MAC: ModAlity Calibration for Object Detection

Anonymous Author(s)
Submission Id: 1242*

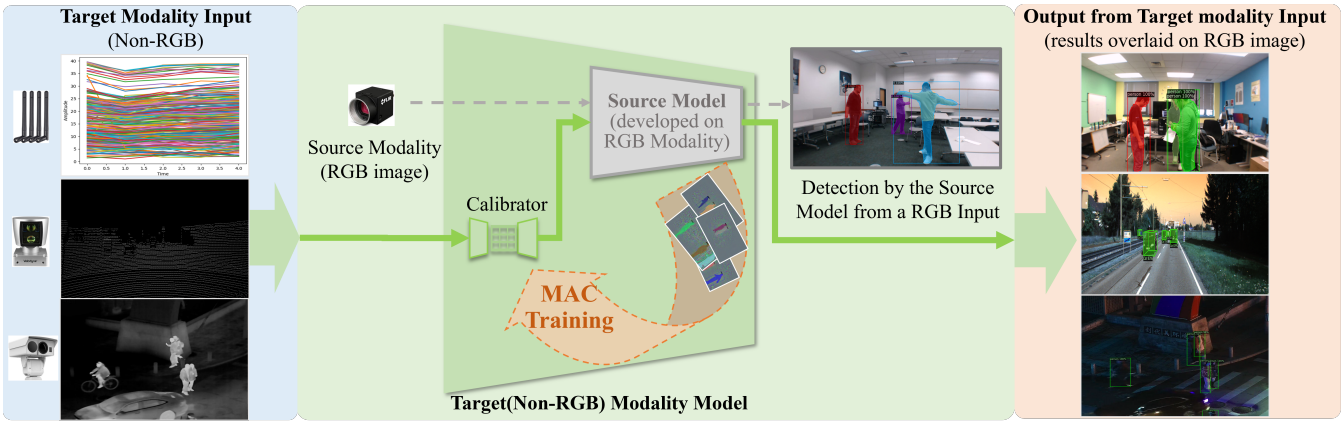

Figure 1: We propose ModAlity Calibration (MAC) for switching input modalities of DNN models. Our "Target Modality Model" is constructed and trained upon a pre-trained source-modality (RGB-input) model by "MAC Training". The dashed arrows "⇢" are for training only. In inference, each target-modality(Non-RGB) input is processed along the thick solid green arrows "→".

ABSTRACT

The flourishing success of Deep Neural Networks(DNNs) on RGB-input perception tasks has opened unbounded possibilities for non-RGB-input perception tasks, such as object detection from wireless signals, lidar scans, and infrared images. Compared to the matured development pipeline of RGB-input (source modality) models, developing non-RGB-input (target-modality) models from scratch poses excessive challenges in the modality-specific network design/training tricks and labor in the target-modality annotation. In this paper, we propose ModAlity Calibration (MAC), an efficient pipeline for calibrating target-modality inputs to the DNN object detection models developed on the RGB (source) modality. We compose a target-modality-input model by adding a small calibrator module ahead of a source-modality model and introduce MAC training techniques to impose dense supervision on the calibrator. By leveraging (1) prior knowledge synthesized from the source-modality model and (2) paired {target, source} data with **zero** manual annotations, our target-modality models reach comparable or better metrics than baseline models that require 100% manual annotations. We demonstrate the effectiveness of MAC by composing the WiFi-input, Lidar-input, and Thermal-Infrared-input models upon the pre-trained RGB-input models respectively.

ACM MM23, 2023, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**.

KEYWORDS

modality calibration, object detection, model inversion

ACM Reference Format:
Anonymous Author(s). 2018. MAC: ModAlity Calibration for Object Detection. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM23).* ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Although research on the DNN-based perception problem has been largely focused on RGB-input models, there are wide scenarios in that non-RGB sensors have clear advantages over RGB cameras. For instance, wireless signals [61, 71] can easily penetrate furniture occlusion and identify human bodies for their Dielectric properties, while being lighting-free, occlusion-resistant, and privacy-friendly compared to cameras. Lidar scans [14, 57] contain depth information, enabling more accurate and robust object localization than RGB under low light or bad weather. Thermal InfRared(TIR) cameras [26, 27] capture near-infrared (0.75-1.3μm) or long-wavelength infrared (7.5-13μm) signals, which makes, in particular, human bodies more visible than in RGB images and more robust to the visible spectrum interference.

Thanks to the decade of work on RGB-input DNN models, researchers has accumulated extensive resources on image-appeal architectures(e.g., MaskRCNN [19], Yolo5 [28], Swin-Transformer [42]), pre-train-datasets (ImageNet [6], MS-COCO [38], OpenImages [33]), pre-train weights, training tricks [15, 20, 70] and code repos(e.g., Detectron2 [65], mmDetection[4]). Unfortunately, non-RGB-input

models cannot be **directly** built upon above RGB resources. Instead, one usually needs new DNN designs[50, 61, 71], from-scratch training, and new data collection/annotation of non-RGB sensors at the comparable scale of RGB databases above.

In this paper, we propose ModAlity Calibration (MAC) for calibrating target-modality inputs to a DNN model developed on the source modality. Figure 1 shows the main idea of MAC. Ahead of a source-modality model, we add a small target-modality-input calibrator module, composing a **[Calibrator|Source]**-structured target-modality model. The calibrator transforms a target-modality input into a source-modality-like tensor highlighting the foreground, which is then mapped to object detection results by the source module. Trained on {source, target} input pairs of **zero** manual annotation, the MAC target-modality models reach comparable or better metrics on WiFi, Lidar, and Thermal Infrared than the baselines that require 100% manual annotation. This is achieved by our MAC training techniques that learn prior knowledge from the enclosed source module and iteratively regularize gradients on the calibrator layers. MAC training helps the detection task by mimicking the foreground features of the source modality inputs.

Summary of Contributions:

- MAC, A simple pipeline, is proposed for building Non-RGB-input models upon pre-trained RGB-input models in reducing the DNN design efforts and training data.
- MAC training techniques are proposed to address the special vanishing gradient problem arising by adding a calibrator to a pre-trained RGB perception model. MAC training introduces strong and dense gradients that significantly improve the target-modality model metrics and reduce the need for annotations.
- Compared with target-modality-input models, (i.e., WiFi, Lidar, Thermal Infrared) under naive training, the MAC training techniques achieve comparable or better metrics without manual annotation (MAC-self-supervised), and significantly better metrics with manual annotation(MAC-supervised).

2 REALATED WORKS

For conciseness, we only list object detection work on the three Non-RGB target modalities (WiFi, Lidar, and Thermal) involved in our experiments.

Modality-specific Perception. In most two-stage approaches, Non-RGB inputs are converted to RGB images before feeding to an RGB-input model. Researchers of [10, 30, 31] only convert Wifi signals to low-resolution ($< 160\times120$) RGB images by over-fitting a few antenna layouts. No codes or data are available. Points2Pix [44] translates Lidar points to RGB images with a conditional GAN. The infrared images were re-colored to RGB in [37, 52] by image translation [29].

In single-stage approaches, the whole model is only designed and trained on a non-RGB modality. Due to the lack of spatial representation, the WiFi-input models are mostly focused on coarse-granularity tasks such as crowd counting [7, 40] or single-person activity recognition [35, 64]. [61] develop pioneer WiFi-specific DNNs for multi-person segmentation and pose estimation. Lidar-input models are specific to point-clouds representations, such as the Point View [49, 50, 55], the Bird's Eye View [34, 67] and the Range View([12, 43]). The range view is popular for its low quantization error and computational costs. Thermal images are close to RGB spatially, enabling [26, 27] to train RGB-input models on the TIR inputs.

Unlike existing two-stage approaches, we work on perception tasks that learn foreground representation instead of RGB appearances. By enclosing the pre-train source model in the target model, we simplify the design effort and reduce target-modality annotations of the single-stage approaches.

