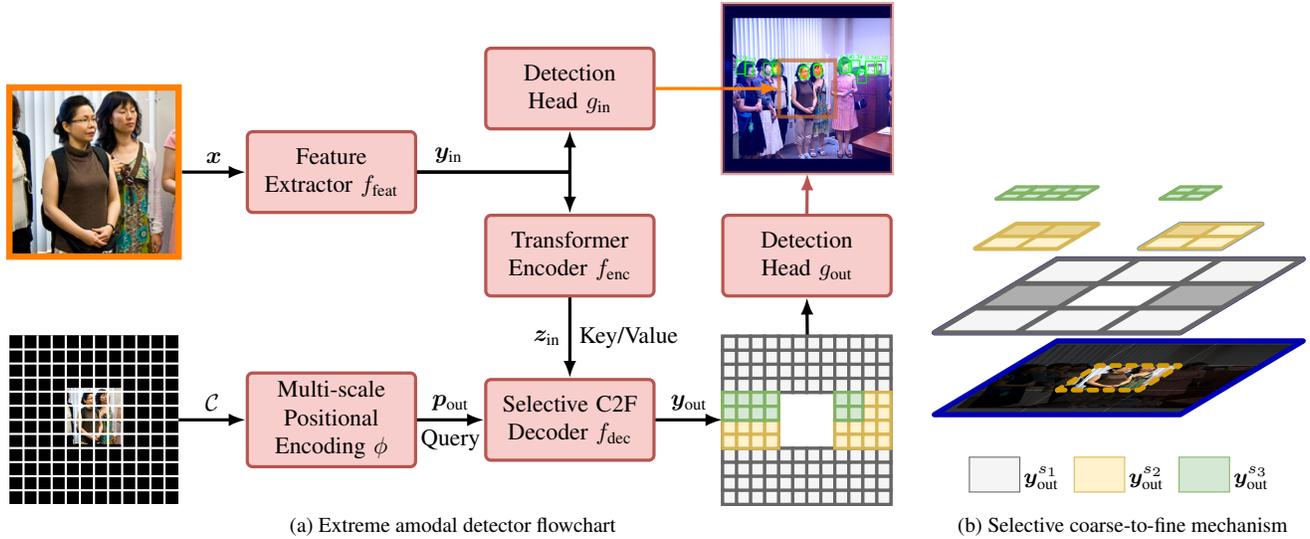
(b) Selective coarse-to-fine mechanism

Figure 2. Overview of our extreme amodal detector. (a) Flowchart of our approach. Given an input image, a feature map is extracted, from which a dedicated in-image detection head infers object boxes and a face probability heatmap. Separately, a transformer encoder–decoder shares information from the image to the extended area around the image. We propose an efficient selective coarse-to-fine decoder that starts with low resolution out-of-image positional encodings as the input tokens, then refines a selected subset of these tokens at higher resolutions. A second detection head uses these tokens to infer the out-of-image object boxes and heatmap. (b) Illustration of our selective coarse-to-fine mechanism. We first query the low-resolution regions, then use a scoring network to rank these regions and select the top-$\mu\%$ to be refined at a higher resolution, until at the same resolution as the input image feature map.

The main novelty of the approach arises from the transformer decoder, which will now be outlined in detail.

Selective course-to-fine (C2F) decoder. Sharing information between the image and the extended region beyond the image is challenging for two reasons: (a) high computational cost: if using the same resolution, the extended region has $(K^2 - 1)\times$ more tokens than the input image; and (b) object sparsity: only a small proportion of image patches contain objects (in our dataset, fewer than 1% of the 16×16 pixel patches contain faces). However, it is not possible to know which patches contain objects in advance. To address this, we propose a selective coarse-to-fine mechanism: first query the extended region at low resolution, then use a scoring network to select promising regions for refinement.

The approach is as follows. As indicated in Equation (3), the transformer decoder receives the in-image features z_{in}, which are projected into keys and values, and the positional encodings p. For the first decoder layer, low-resolution, coarse positional encodings $p_{\text{out}}^{s_1}$ from the extended region around the image are projected into queries. The positional encodings are given by

$$p_{\text{out}}^{s_i} = \{\text{avgpool}(p, s_i)(u, v) \mid (u, v) \in \mathcal{C}_{\text{out}}\}, \quad (6)$$

where $\text{avgpool}(\cdot, s_i)$ denotes average pooling with an $s_i \times s_i$ window, $\mathcal{C}_{\text{out}}$ denotes the out-of-image coordinates, and $s_i \in \mathcal{S}$ are a sequence of coarse-to-fine scales. The decoder layer uses ROPE positional encodings [8] to facilitate

cross-attention between within-image and out-of-image tokens at the requisite scale. After the first 2-layer decoder block f_{decblk}, a scoring network f_{score} predicts which tokens to refine at a higher resolution, retains only the top-$\mu^{s_i}\%$ tokens, and duplicates these to match the number required by the next resolution level.

In summary, initialization sets $x_{\text{out}}^{s_1} \leftarrow p_{\text{out}}^{s_1}$ and then the per-block computations proceed as

$$y_{\text{out}}^{s_i} = f_{\text{decblk}}(x_{\text{out}}^{s_i}, p_{\text{out}}^{s_i}, z_{\text{in}}, p_{\text{in}}) \quad (7)$$

$$x_{\text{out}}^{s_{i+1}} = f_{\text{score}}(y_{\text{out}}^{s_i}, \mu^{s_i}). \quad (8)$$

The output features at each scale are aggregated by summing upsampled (if necessary) feature maps,

$$y_{\text{out}} = \sum_{i=1}^{|\mathcal{S}|} \uparrow (y_{\text{out}}^{s_i}). \quad (9)$$

6. Experiments

In this section, we evaluate our approach on our EXAFace dataset and compare it with an object detector baseline and two generation-based methods. Our method outperforms all compared approaches while being significantly more efficient than those that require image generation. We also analyze our design choices and report failure cases.

6.1. Experiment setup

Detection metrics. Average precision (AP) and mean absolute error (MAE) are reported to evaluate the accuracy

Method	AP↑	AP$_t$↑	AP$_o$↑	AP$_{o+}$↑	AP$_{o-}$↑	MAE↓	MAE$_t$↓	MAE$_o$↓	MAE$_{o+}$↓	MAE$_{o-}$↓	mIoU$_o$↑	AR$_o$↑	SE$_o$↓	CE$_o$↓
Uniform	–	–	–	–	–	–	–	–	–	–	8.80	51.71	100	100
Oracle-GT	100	100	100	100	100	0.00	0.00	0.00	0.00	0.00	100	100	58.68	58.68
Oracle-YOLOH	44.79	61.70	36.34	49.83	22.85	7.55	2.07	10.65	2.54	13.60	28.63	44.56	91.96	78.74
YOLOH [23]	10.20	30.60	0.01	0.01	10^{-3}	17.37	2.78	26.11	6.87	33.11	17.23	19.01	96.90	94.01
Pix2Gestalt [15]	11.30	33.43	0.24	0.48	10^{-3}	17.38	2.83	26.10	6.63	33.18	17.75	20.25	96.54	93.31
Outpaint [17]	4.93	11.54	**1.62**	**2.47**	**0.76**	**14.69**	2.07	**21.94**	**3.48**	**28.67**	20.53	25.03	96.41	90.18
Ours	**23.07**	**66.69**	1.26	2.17	0.34	17.83	**2.01**	27.43	4.53	35.77	18.70	**27.17**	93.99	88.16

Table 2. Extreme amodal detection performance on the test set of our MS COCO-based dataset. We report the average precision (AP), the mean absolute error (MAE) of the nearest bounding box center, the mean intersection-over-union (mIoU), the average recall (AR), the self-entropy (SE), and the cross-entropy (CE). The data subsets truncated (t), outside (o), outside with evidence (o+), and outside without evidence (o-) are indicated by subscripts. The metrics that are most meaningful for assessing performance on the different data subsets are shaded. Detection metrics like AP are appropriate for evaluation of the truncated faces, since the realization of the conditional distribution (our "ground-truth") is very close to the true distribution near the image. However, further from the image, this realization no longer captures all modes of the true distribution, and so AR, CE and SE are more meaningful measures of performance in this regime.