Domain Adaption. By definition[1], domain adaptation mainly addresses the shift of data sampling distribution. Modality adaption, on the other hand, mainly addresses the change in the spatial, temporal, and physical nature of the inputs. Nevertheless, some works still consider modality adaption as a special case of domain adaption especially in the case of image-to-image translation [9, 13, 45, 46, 48, 66]. All these works aim to generate realistic image pixels in new domains.

Our work applies to any non-image modalities such as WiFi signals. Instead of pursuing realistic pixels, we ONLY generate foreground-associated features contributing to the detection tasks, which aims to reduce the efforts in DNN design and data annotation. Moreover, for simplicity, this work is presented under the assumption that the source and target modality data are sampled from the same foreground/background distributions, such that MAC is only focused on reducing the discrepancy between modalities.

Knowledge Distillation between Models. Teacher-to-student Knowledge Distillation (KD) is a popular approach [8, 21]. The student models mimic the teacher models in predictive probabilities [21, 36], intermediate features [54, 63], or attention maps [3, 41, 60, 68]. When the teacher and student models have different input modalities [17], KD requires the same amount of annotated source-target data as those in the teacher training. All KD methods run the teacher and student in parallel during training.

Unlike KD, our [Calibrator|Source] target model is initialized and supervised by the enclosed Source module, which requires neither an independent teacher inference dataflow nor fully annotated target-modality data. Ablation study shows that MAC clearly performs better.

Adversarial Training. To improve robustness on imbalanced datasets, many adversarial training strategies explicitly produce hard features/samples: auto-augmentation [74], co-mixup [32], random erasing [72], representation self-challenging [23, 24], reverse attention [5]. Regulators such as cross-layer consistency [22, 62] and self-distillation mechanism [25] were also very effective. Generative Adversarial Networks(GAN) implicitly produce hard samples from a discriminator and are recently extended to the object detection tasks by [39, 51].

We improve target model robustness by synthesizing image-like foreground representation and regularising gradients of the enclosed source model.

Vector Quantized Representation. VQ-VAE [53, 58] and VQ-GAN [11] show that a quantized latent space provides a compact

[1] https://en.wikipedia.org/wiki/Domain_adaptation

representation of natural images, language, and audio/video sequence while using a relatively small number of parameters, making them efficient to train and use.

We extend VQVAE to learn the foreground representation shared between modalities.

3 MODALITY CALIBRATION (MAC)

Problem Definition: MAC for the object detection task:

- **Source model**: $S(\cdot) : \mathbf{I} \rightarrow \mathbf{Y}$ maps one RGB image $\mathbf{I} \in \mathfrak{R}^{[width \times height \times 3]}$ to the object locations and categories $\mathbf{Y} = [object\_bboxes, object\_mask, object\_class]$.
- **Target model**: $T(\cdot) : \mathbf{X} \rightarrow \mathbf{Y}$ maps one target modality tensor $\mathbf{X} \in \mathfrak{R}^{[width\_T \times height\_T \times channel\_T]}$ to $\mathbf{Y}$.
- **MAC target model**: $T(\cdot) : \{C(\cdot)|S(\cdot)\}$, where the "Calibrator" module $C(\cdot) : \mathbf{X} \rightarrow \mathbf{J}$ produces an image-like tensor $\mathbf{J} \in \mathfrak{R}^{[width \times height \times 3]}$. The "Source" module $S(\cdot)$ maps $\mathbf{J}$ to $\mathbf{Y}$.

The goal of MAC is to train a MAC target model $\{C(\cdot)|S(\cdot)\}$ given a pre-trained source model $S(\cdot)$ and a set of $\{\mathbf{X}, \mathbf{I}\}$ pairs. (See the framework in Figure 2).

For simplicity, we assume that the source and target modality data are sampled from the same $\mathbf{Y}$ (foreground/background) distributions, such that **MAC is *only* focused on reducing the discrepancy between modalities.**

3.1 Reasoning for the $\{C(\cdot)|S(\cdot)\}$ Target Model

If we follow the development procedure of $S(\cdot)$ to develop a new $T(\cdot)$, we need a modality-specific multi-resolution feature extractor (comparable to ResNet), which is coupled with a task-specific output head (Such as the anchor-based or transformer-based bounding box regressors) by multi-resolution skip connections. One also needs annotated $\{\mathbf{X}, \mathbf{Y}\}$ data comparable to the amount of the $\{\mathbf{I}, \mathbf{Y}\}$ data used in the source model $S(\cdot)$ training.

Under MAC(Figure 2), we only design calibrator $C(\cdot)$ that produces a *single-resolution* tensor $\mathbf{J}$ feeding to the source module $S(\cdot)$ enclosed in $T(\cdot)$.

Foreground encoding in $C(\cdot)$: Since it is $S(\cdot)$'s expertise to locate and classify objects, $C(\cdot)$ only needs to pass to $S(\cdot)$ some foreground-sensitive features $\mathbf{J}$, i.e., the edges/textures highlighting all the object categories. To generate such foreground features, we revisit the insight of the image-wise Class Activation Maps [73]: all pixels of each object category can be mapped to an element of a probability vector by Softmax activation, which highlights foreground pixels assuming category-wise features follow a multi-modal distribution. In order to preserve the internal spatial layout of objects, we encode pixel patches(object parts) under the multi-modal distribution. This is done by a encoder-decoder structure $C(\cdot) : \{E_T, D_T\}$ with a quantized latent space inspired by VQ-VAE[58] (see Figure 2). We set the latent space tensor size to $[width/8, height/8, channel]$, in which every $[1 \times 1 \times channel]$ vector is hard-coded to one of the multi-modal centers of local patches, denoted as codebook $\{B_i \in \mathfrak{R}^{1\times1\times channel}\}|_{i=1,...,p}$. Mapping such a VQ-VAE-like latent space to image-like feature $\mathbf{J}$, the calibrator decoder D_T simply takes the same structure as the standard VQ-VAE decoder. The calibrator encoders E_T for target-modality inputs are similar to the standard VQ-VAE encoder with minor adjustments below.

Target-Modality Encoder E_T: The WiFi signal corresponding to one synchronized RGB image [61], is represented as the Channel State Information (CSI) [18] tensor [samples, transmitters, receivers, sub-carriers]. The CSI tensor elements has no spatial dependence as RGB pixels. In this case, we construct E_T by adding Wi2Vi [31] layers in front of a VQ-VAE encoder. For other target modalities (infrared and Lidar range images) that have the image-like tensor shape, we directly use the VQ-VAE encoder structure for E_T.

Remarks: To enable $T(\cdot) : \{C(\cdot), S(\cdot)\}$ to produce the same $\mathbf{Y}$ as that of a pre-trained source model $S(\cdot)$, the latent space of $C(\cdot)$ has to contain the same semantics as those encoded in $S(\cdot)$. This opens the possibility to learn $C(\cdot)$ from $S(\cdot)$, without requiring a huge set of $\{\mathbf{X}, \mathbf{Y}\}$ training data.

3.2 MAC Training

The devil lies in the training of the MAC target model $T(\cdot)$. There are two *naïve* sources of supervision: **(1)** Given the $\{\mathbf{X}, \mathbf{I}\}$ pairs, one may pre-train $C(\cdot)$ to approximate $\mathbf{I}$. Since there are usually more background pixels than foreground pixels, the pre-trained $C(\cdot)$ outputs $\mathbf{J}$ may largely be background textures that are irrelevant to the detection tasks. **(2)** Given abundant $\{\mathbf{X}, \mathbf{Y}\}$ training pairs, one may randomly initialize $C(\cdot)$ and update all $T(\cdot)$ layers using the gradients back-propagated from the source model losses. Such a strategy suffers from the vanishing gradient problem [1, 16] and is contradictory to the common DNN training practice (for instance, ImageNet-pretrained Resnet + randomly initialized Mask-RCNN). In fact, the randomly initialized $C(\cdot)$ should receive the strong gradients for updating, while the pre-trained weights of $S(\cdot)$ should be preserved by updating with weak gradients. Figure 3(a) shows that the *naïve* training produces neither foreground-sensitive features nor any clear gradients at $\mathbf{J}$ comparable to the high-level gradients, e.g., "Avg. Gradients of Res50-p4to6".

To address the above issues, we propose the MAC training techniques below (Also see diagram in Figure 2).

Road-map: MAC training techniques address the special vanishing gradient problem arising by adding a calibrator to a pre-trained RGB perception model. Source Model Inversion(SMI) discovers the dense image-like foreground features. Foreground Semantics Reconstruction(FSR) introduces dense image-like supervision by pre-training the calibrator. Decayed Semantic Supervision (DSS) uses $\mathbf{J}_T$ to provide strong gradients for foreground supervision, then let the weaker gradients from $\mathcal{L}_S$ amend the acute details. Skipped Inverted Attention (SIA) channels the high-level gradient-based attention from §(·) back to the low-level feature layers of (·). Along with section 3.2, Fig. 3 and Table 4 demonstrate how the gradients are improved by each MAC training technique.

3.2.1 Source Model Inversion(SMI). We first guide the calibrator $C(\cdot)$ to produce foreground-sensitive features leveraging a pre-trained source model $S(\cdot)$. Given the pre-trained $S(\cdot)$ and the object annotation $\mathbf{Y}$, we conduct model inversion [2, 59] to generate $\mathbf{J}_S$ by minimizing the $S(\cdot)$ losses,

$$\mathcal{L}_S(S(\mathbf{J}_S), \mathbf{Y}) = \lambda_{bbox}\mathcal{L}_{bbox} + \lambda_{cls}\mathcal{L}_{cls} + \lambda_{mask}\mathcal{L}_{mask}. \quad (1)$$

This problem is solved by image-wise optimization: freezing all the $S(\cdot)$ layers, computing gradient from $\mathcal{L}_S(\mathbf{J}_S, \mathbf{Y})$ and only updating $\mathbf{J}_S$ from its random initialization. After the optimization converges,

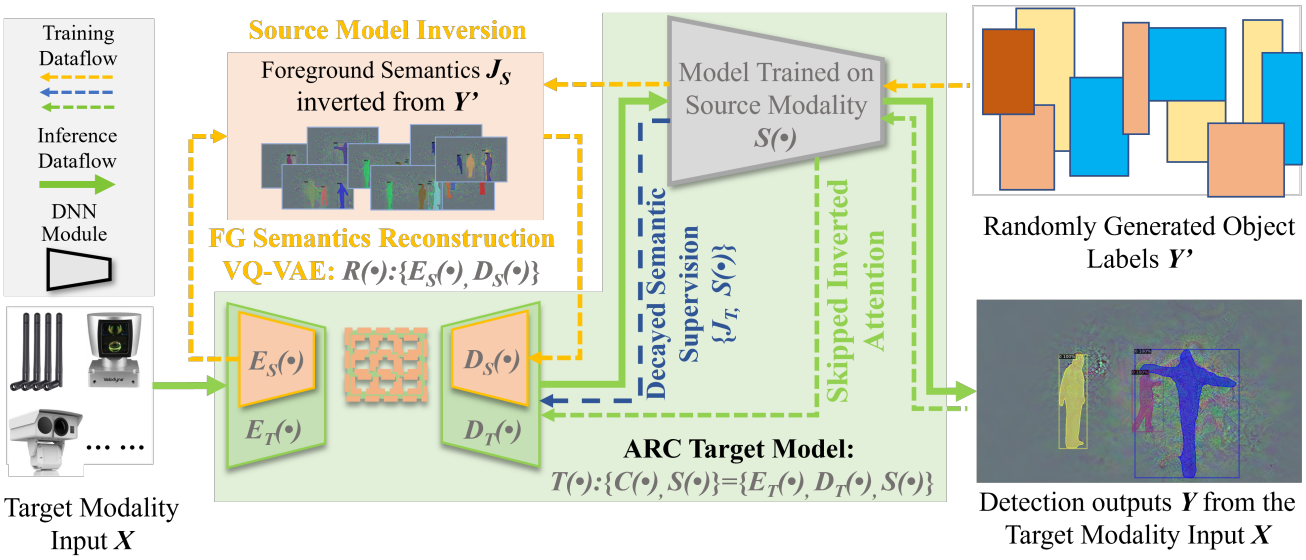

Figure 2: ModAlity Calibration(MAC) Framework. Our MAC target model $T(\cdot) : \{E_T, D_T, S\}$ is composed of calibrator $C(\cdot) : \{E_T, D_T\}$ and source model $S(\cdot)$, with its inference dataflow along the solid green arrows ("⟶"). The MAC training includes three types of supervision(along the dashed arrows "⇠⇢") : (a) Foreground(FG) Semantic Reconstruction(FSR), which leverages the source model prior: the foreground(object) semantics $\mathbf{J}_S$ synthesized from $S(\cdot)$ and self-reconstructed by an auxiliary VQ-VAE, $R(\cdot) : \{E_S, D_T\}$. (b) Decayed Semantic Supervision(DSS), which regularizes the supervision from precomputed foreground semantics $\mathbf{J}_T$ on target modality training data and from source model $S(\cdot)$. (c) Skipped Inverted Attention(SIA), which improves the attention of $C(\cdot)$ output using high-level $S(\cdot)$ layer gradients. See details in the MAC Training section.

$\mathbf{J}_S$ is synthesized as a most probable and style-invariant "image" from which $S(\cdot)$ can detect $\mathbf{Y}$. $\mathbf{J}_S$ only captures the edge-like patterns of the foreground (See pixels marked in colors Figure 3(b)). We call $\mathbf{J}_S$ the **Foreground Semantics**.

Compared with the original RGB pixels, we see $\mathbf{J}_S$ a clean foreground-focused supervision to guide $C(\cdot)$ training. Compared with the small gradient amplitude from $\mathcal{L}_S$, the amplitude of $\mathbf{J}_S$ is comparable to $\mathbf{I}$ resulting in strong gradients on the $C(\cdot)$ layers.

To increase the diversity of $\mathbf{J}_S$, we also randomly generate the object layouts in $\mathbf{Y}'$ of different bbox locations (instance masks for Mask-RCNN) and class labels. The random object labels $\mathbf{Y}$ introduce diverse object locations, sizes, and co-occurrences. In addition, the noise diversity of the foreground is also introduced by the random $\mathbf{J}$ initialization of model inversion from different $\mathbf{Y}'$s.

3.2.2 Foreground Semantics Reconstruction(FSR). Next, to inject the prior knowledge in J_S to calibrator $C(\cdot)$, we train an auxiliary VQ-VAE, $R(\cdot) = \{E_S, D_S\} : \mathbf{J}_S \rightarrow \mathbf{J}_S$ that encodes and reconstructs J_S. Then we share the $R(\cdot)$'s VQ codebook with $C(\cdot)$, and initialize the $C(\cdot)$'s decoder $D_T(\cdot)$ by the $R(\cdot)$'s decoder $D_S(\cdot)$ weights. We call the initialization method of $C(\cdot)$ as **Foreground Semantics Reconstruction(FSR)**.

In addition, our target model $T(\cdot)$ encloses $S(\cdot)$ as a module, which can explicitly inherit the source model knowledge by initializing it with the pre-trained source model weights.

Finally, to accommodate both above priors incorporated into $C(\cdot)$ and $S(\cdot)$, we train $T(\cdot)$ in a **two-stage update** strategy: (i) fix $S(\cdot)$ and only update $C(\cdot)$ until it converges; (ii) continue training by updating both $C(\cdot)$ and $S(\cdot)$.

Remarks: (i) $\mathbf{J}_S$ is synthesized from a pre-trained source model over synthetic $\mathbf{Y}'$. No manually annotated source inputs $\mathbf{I}$ are needed. **(ii)** Given that the $R(\cdot)$ training requires no annotation on $\mathbf{X}$, and the $S(\cdot)$ module can provide pseudo ground truth $\mathbf{Y}_{pseudo}$ for the $\{\mathbf{X}, \mathbf{I}\}$ pairs, the overall training of $T(\cdot)$ is **self-supervised** (**zero** manual annotation on $\mathbf{X}$).