Method	#Params $\times 10^6$	Memory (MB)	FLOPs $\times 10^9$	Latency (ms)	Throughput (s^{-1})	VRAM (MB)
YOLOH	**42.8**	**164**	**20.4**	**9.0**	**111.9**	**428**
Pix2Gestalt	3.5k	7k	452k	7.2k	0.3	31k
Outpaint	7.3k	14k	467k	7.4k	0.1	31k
Ours	67.8	259	47.6	161.6	6.2	728

Table 3. Inference efficiency on a single L40S GPU. We report the number of parameters, the memory size of the parameters, the computational cost, the latency at the 95th percentile, throughput in iterations per second, and peak VRAM usage. Generative pipelines (Pix2Gestalt and Outpaint) require orders of magnitude more parameters and FLOPs, resulting in prohibitive latency and memory consumption.

of the predicted bounding boxes. AP is given at a 25% intersection-over-union (IoU) threshold, a looser threshold than is used for the standard detection task since extreme amodal detection is considerably more challenging. MAE measures how far the predicted object centers are from the ground-truth centers, where predictions and ground-truth centers are paired using the Hungarian algorithm. We report the MAE normalized by the diagonal of the input image so that it is independent of the image resolution. Since we necessarily evaluate with respect to a realization of the ground-truth conditional distribution, these metrics are only reliable measures close to the input image, where the realization approximates the conditional distribution. Therefore, they are suitable only for evaluating truncated faces.

Localization metrics. Heatmap IoU, average recall (AR), cross-entropy (CE), and self-entropy (SE) are reported to evaluate the accuracy of the predicted heatmaps outside of the image. Since we evaluate with respect to a realization of the true distribution, AR, CE and SE are the most relevant metrics for assessing performance. That is, a prediction that has modes in addition to those of the observed sample of the ground-truth distribution should still be considered good, and this can be assessed using AR and CE. Equally, it is important to check that the prediction is not uniform by consulting the self-entropy.

Compared methods. We compare our method with three baselines/oracles and three state-of-the-art approaches. The baselines include a uniform heatmap prediction (Uniform), where the presence of a face is set to be equally likely at all locations in the expanded region; an oracle that yields the ground-truth realization (Oracle-GT); and an oracle that applies the YOLOH object detector [23] to the real extended image (Oracle-YOLOH). The compared methods include YOLOH [23], given a black-padded input image the size of the required output; Pix2Gestalt [15], a method that amodally completes partially occluded bodies, given the ground-truth in-image masks, resulting in an extended image that is passed to the YOLOH detector; and Outpainting, similar to Bhattacharjee et al. [2], where a diffusion model generates many samples of outpainted images, with text prompts generated by a vision–language model (VLM), which are passed to the YOLOH detector whose predictions are aggregated. The diffusion model and VLM used for this model are almost certain to have seen the extended images in our test set. Note that all methods use the same YOLOH detector that we trained on our dataset to predict bounding boxes and heatmaps of faces and bodies.