3.2.3 Decayed Semantic Supervision (DSS). The Decayed Semantic Supervision is used to regularize $\mathcal{L}_S$ gradients with image-space semantic supervision. Given a pre-trained $S(\cdot)$ and either $\{\mathbf{X}, \mathbf{Y}_{pseudo}\}$ (the self-supervised MAC) or $\{\mathbf{X}, \mathbf{Y}\}$ (the supervised MAC) as the target model training data, we invert $S(\cdot)$ to generated $\mathbf{J}_T$ as GT to directly supervise $C(\cdot)$. The $C(\cdot)$ output $\mathbf{J}$ is an image-like tensor, therefore $C(\cdot)$ can be trained with image-based losses against J_T along with the source loss $\mathcal{L}_S$, leading to the Semantic Supervision (**SS**) loss,

$$\mathcal{L}_{SS}(\mathbf{J}, \mathbf{J}_T) = SSIM(\mathbf{J}, \mathbf{J}_T) + L_1(\mathbf{J}, \mathbf{J}_T) + \mathcal{L}_S, \quad (2)$$

where $SSIM(\cdot, \cdot)$ is the structural similarity loss and $L_1(\cdot, \cdot)$ is the L1-norm loss (Both losses provide **higher amplitudes gradients** than gradients back-propagated from $\mathcal{L}_S$). Due to the image-wise optimization nature of model inversion, each $\mathbf{J}_T$ over-fit different noise of the source model $S(\cdot)$. When training on all $\mathbf{J}_T$ samples, $C(\cdot)$ tends to produce averaged foreground semantics smoothing out sample-specific details that may be critical to detection.

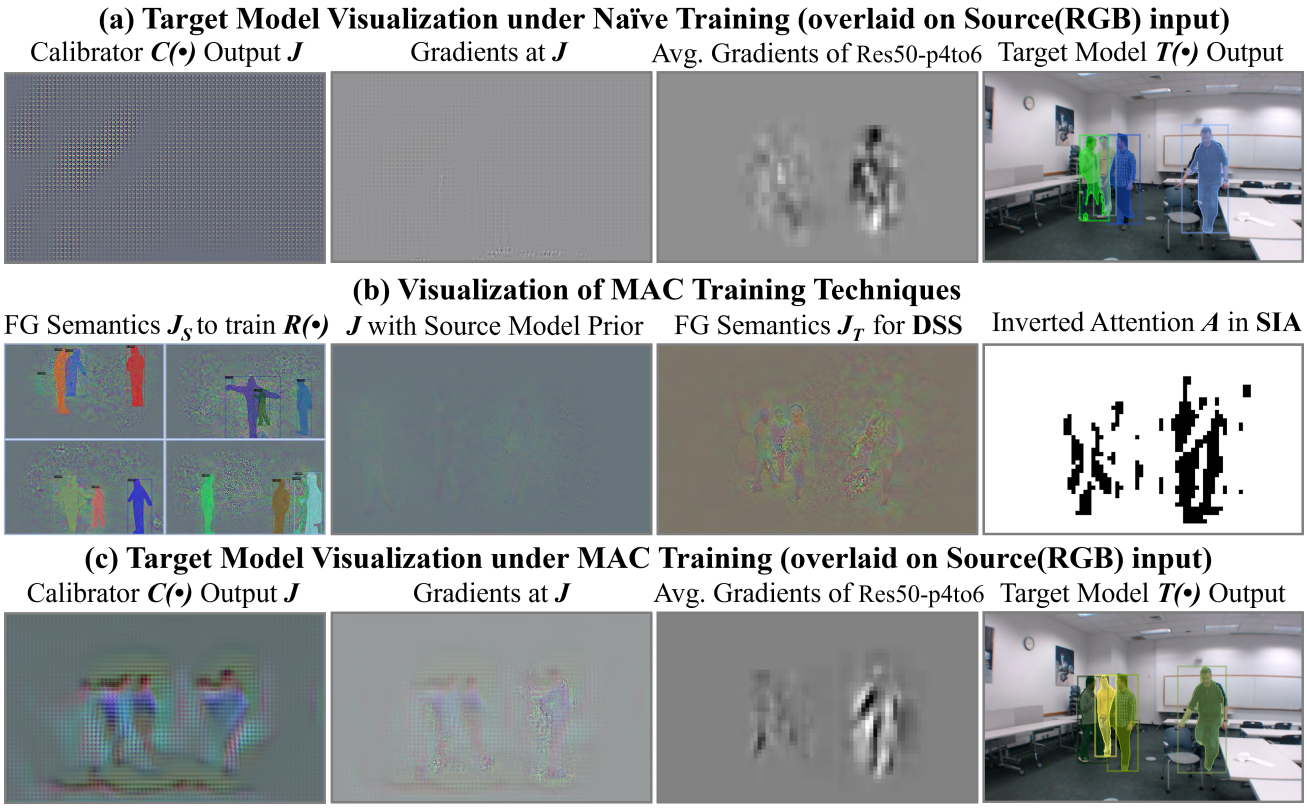

Figure 3: Training strategy examples on a WiFi-input target model $T(\cdot) : \{C(\cdot), S(\cdot)\}$. Here, the source model $S(\cdot)$ is an RGB-input ResNet50-FPN-Mask-RCNN. (a) Target model visualization under *Naïve* training shows neither foreground-sensitive features in $C(\cdot)$ output J nor any clear gradients at J compared to "Avg. Gradients of Res50-p4to6" (the high-level gradients visualized by resizing and averaging on one channel). (b) Visualization of MAC training techniques (described the MAC Training section). (c) Target model visualization under MAC training shows clear foreground-sensitive features, strong gradients at J and better detection. Note that, in "Target Model T(·) Output" of the 3rd row, all persons are detected with better masks than "Target Model T(·) Output" of the 1st row. Please zoom in to see all the instance masks.

We propose a simple fix, called Decayed Semantic Supervision(**DSS**), to such a problem:

$$\mathcal{L}_{DSS} = \lambda_{DSS}(SSIM(\mathbf{J}, \mathbf{J}_T) + L_1(\mathbf{J}, \mathbf{J}_T)) + \mathcal{L}_S, \tag{3}$$

where λ_{DSS} is a scalar that continuously decays with the increase of iterations (see supplementary materials).

How DSS works? The foreground semantics $\mathbf{J}_T$ only contain relatively clean foreground features reconstructed from Y, but the background features are random due to the image-wise optimization of SMI. When training over all the data, using $\mathbf{J}_T$ throughout all training iterations will contaminate the target model with random background noise. By smoothly decaying lambda_DSS, $\mathbf{J}_T$ provides strong gradients/supervision on foreground features in the early iterations, then the gradients from the source model losses $\mathcal{L}_S$ amend the acute overall details in later iterations. The effectiveness of **DSS** is qualitatively shown in Figure 3(b) and quantitatively evaluated in Table 4.

3.2.4 Skipped Inverted Attention (SIA). The Skipped Inverted Attention (SIA) is applied to amplify and balance the gradients from $\mathcal{L}_S$. The strong gradients at the high-level layers of $S(\cdot)$ (see "Avg. Gradients of Res50-p4to6 " in Figure 3(a) of a ResNet-FPN-MaskRCNN $S(\cdot)$ module) does not propagate into strong "Gradients at J".

To address this issue, we generate a 2D inverted attention mask **A** from the above gradients **G** and skip the earlier ResNet layers backward to supervise $C(\cdot)$. Formally, from $\mathbf{G} \in \Re^{width \times height \times channel}$, $\mathbf{A}(\mathbf{G}) \in \Re^{width \times height}$ is computed as,

$$\mathbf{a}(i) = \begin{cases} 0, & \text{if } \sum_{j=1}^{channel} g(i,j) \geq q \\ 1, & \text{otherwise} \end{cases} \tag{4}$$

where g_p is a scalar of the $(100-p)^{\text{th}}$ percentile of $\sum_{j=1}^{channel} \mathbf{G}(:, j)$, $(i \in [1, \cdots, width \times height])$. Low **G** and high **A** values denote under-represented regions by $S(\cdot)$ marked by white pixels in Figure 3(b)"Inverted Attention A in SIA". Forwarding the element-wise masked feature $\mathbf{A} \odot \mathbf{J}$ through $S(\cdot)$, $T(\cdot)$ is updated by the **SIA** loss

$$\mathcal{L}_{SIA} = \mathcal{L}_S(S(\mathbf{A} \odot \mathbf{J}), \mathbf{Y}). \tag{5}$$

Training with the $\mathbf{A} \odot \mathbf{J}$-induced loss $\mathcal{L}_{SIA}$, $C(\cdot)$ is forced to balance feature learning in all regions (See the strong foreground gradients at **J** in Figure 3(c)).

In summary, the MAC Training procedure consists of initializing $T(\cdot)$ with **(1)** and updating $T(\cdot)$ with **(2-3)**. Figure 3(c) shows that MAC training produces foreground-focused features in **J**, stronger gradients at **J**, and better instance mask detection than the *Naïve* training in Figure 3(a).

4 EXPERIMENTS

We use the following keywords throughout this section. **"Standard"** refers to the source model training strategy (backbone pre-trained on Imagenet and $S(\cdot)$ fully updated with $\mathcal{L}_S$). **"MAC training"** refers to our techniques. **"MAC-Self-supervised":** Training on Pseudo-GT generated by inference $S(\cdot)$ with the source inputs of the target-source pairs in the target-modality training set. **"MAC-Supervised":** Training on manually annotated GT of the target-modality training set. **"MAC-Semi-Supervised":** Training on Pseudo-GT and a subset of manually annotated GT. All implementation details are described in the supplementary materials.

4.1 WiFi-input Target Model

In Table 1, we build WiFi-input models upon an RGB-input MaskRCNN model(source model) on the Person-in-Wifi(**PiW**) [61] dataset[2]. The PiW dataset contains synchronized RGB videos (20FPS) and Channel State Information(CSI) sequences (100Hz) of the Wifi signal. There are 16 indoor layouts 16 with multiple persons captured. One RGB frame (resized to $[3, 640, 384]$) corresponds to a CSI tensor ([CSI_samples, transmitters, receivers, sub-carriers]=$[5, 3, 3, 30]$). The model in [61] only produces image-wise semantic mask and body-joint heatmaps and cannot be directly compared on person detection metrics. No common ground truth was annotated to compare the RGB-input model and WiFi-input model.

To overcome these drawbacks, we propose the following setting: **(1)** All models are trained on and evaluated against a common ground truth **X101-GT**, which is generated by an MS-COCO-pre-trained ResNeXt101-FPN-MaskRCNN-32x8d(x3) model in Detectron2 [65] on RGB inputs. **(2)** Source model $S(\cdot)$: an RGB-input R50-FPN-MaskRCNN pre-trained on MS-COCO and fine-tuned on the PiW data. **(3)** Target model baseline ("PiW$|S(\cdot)$" and "Wi2Vi$|S(\cdot)$"): composed by adding the CSI-to-RGB modules of PiW [61] and Wi2Vi [31] to $S(\cdot)$ respectively. We train the target baselines by randomly initializing their CSI-to-RGB modules, initializing $S(\cdot)$ with source model weights, and training with the "Standard" strategy. Pre-training the Wi2Vi module to synthesize the foreground-cropped images [31], denoted by "RGB-FG-pre-trained WiVi", does not work well on the multi-person and multi-layout PiW data.

"Our target models $C(\cdot)|S(\cdot)$" were trained by "MAC" under three configurations: "MAC-Self-supervised" produces box and mask mAP of $[71.21, 63.67]$ using **0%** target-modality annotation, which outperforms the best target baseline metrics [68.12, 54, 79] on 100% target-modality annotation. This shows that MAC effectively transferred priors from the strong RGB-input model to the CSI-input model. Compared to the best baseline(Wi2Vi), the overheads of MAC models are 10% in Flops and 3% in #Para. "MAC-Semi-supervised" and "MAC-Supervised" outperforms all other target models using 10% and 100% target-modality annotation, respectively. Figure 4(a) visualizes the results.

[2]https://www.donghuang-research.com/publications

4.2 Lidar-input Target Model

We build Lidar Range-input models upon an RGB-input DD3D model [47] on the Kitti-3D dataset [14]. Since there is no 360-degree RGB coverage that matches the 360-degree Lidar scans, we only evaluate results on the **frontal-view sector** of the Lidar scans[3] overlapped the RGB Camera#2 field-of-view. The 32-beam Lidar scans at 1-degree horizontal resolution, creating 32×90 points in the frontal-view sector, which are then projected to the 384×1270 pixel grid to match the RGB pixels. Following [12, 43], the missing range pixels are filled by a fixed depth value of 80 (meters). Following [47], we report $AP|_{40}$ [56] computed on the training|testing split of 3712|3769 samples. As shown in Figure 4(b), the range image has very sparse depth pixels with no visual appearance.

In Table 2, the "Target model baseline", directly training a DD3D on Range-inputs, produces higher metrics than the RGB-input Source models, which indicates the advantage of the Lidar over the RGB camera. It appears that the weak RGB-input model cannot provide good prior knowledge to develop the Lidar-input model. However, the MAC-Self-supervised target model, with **0%** target-modality annotation, still outperforms the target baseline trained on 100% annotation. Trained on 10% and 100% target-modality annotations respectively, our MAC-Self-Supervised and MAC-Supervised target models easily outperform all other models in all $AP|_{40}$ metrics. Figure 4(b) visualizes the results.

4.3 Infrared-input Target Model

On the LLVIP dataset [27] [4], we show how MAC works when the source modality(RGB) contains far less information than the target modality(Thermal InfRared (TIR)) under low-light vision. We used the official training/testing split containing 15488 RGB-TIR pairs with manually labeled 2D bounding boxes.

In Table 3, we used the R50-FPN-FastRCNN (input size 1024×1280) for both the source model (RGB input) and target model baseline (TIR input). Both models were pre-trained on MS-COCO and fine-tuned on the RGB and TIR inputs respectively. The MAC target models are trained with the MAC training algorithm. The source model produces Bbox Average Precision(Box AP) of 43.83, which is clearly inferior to the target model baseline (Box AP 55.58), therefore may not provide strong priors or correct Pseudo labels for the target model. However, the MAC-Self-supervised target model still gets Box AP of 55.63 using **0%** target-modality annotation, which is comparable to the target baseline trained on 100% annotation. With only 10% annotation, the MAC-Semi-supervised model easily outperforms the target baseline. With 100% annotation, the MAC-supervised target model outperforms all other models. Figure 4(c) visualizes the results. In this experiment, the higher MAC overheads are due to the large input size, which could be reduced by concatenating $C(\cdot)$ with the smaller feature maps of $S(\cdot)$.