Implementation details. Our extreme amodal detector extends the pre-trained YOLOH [23] detector's feature extractor and detection head with a two-layer transformer encoder and a two-layer selective C2F transformer decoder. Transposed convolutions are used for upsampling, and the scoring network shares the same architecture as the YOLOH detection head. The expansion ratio is $K = 3$ and
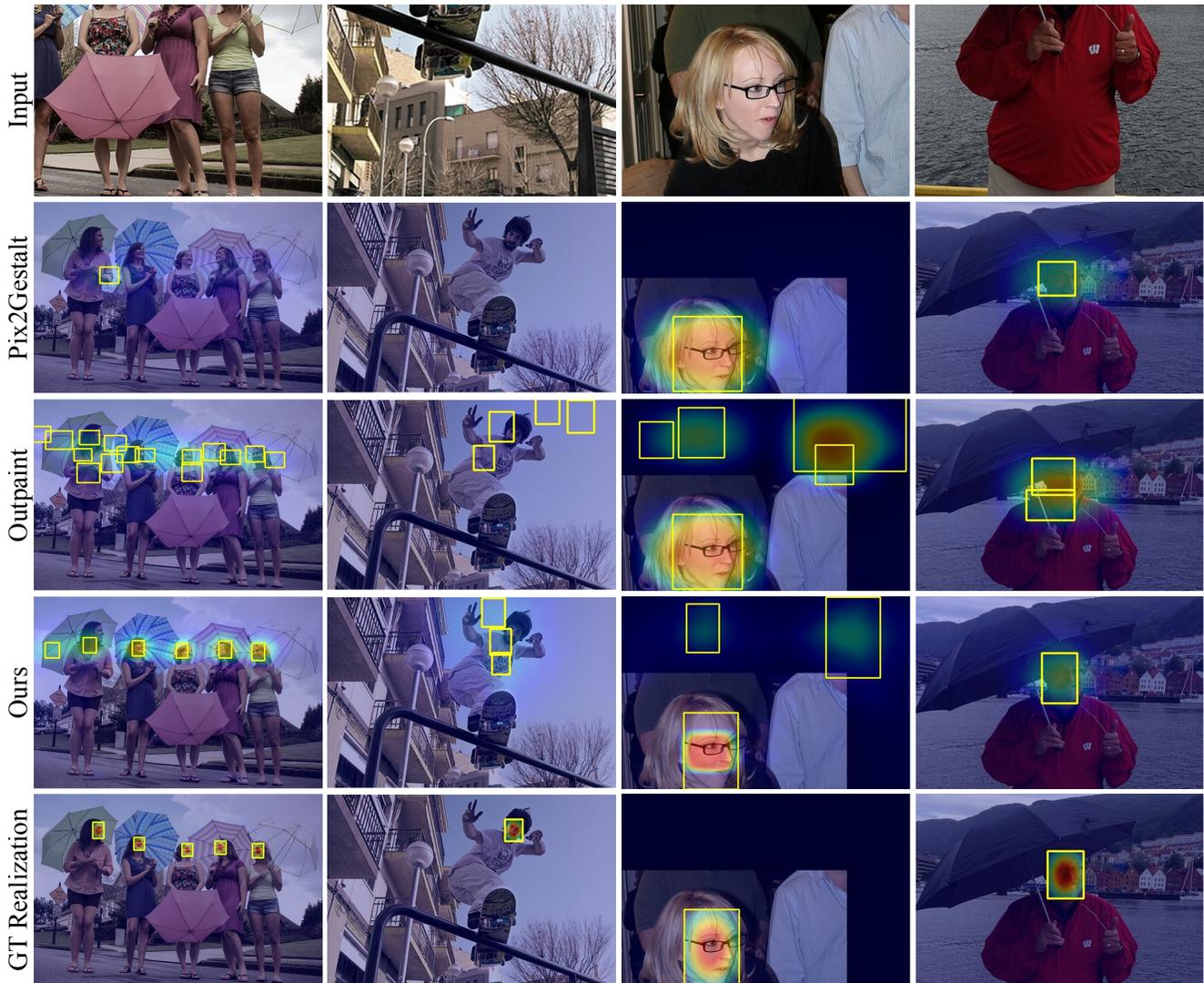
Figure 3. Qualitative results. The final row shows samples from the ground-truth conditional distributions. Our model effectively leverages contextual cues—such as nearby people (example 1), objects like a skateboard (example 2), or partial body evidence (example 4)—to infer completely unseen faces. In example 1, the model correctly extends predictions to the left, where a partial person is visible, but not to the right, demonstrating awareness of scene context and typical human height. Example 3 highlights the model's generalization to real-world scenarios. Unlike other examples where inputs are synthetically cropped from complete images, this example is naturally truncated (i.e., the faces were never captured in the original photo). Our model successfully generates plausible faces despite the lack of ground truth, demonstrating its practical utility for real-world photo expansion. Compared to our model, Pix2Gestalt struggles without large visible body parts, while the outpainting pipeline can infer outside faces but yields noisier and less consistent results.

the multi-scale refinement set is $S = (2, 1)$. Input images are resized to 320×320 and normalized, without additional augmentation. The model is optimized with AdamW, with the momentum parameter set to 0.9, weight decay set to 10^{-2}, and learning rates set to 0.024 for the transformer and detection head, and 0.004 for the YOLOH backbone. We use a warm-up scheduler [22], where the learning rate is scaled with the embedding dimension and number of warm-up steps (20% of the total). A decay factor of 0.1 is applied

after 100 steps. The model is trained for 14 epochs on four A100 GPUs with a batch size of 64. For the ablation study, we train for 8 epochs on 25% of the EXAFace dataset with a batch size of 32 on two 2080Ti GPUs.