[3]LaserNet [43], RangeDet [12] only reported results on 360-degree Range-inputs with no pre-trained model released to evaluate the frontal sector range data.
[4]https://github.com/bupt-ai-cz/LLVIP

Model (on **x101-GT**)	Input	Training Strategy	Box mAP ↑	Mask mAP ↑	Target-Modality Annot. (%) ↓	Inference Flops|#Para.
Source models $S(\cdot)$						
R50-FPN-MaskRCNN	RGB	Coco-pretrain [65] + Standard	82.36	87.94	-	61.42G|44.30M
Target models baselines						
PiW [61]|$S(\cdot)$	CSI	Source Init. $S(\cdot)$+ Standard	59.86	45.08	100	62.27G|45.0M
Wi2Vi [31]|$S(\cdot)$	CSI	RGB-FG-pretrained Wi2Vi [31]	0.12	0.09	100	63.22G|49.82M
Wi2Vi [31]|$S(\cdot)$	CSI	Source Init. $S(\cdot)$+Standard	68.12	54.79	100	63.22G|49.82M
Ours target models $C(\cdot)\|S(\cdot)$						
MAC-Self-supervised	CSI	MAC training	71.21	63.67	**0**	70.10G|51.34M
MAC-Semi-Supervised	CSI	MAC training	74.65	65.86	10	70.10G|51.34M
MAC-Supervised	CSI	MAC training	**77.38**	**66.49**	100	70.10G|51.34M

Table 1: WiFi CSI-input model results on the Person-in-Wifi(PiW) dataset (all 16 layouts). All models were trained and evaluated against GTs generated by X101-FPN-MaskRCNN(×3) in Detectron2 model zoo. The best target model metrics are in bold.

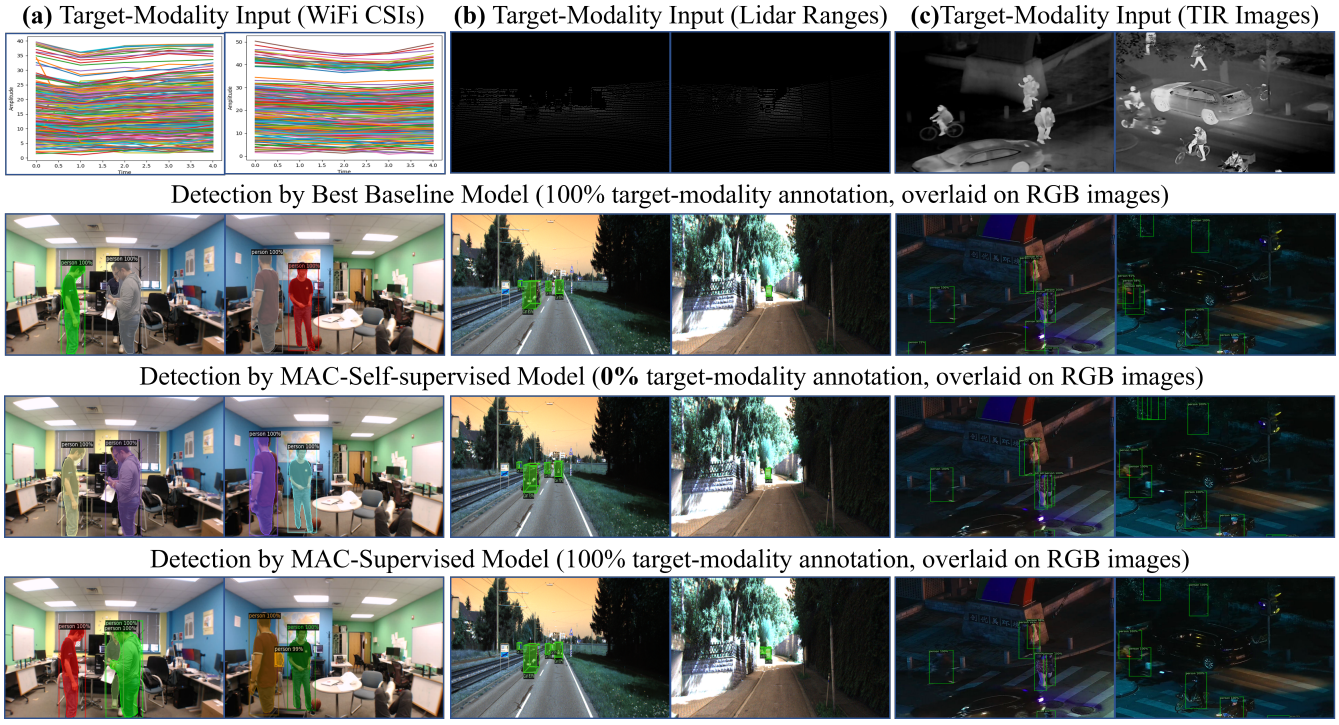

Figure 4: Qualitative results on three target-modality inputs (a) WiFi CSIs (amplitude sequences corresponding to each frame), (b) Lidar Ranges (sparse depth points projected on the pixel grid) and (c) Thermal InFraRed (TIR) images.

Models	Input	Training	Car BEV AP|$_{40}$ [Easy, Med., Hard] ↑	Car 3D-Bbox AP|$_{40}$ [Easy, Med., Hard] ↑	Target-Modality Annot. (%) ↓	Inference Flops|#Para.
Source models $S(\cdot)$						
DD3D-DLA34 [47](github)	RGB	Standard	[31.7, 24.4, 21.7]	[22.6, 17.0, 14.9]	-	109.9G|25.6M
Target model baselines						
DD3D-DLA34	Range	Standard	[40.7, 25.4, 22.0]	[29.4, 17.9, 14.9]	100	109.9G|25.6M
Ours target models $C(\cdot)\|S(\cdot)$						
MAC-Self-supervised	Range	MAC	[41.5, 26.1, 23.2]	[30.2, 18.2, 15.4]	**0**	123.6G|28.1M
MAC-Semi-Supervised	Range	MAC	[43.6, 27.3, 25.5]	[32.1, 19.8, 16.8]	10	123.6G|28.1M
MAC-Supervised	Range	MAC	**[46.3, 33.4, 30.9]**	**[35.4, 21.8, 19.9]**	100	123.6G|28.1M

Table 2: Lidar Range-input model results on the KITTI validation set. Metrics (Car metrics only) are evaluated on the frontal sector of Lidar scans matching the RGB Camera 2 field-of-view. The best target model metrics are in bold.

4.4 Ablation Study

In Table 4, we conduct an ablation study of the MAC training techniques on the "2018_10_17_2" subset of the PiW dataset. We start with the baseline training strategy: "RandInit+$\mathcal{L}_S$", and add MAC or alternative techniques in four groups. The cumulative relation among groups is denoted by their indents of "+"s. Each

Models	Input	Training	Box AP ↑	Target-Modality Annot. (%) ↓	Inference Flops|#Para.
Source model $S(\cdot)$					
R50-FPN-FasterRCNN	RGB	Standard	43.83	-	255G|41.70M
Target model baseline					
R50-FPN-FasterRCNN	TIR	Standard	55.58	100	255G|41.70M
Ours target models $C(\cdot)\|S(\cdot)$					
MAC-Self-supervised	TIR	MAC	55.63	**0**	302G|44.35M
MAC-Semi-Supervised	TIR	MAC	57.08	10	302G|44.35M
MAC-Supervised	TIR	MAC	**58.05**	100	302G|44.35M

Table 3: Thermal InfRared TIR-input model results on the LLVIP dataset. Best in bold.

group is added upon its previous MAC techniques. For instance, "Feature-based KD..." and "Source Init. $S(\cdot)$..." are both added upon "FSR pretrained $C(\cdot)$...".

Training Strategies on $C(\cdot)\|S(\cdot)$	Box AP ↑	Mask AP ↑	Target-Modality Annot. (%)↓
Rand. Init.+Standard (Baseline)	75.28	66.26	100
MAC (bold rows) vs. Alternative techniques			
+ w/o FSR-pretrained $C(\cdot)$	76.69	67,03	100
+ FSR-pretrained $C(\cdot)$	**78.83**	**68.73**	100
+ Feature-based KD [69] from $S(\cdot)$	70.90	54.43	100
+ Source Init. $S(\cdot)$ **and two-stage update**	**80.36**	**70.55**	100
+ SS loss $\mathcal{L}_{SS}$ (Eq.(2))	81.03	71.09	100
+ DSS loss $\mathcal{L}_{DSS}$ (Eq.(3))	**81.60**	**71.62**	100
+ RSC [24] on the $C(\cdot)$ output layer	80.94	71.21	100
+ RSC [24] on the Res-50-p5 layer	81.89	71.92	100
+ SIA(Eq.(5)	**82.15**	**72.33**	100
MAC-Self-supervised	75.14	65.96	**0**

Table 4: Ablation study of training techniques on MAC target model $C(\cdot)|S(\cdot)$ on the "2018_10_17_2" subset of PiW. The cumulative relation among groups of "MAC vs.Alternative techniques" are denoted by their indents of "+"s. Each group of techniques is added upon the MAC techniques (bold rows) of the previous group.

Within each compared group, the MAC technique (bold rows) produces better metrics than their alternative counterparts. "FSR-pretrained $C(\cdot)$" and "Source Init. $S(\cdot)$ and two-stage update" provide better priors to the target model than"w/o FSR" and "Feature-based KD". SIA is better than RSC [24] which computes and applies attention on the same layer ("the C(·) output layer" or "the Res-50-p5 layer"). Using 100% target-modality annotation, our final model (after applying SIA) produces [82.15, 72.33], which is significantly better than the baseline ([75.28, 66.26]). Trained on the Pseudo GT generated by $S(\cdot)$, MAC-Self-supervised produces [75.14, 65.96] comparable to the baseline that is trained on 100% manual target-modality annotation. Besides using the same number of pseudo images as the real images, we also tried fewer (0, 1/2) pseudo images and got [77.13,68.25] and [79.83,70.12] respectively. We also tried vanilla VAE as a calibrator and only got [0.5, 0.3] due to the VAE's inferior fine-grain reconstruction ability than VQVAE.

In summary, DSS is proven better than SS, and also improves upon "Source Init. S(·) and two-stage update". SIA is proven better than two RSC variants, and also improves upon "DSS loss LDSS (Eq.(3))". "MAC-Self-supervised", that uses 0 manual annotations, has comparable performance as "Baseline" that uses 100% manual annotations. Using 100% annotations, "MAC-Supervised" is significantly better than "Baseline".

5 DISCUSSION

Modality Calibration vs. Domain Adaption: By definition[5], domain adaptation mainly addresses the shift of data sampling distribution. Our modality calibration problem mainly addresses the change in the spatial, temporal, and physical nature of the inputs. Although some work consider image translation as both modality calibration and domain adaption, our work applies to any non-image modalities such as WiFi signals. Instead of pursuing realistic pixels in image domain, we ONLY generate foreground-associated features contributing to the detection tasks, which aims to reduce the efforts in DNN design and data annotation. Moreover, for simplicity, this work is presented under the assumption that the source and target modality data are sampled from the same foreground/background distributions, such that MAC is only focused on reducing the discrepancy between modalities.

Pros. and Cons. of MAC: By adding a calibrator to a pre-trained RGB perception model, MAC introduces an efficient pipeline for switching input modalities of DNNs, while bringing back the old deep learning challenge: the vanishing gradient problem [1, 16]. To address this challenge, multiple MAC training techniques, FSR, DSS, and SIA, are developed to generate strong and dense gradients.

MAC-Self-supervised vs. MAC-Supervised: In "MAC-Self-supervised" (see the cases of "Target-Modality Annot." 0% in Table 1-4) the GTs for training the MAC target model are the pseudo annotations produced by inferencing a pre-trained source model. There is NO manual GT for the target modality. "MAC-Supervised" uses the real manually annotated GTs (see the case of "MAC-Supervised" with "Target-Modality Annot." 100% in Table 1-4). This presents us with options to balance the performance and the effort of annual data annotation.

General use of MAC: The same FSR, DDS, and SIA training techniques can be used to develop all non-RGB modality models. Only different calibrators need to be developed for different non-RGB modalities.

6 CONCLUSIONS

We proposed MAC, an efficient pipeline for switching input modalities of DNNs. In training a target-modality model, MAC leverages the prior knowledge from the source-modality model and requires as few as **zero** target-modality annotations. The MAC components (FSR, DSS, SIA) could potentially be used to compose any cross-modality models, for instance, using DALL-E-mini as a calibrator to compose a text-input model.

Potential negative social impact: If the source model were trained on private RGB images, data-privacy concerns may arise from the source model inversion operation in MAC.

REFERENCES

[1] Y. Bengio, P. Simard, and P. Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks* 5, 2 (1994), 157−−166.
[2] Ang Cao and Justin Johnson. 2021. Inverting and Understanding Object Detectors. *arXiv:2106.13933* (2021).
[3] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. 2017. Learning efficient object detection models with knowledge distillation. In *NIPS*. 742–751.

[5] https://en.wikipedia.org/wiki/Domain_adaptation

[4] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. 2019. MMDetection: Open MMLab Detection Toolbox and Benchmark. *arXiv:1906.07155* (2019).

[5] Shuhan Chen, Xiuli Tan, Ben Wang, Huchuan Lu, Xuelong Hu, and Yun Fu. 2020. Reverse Attention Based Residual Network for Salient Object Detection. *TIP* 29 (2020), 3763–3776.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

[7] S. Depatla and Y. Mostofi. 2018. Crowd Counting Through Walls Using WiFi. In *Proceedings of IEEE International Conference on Pervasive Computing and Communications*.

[8] Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyan Wu, and Rama Chellappa. 2019. Learning without memorizing. In *CVPR*. 5138–5146.

[9] Qi Dou, Cheng Ouyang, Cheng Chen, Hao Chen, and Pheng-Ann Heng. 2018. Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*. 691–697.

[10] Michael Drob. 2021. RF PIX2PIX Unsupervised Wi-Fi to Video Translation. In *arXiv:2102.09345*.

[11] Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming Transformers for High-Resolution Image Synthesis. *arXiv:2012.09841* (2021).

[12] Lue Fan, Xuan Xiong, Feng Wang, Naiyan Wang, and ZhaoXiang Zhang. 2021. RangeDet: In Defense of Range View for LiDAR-Based 3D Object Detection. In *ICCV*.

[13] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. 2021. StyleGAN-NADA: CLIP-Guided Domain Adaptation of Image Generators. arXiv:2108.00946 [cs.CV]

[14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *CVPR*.

[15] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. 2020. Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation. *arXiv:2012.07177* (2020).

[16] X. Glorot and Y. Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*.

[17] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. 2016. Cross Modal Distillation for Supervision Transfer. In *CVPR*.

[18] Daniel Halperin, Wenjun Hu, Anmol Sheth, and David Wetherall. 2011. Tool release: Gathering 802.11 n traces with channel state information. *ACM SIGCOMM Computer Communication Review* 41, 1 (2011), 53–53.

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2017. Mask R-CNN. *CoRR* abs/1703.06870 (2017). arXiv:1703.06870 http://arxiv.org/abs/1703.06870

[20] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. 2018. Bag of Tricks for Image Classification with Convolutional Neural Networks. *arXiv:1812.01187* (2018).

[21] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv:1503.02531* (2015).

[22] Yuenan Hou, Zheng Ma, Chunxiao Liu, and Chen Change Loy. 2019. Learning lightweight lane detection cnns by self attention distillation. In *ICCV*. 1013–1021.

[23] Zeyi Huang, Wei Ke, and Dong Huang. 2020. Improving Object Detection with Inverted Attention. In *WACV*.

[24] Zeyi Huang, Haohan Wang, Eric P. Xing, and Dong Huang. 2020. Self-Challenging Improves Cross-Domain Generalization. In *ECCV*.

[25] Zeyi Huang, Yang Zou, BVK Kumar, and Dong Huang. 2020. Comprehensive Attention Self-Distillation for Weakly-Supervised Object Detection. *NIPS* 33 (2020).

[26] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. 2015. Multispectral Pedestrian Detection: Benchmark Dataset and Baselines. In *CVPR*.

[27] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. 2021. LLVIP: A Visible-infrared Paired Dataset for Low-light Vision. In *Proceedings of the ICCV*. 3496–3504.