The baseline YOLOH detector with a dilated ResNet-50 backbone and a CNN-based decoder is trained for 14 epochs on pseudo-labeled COCO [13] faces and bodies to predict bounding boxes and heatmaps. The input resolution is 320×320, random horizontal flip and random shift aug-

Method	AP$_t\uparrow$	MAE$_t\downarrow$	AR$_o\uparrow$
Ours	62.73	2.19	26.64
w/o average pooling	61.13	2.28	25.67
w/o multi-scale	61.28	2.25	25.24

Table 4. Ablation study. Here, "w/o average pooling" replaces average pooling with center sampling for downsampling the positional encodings, and "w/o multi-scale" restricts the decoder to a single scale. Both components improve performance across all three metrics: average pooling contributes more to bounding box localization (AP$_t$, MAE$_t$), while the multi-scale selective C2F mechanism yields greater gains in heatmap quality (AR$_o$).
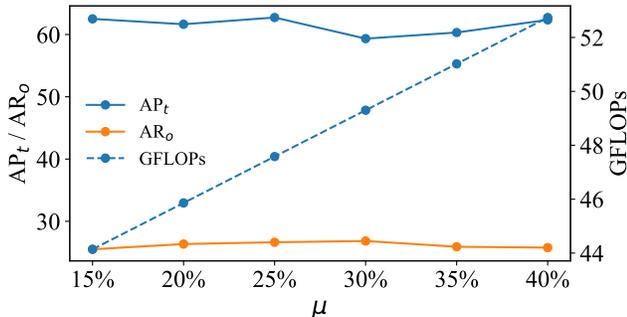

Figure 4. Sensitivity analysis of the percentage of retained tokens μ at scale $\mathcal{S} = (2)$. The metrics are relatively insensitive to μ, so we select $\mu = 25\%$, which is computationally efficient without sacrificing performance. The original data is shown in the appendix (Tab. 5).
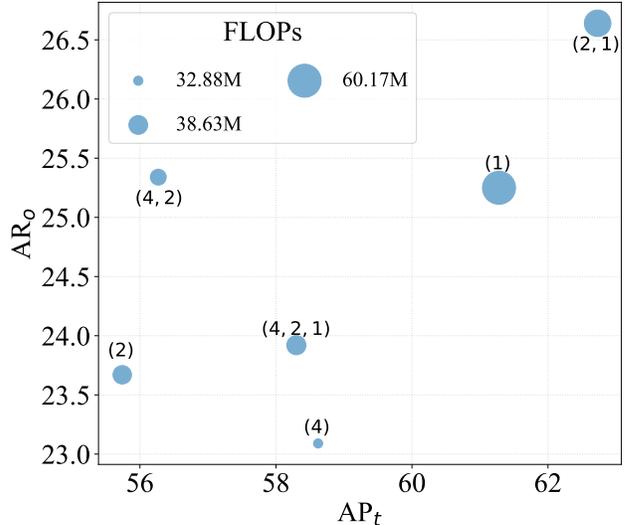

Figure 5. Analysis of multi-scale settings. We evaluate three scales $s = 1, 2, 4$ and their combinations $\mathcal{S} = (4, 2), (2, 1), (4, 2, 1)$. The results show that $\mathcal{S} = (2, 1)$ yields the highest AP$_t$ and AR$_o$, and is therefore adopted as our default setting. Original data is shown in the appendix (Tab. 6).

mentations are applied, the learning rate is 0.03, the warm-up iterations are 1200, step decays of 0.003 and 0.0003 at the 8th and 11th epochs are applied, and the batch size is 32 on two 2080Ti GPUs. For the generative baselines, we use the official Pix2Gestalt [15] checkpoint, following the gradual completion strategy of [26]. Masks touching image boundaries are iteratively extended by 10% until completion, and multiple bodies are completed sequentially and merged. For the outpainting pipeline, BLIP2 [12] generates text captions as prompts, which are fed into SDXL [17] for image extrapolation.

6.2. Results

Quantitative and qualitative results are given in Tab. 2 and Figure 3, respectively. Our model consistently outperforms all comparison methods, while also having significantly better inference efficiency than generative methods (Tab. 3). It is important to note that since we evaluate the performance on a *realization* of the ground-truth conditional distribution, AP and MAE are not suitable for measuring the detection performance outside the image, though they are appropriate for truncated faces where the true realization and true distribution overlap. For faces outside the image frame, heatmap metrics like average recall and cross-entropy are

more suitable, since they do not punish the prediction of additional modes beyond those contained in the realization, unlike the mIoU metric. This is desirable because the true conditional distribution is likely to have more modes than a realization: there are multiple possible plausible configurations. However, these metrics should be considered in parallel with self-entropy to verify that the model is not predicting a near-uniform distribution, which is also implausible. In Tab. 2, we shade the columns that are most meaningful for assessing performance on this task. Our approach exhibits a strong ability to predict face locations, whether or not there is direct visual evidence.

The outpainting pipeline also achieves strong alignment with the realized ground-truth distribution of outside faces, outperforming our approach on AP$_o$ and MAE$_o$, albeit with $10000\times$ the FLOPS. While these metrics are not suitable for measuring performance with respect to the true distribution, they should also be interpreted with some caution regardless: there is very likely information leakage, since BLIP2 [12] is trained on COCO and SDXL [17] is likely to have been trained on COCO. Therefore, the model is almost certain to have seen the extended images in our test set. A visual example of outpainting is shown in the appendix (Figure 7). In contrast, Pix2Gestalt [15] often fails to amodally complete the truncated part of the face. This is expected, since the model is trained for in-frame occluder removal, not for occlusions caused by the camera's field-of-view. A visual example of a completion by Pix2Gestalt is shown in the appendix (Figure 8). Finally, it is interest-
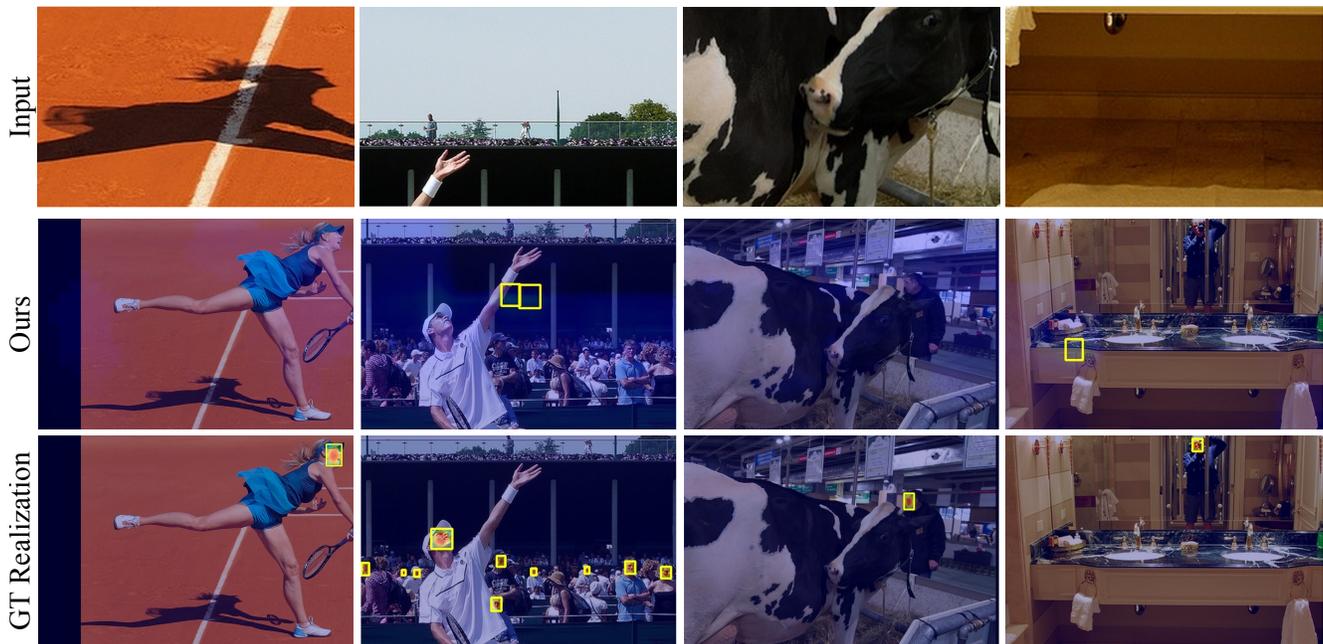
Figure 6. Failure cases. Our model struggles to predict outside faces when contextual cues are weak. In the first and second examples, strong appearance evidence is present but location cues are limited. In the third and fourth examples, no appearance evidence is available, making the presence and location of an outside face ambiguous—even for human observers.