[28] Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, ChristopherSTAN, Liu Changyu, Laughing, tkianai, Adam Hogan, lorenzomammana, yxNONG, AlexWang1900, Laurentiu Diaconu, Marc, wanghaoyang0106, ml5ah, Doug, Francisco Ingham, Frederik, Guilhen, Hatovix, Jake Poznanski, Jiacong Fang, Lijun Yu, changyu98, Mingyu Wang, Naman Gupta, Osama Akhtar, PetrDvoracek, and Prashant Rai. 2020. *ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements*. https://doi.org/10.5281/zenodo.4154370

[29] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *Proc. CVPR*.

[30] Sorachi Kato, Takeru Fukushima, T. Murakami, H. Abeysekera, Yusuke Iwasaki, T. Fujihashi, Takashi Watanabe, and S. Saruwatari. 2021. CSI2Image: Image Reconstruction From Channel State Information Using Generative Adversarial Networks. *IEEE Access* 9 (2021), 47154–47168.

[31] Mohammad Hadi Kefayati, Vahid Pourahmadi, and Hassan Aghaeinia. 2020. Wi2Vi: Generating Video Frames from WiFi CSI Samples. *IEEE Sensors Journal* 20, 19 (2020), 11463–11473.

[32] JangHyun Kim, Wonho Choo, Hosan Jeong, and Hyun Oh Song. 2021. Co-Mixup: Saliency Guided Joint Mixup with Supermodular Diversity. In *International Conference on Learning Representations*.

[33] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV* (2020).

[34] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. 2019. Pointpillars: Fast encoders for object detection from point clouds. In *CVPR*. 12697–12705.

[35] Heju Li, Xin He, Xukai Chen, Yinyin Fang, and Qun Fang. 2019. Wi-Motion: A Robust Human Activity Recognition Using WiFi Signals. *IEEE Access* 7 (2019), 153287 – 153299.

[36] Zhizhong Li and Derek Hoiem. 2017. Learning Without Forgetting. *PAMI* 40, 12 (2017), 2935–2947.

[37] Matthias Limmer and Hendrik P.A. Lensch. 2019. Infrared Colorization Using Deep Convolutional Neural Networks. *arXiv:1604.02245* (2019).

[38] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *ECCV* (2014-01-01). Zürich. /se3/wp-content/uploads/2014/09/coco_eccv.pdf,http://mscoco.org

[39] Lanlan Liu, Michael Muelly, Jia Deng, Tomas Pfister, and Jia Li. 2019. Generative Modeling for Small-Data Object Detection. In *ICCV*.

[40] Shangqing Liu, Yanchao Zhao, Fanggang Xue, Bing Chen, and Xiang Chen. 2019. DeepCount: Crowd counting with WiFi via deep learning. *arXiv:1903.05316* (2019).

[41] Xialei Liu, Hao Yang, Avinash Ravichandran, Rahul Bhotika, and Stefano Soatto. 2020. Continual Universal Object Detection. *arXiv:2002.05347* (2020).

[42] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *arXiv:2103.14030* (2021).

[43] Gregory P Meyer, Ankit Laddha, Eric Kee, Carlos VallespiGonzalez, and Carl K Wellington. 2019. LaserNet: An Efficient Probabilistic 3D Object Detector for Autonomous Driving. In *CVPR*. 12677–12686.

[44] Stefan Milz, Martin Simon, Kai Fischer, and Maximillian Pöpperl. 2019. Points2Pix: 3D Point-Cloud to Image Translation using conditional Generative Adversarial Networks. *arXiv:1901.09280* (2019).

[45] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. 2018. Image to Image Translation for Domain Adaptation. In *CVPR*.

[46] Luigi Musto and Andrea Zinelli. 2020. Semantically Adaptive Image-to-image Translation for Domain Adaptation of Semantic Segmentation. In *BMVC*.

[47] Dennis Park, Rares Ambrus, Vitor Guizilini, Jie Li, and Adrien Gaidon. 2021. Is Pseudo-Lidar needed for Monocular 3D Object detection?. In *ICCV*.

[48] Fabio Pizzati, Raoul de Charette, Michela Zaccaria, and Pietro Cerri. 2020. Domain Bridge for Unpaired Image-to-Image Translation and Unsupervised Domain Adaptation. In *WACV*.

[49] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. 2016. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. *arXiv:1612.00593* (2016).

[50] Charles R Qi, Li Yi, Hao Su, and Leonidas J Guibas. 2017. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. *arXiv:1706.02413* (2017).

[51] Jakaria Rabbi, Nilanjan Ray, Matthias Schubert, Subir Chowdhury, and Dennis Chao. 2020. Small-Object Detection in Remote Sensing Images with End-to-End Edge-Enhanced GAN and Object Detector Network. *Remote Sensing* 12, 9 (2020), 1432.

[52] Rahul Rajendran, Thaweesak Trongtirakul, Thaweesak Trongtirakul, Karen Panetta, and Sos Agaian. 2019. A pixel-based color transfer system to recolor nighttime imagery. In *Mobile Multimedia/Image Processing, Security, and Applications*.

[53] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. 2019. Generating Diverse High-Fidelity Images with VQ-VAE-2. *arXiv:1906.00446* (2019).

[54] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2014. Fitnets: Hints for thin deep nets. *arXiv:1412.6550* (2014).

[55] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. 2019. PointRCNN: 3D Object Proposal Generation and Detection From Point Cloud. In *CVPR*.

[56] Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Manuel Lopez Antequera, and Peter Kontschieder. 2020. Disentangling monocular 3d object detection: From single to multiclass recognition. *PAMI* (2020).

[57] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, and et al. 2020. Scalability in Perception for Autonomous Driving: Waymo Open Dataset. In *CVPR*.

[58] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural Discrete Representation Learning. *arXiv:1711.00937* (2017).

[59] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba. 2013. HOGgles: Visualizing Object Detection Features. *ICCV* (2013).
[60] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. 2017. Residual attention network for image classification. In *CVPR*. 3156–3164.
[61] Fei Wang, Sanping Zhou, Stanislav Panev, Jinsong Han, and Dong Huang. 2019. Person-in-WiFi: Fine-grained Person Perception using WiFi. In *ICCV*.
[62] Lezi Wang, Ziyan Wu, Srikrishna Karanam, Kuan-Chuan Peng, Rajat Vikram Singh, Bo Liu, and Dimitris N Metaxas. 2019. Sharpen focus: Learning with attention separability and consistency. In *ICCV*. 512–521.
[63] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. 2019. Distilling Object Detectors With Fine-Grained Feature Imitation. In *CVPR*. 4933–4942.
[64] Wei Wang, Alex X. Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu. 2017. Device-free Human Activity Recognition Using Commercial WiFi Devices. *IEEE Journal on Selected Areas in Communications* 35, 5 (2017), 1118–1131.
[65] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2.
[66] Xinpeng Xie, Jiawei Chen, Yuexiang Li, Linlin Shen, Kai Ma, and Yefeng Zheng. 2020. Self-Supervised CycleGAN for Object-Preserving Image-to-Image Domain Adaptation. In *ECCV*.
[67] Yan Yan, Yuxing Mao, and Bo Li. 2018. Second: Sparsely embedded convolutional detection. *Sensors* (2018).
[68] Sergey Zagoruyko and Nikos Komodakis. 2016. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv:1612.03928* (2016).
[69] Linfeng Zhang and Kaisheng Ma. 2021. Improve Object Detection with Feature-based Knowledge Distillation: Towards Accurate and Efficient Detectors. In *ICLR*.
[70] Zhi Zhang, Tong He, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. 2019. Bag of Freebies for Training Object Detection Neural Networks. *arXiv:1902.04103* (2019).
[71] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. 2018. Through-wall human pose estimation using radio signals. In *CVPR*. 7356–7365.
[72] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. 2017. Random erasing data augmentation. *arXiv:1708.04896* (2017).
[73] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *CVPR*. 2921–2929.
[74] Barret Zoph, Ekin D. Cubuk, Golnaz Ghiasi, Tsung-Yi Lin, Jonathon Shlens, and Quoc V. Le. 2020. Learning Data Augmentation Strategies for Object Detection. In *ECCV*.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009