ing that our approach outperforms the YOLOH oracle that receives the extended image for truncated faces. This is attributable to the input resolution: both methods process a 320×320 image, but the resolution of the cropped region is effectively higher for our approach.

6.3. Ablation study and analysis

In Tab. 4, we ablate two design choice: the positional encoding downsampling strategy of average pooling is replaced with center sampling, and the multi-scale decoding strategy is replaced by a single scale. The results indicate that replacing either of these design choices with simpler approaches leads to significantly poorer performance.

Figure 5 presents the analysis of different multi-scale strategies. Among the explored settings, the $(2, 1)$ configuration achieves the best overall performance, and is therefore adopted as our default. Figure 4 shows the effect of varying μ, where it is clear that the metrics are relatively insensitive to this hyperparameter choice. This confirms that our selection mechanism is computationally advantageous without sacrificing accuracy.

6.4. Limitations and discussion

Several failure cases of our method are shown in Figure 6. This highlights one limitation of our approach, that it struggles when the contextual cues are weak, such as a person's shadow but no body. This may stem from insufficient training data to capture such rare examples, or from the inherent ambiguity in these scenarios. Another limitation is that our approach predicts the conditional distribution of a face outside the image, but cannot be used to sample multiple co-occurring faces. In contrast, the outpainting method samples co-occurring faces and so retains these useful correlations. This may limit the use of our approach in some downstream applications, where we may wish to know about the plausible configurations of multiple objects. A final limitation is that we have only considered the class of human faces. However, our approach is not tailored specifically to faces, and should easily extend to other classes.

7. Conclusion

In this paper, we proposed extreme amodal face detection, a new task that requires the model to detect and localize faces that are outside the image or truncated by the image frame. We construct the new EXAFace dataset for training and evaluating models on this task and propose a heatmap-based extreme amodal object detector with a novel selective coarse-to-fine decoder. The results indicate that our approach outperforms other related methods, while requiring orders of magnitude less compute and memory. This work points to the feasibility of efficiently inferring the presence of unseen objects, with possible applications in, for example, safer robot navigation, active surveillance, and realistic image expansion.

References

[1] Jiayang Ao, Yanbei Jiang, Qiuhong Ke, and Krista A Ehinger. Open-world amodal appearance completion. In *CVPR*, pages 6490–6499, 2025. 2

[2] Subhransu S Bhattacharjee, Dylan Campbell, and Rahul Shome. Believing is seeing: Unobserved object detection using generative models. In *CVPR*, pages 19366–19377, 2025. 2, 3, 5

[3] MacDonald Cheyenne. A food delivery robot's footage led to a criminal conviction in la, 2023. Retrieved October 29, 2023, from Engadget website: https://www.engadget.com/a-food-delivery-robots-footage-led-to-a-criminal-conviction-in-la-190854339.html. 2

[4] Tai-Yin Chiu, Yinan Zhao, and Danna Gurari. Assessing image quality issues for real-world problems. In *CVPR*, pages 3646–3656, 2020. 1

[5] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. RetinaFace: single-shot multi-level face localisation in the wild. In *CVPR*, 2020. 3

[6] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, page 1322–1333, New York, NY, USA, 2015. Association for Computing Machinery. 2

[7] Dave Gershgorn. Nothing pixelated will stay safe on the internet, 2016. Retrieved October 29, 2023, from Quartz website: https://qz.com/779625/none-of-your-pixelated-or-blurred-information-will-stay-safe-on-the-internet. 2

[8] Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *ECCV*, pages 289–305. Springer, 2024. 3, 4

[9] Cheng-Yen Hsieh, Kaihua Chen, Achal Dave, Tarasha Khurana, and Deva Ramanan. TAO-Amodal: a benchmark for tracking any object amodally. *arXiv preprint arXiv:2312.12433*, 2023. 2

[10] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *ECCV*, pages 18–35. Springer, 2024. 2

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *CVPR*, pages 4015–4026, 2023. 2

[12] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. pages 19730–19742. PMLR, 2023. 7, 13

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 2, 3, 6

[14] Verma Mimansa. Amazon was fined \$30 million for enabling ring workers to spy on people and keeping kids' alexa records, 2023. Retrieved October 29, 2023, from Yahoo Finance website: https://tech.yahoo.com/business/articles/amazon-fined-30-million-enabling-095400450.html. 2

[15] Ege Ozguroglu, Ruoshi Liu, Dídac Surís, Dian Chen, Achal Dave, Pavel Tokmakov, and Carl Vondrick. Pix2Gestalt: amodal segmentation by synthesizing wholes. In *CVPR*, pages 3931–3940, 2024. 2, 5, 7, 14

[16] Chiara Plizzari, Shubham Goel, Toby Perrett, Jacob Chalk, Angjoo Kanazawa, and Dima Damen. Spatial cognition from egocentric video: Out of sight, not out of mind. In *2025 International Conference on 3D Vision (3DV)*, pages 1211–1221, 2025. 2

[17] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024. 5, 7, 13

[18] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded SAM: assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 2

[19] Hernandez Salvador. A home security tech hacked into cameras to watch people undressing and having sex, prosecutors say, 2021. Retrieved October 29, 2023, from BuzzFeed News website: https://www.buzzfeednews.com/article/salvadorhernandez/home-security-camera-hacked-adt. 2

[20] Das Shanti. Nhs data breach: trusts shared patient details with facebook without consent, 2023. The Observer. Retrieved from https://www.theguardian.com/society/2023/may/27/nhs-data-breach-trusts-shared-patient-details-with-facebook-meta-without-consent. 2

[21] Basile Van Hoorick, Pavel Tokmakov, Simon Stent, Jie Li, and Carl Vondrick. Tracking through containers and occluders in the wild. In *CVPR*, pages 13802–13812, 2023. 2

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 6

[23] Shaobo Wang, Renhai Chen, Hongyue Wu, Xiaozhe Li, and Zhiyong Feng. YOLOH: you only look one hourglass for real-time object detection. *IEEE TIP*, 33:2104–2115, 2024. 5

[24] Davis Wes. A woman and her daughter plead guilty to abortion-related charges supported by meta-provided facebook chats, 2023. Retrieved October 29, 2023, from Verge website: https://www.theverge.com/2023/7/11/23790923/facebook-meta-woman-daughter-guilty-abortion-nebraska-messenger-encryption-privacy. 2

[25] Contributors Wikipedia. Facebook–cambridge analytica data scandal, 2019. Retrieved October 29, 2023, from

Wikipedia website: `https://en.wikipedia.org/wiki/Facebook-Cambridge_Analytica_data_scandal`. 2

[26] Katherine Xu, Lingzhi Zhang, and Jianbo Shi. Amodal completion via progressive mixed context diffusion. In *CVPR*, pages 9099–9109, 2024. 2, 7

[27] Jinze Yang, Haoran Wang, Zining Zhu, Chenglong Liu, Meng Wu, and Mingming Sun. VIP: versatile image outpainting empowered by multimodal large language model. In *ACCV*, pages 1082–1099, 2024. 3

[28] Shaofeng Zhang, Jinfa Huang, Qiang Zhou, zhibin wang, Fan Wang, Jiebo Luo, and Junchi Yan. Continuous-multiple image outpainting in one-step via positional query and a diffusion-based approach. In *ICLR*, 2024. 3

[29] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 11

[30] Chaoyang Zhu and Long Chen. A survey on open-vocabulary detection and segmentation: Past, present, and future. *IEEE TPAMI*, 46(12):8954–8975, 2024. 1

[31] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276, 2023. 1