# RCTDistill: Cross-Modal Knowledge Distillation Framework for Radar-Camera 3D Object Detection with Temporal Fusion

## Supplementary Material

This Supplementary Material provides further information to complement our main paper. We first provide a comprehensive description of our model's implementation details (Section 1), followed by detailed architecture specifics (Section 2). Subsequently, we present extended experimental results, including additional ablation studies and quantitative outcomes (Section 3). Finally, Section 4 presents qualitative results with visualizations.

## 1. Implementation Details

We implemented our model using the MMDetection3D [2] open-source framework. The hyperparameters $\lambda_{\text{RA}}$, $\lambda_{\text{T}}$, and $\lambda_{\text{RD}}$, which control the weights of each loss term in Equation (12), were empirically set to 6, 0.01, and 50, respectively. We measured RCTDistill's FPS on an RTX 3090 GPU, and used the FPS reported in the respective papers for all other methods

**Evaluation Metrics.** We follow the official nuScenes [1] evaluation protocol and report the nuScenes Detection Score (NDS), which integrates mean Average Precision (mAP) with five true-positive submetrics: mean Average Translation Error (mATE), mean Average Scale Error (mASE), mean Average Orientation Error (mAOE), mean Average Velocity Error (mAVE), and mean Average Attribute Error (mAAE).

**Training configuration.** Our baseline model employs a two-stage training strategy: first, training with single-frame input for 6 epochs, followed by incorporating 8 past frames for the subsequent 60 epochs (66 epochs in total). For our proposed RCTDistill, we follow the training protocol of CRKD [10], initializing from a pre-trained baseline model while keeping the teacher model frozen throughout the 60 epochs training process. For the teacher model, we employ CenterPoint with a SECOND backbone. The detailed training configurations for different backbone architectures are summarized in Table 1.

| Configs | ResNet-50 | ResNet-101 | ConvNeXt-B |
|---|---|---|---|
| Image Size | 256×704 | 512×1408 | 512×1408 |
| BEV Size | 128×128 | 256×256 | 256×256 |
| Base Learning Rate | 2e-4 | 1e-4 | 2e-4 |
| Weight Decay | 1e-7 | 1e-2 | 1e-2 |
| Optimizer Momentum | 0.9, 0.999 | 0.9, 0.999 | 0.9, 0.999 |
| Batch Size | 16 | 8 | 16 |
| Training Epochs | 60 | 60 | 60 |
| Gradient Clip | 35 | 5 | 5 |

Table 1. Training settings for different backbone networks.

**Hardware.** We perform experiments using 4× NVIDIA RTX 3090 GPUs for ResNet-50-based models and 4× NVIDIA A100 GPUs for ResNet-101 and ConvNeXt-B-based models.

**Data augmentation.** During training, we utilize comprehensive data augmentation strategies following prior works [4, 6, 8, 9]. For Image-view-space Data Augmentation (IDA), we apply horizontal flipping, scaling ($-0.06$ to $0.11$), and rotations ($\pm 0.54°$). BEV-space Data Augmentation (BDA) includes random flipping along the $X$ and $Y$ axes, scaling (0.95-1.05), and rotations ($\pm 0.3925\,\text{rad}$). We enhance radar data robustness through randomly drop sweeps and points [7]. For LiDAR data augmentation, we apply random flipping ($X$ and $Y$ axes), rotations, scaling, and translations. The detection range is set to $[-51.2\,\text{m}, 51.2\,\text{m}]$ for the $X$ and $Y$ dimensions and $[-5\,\text{m}, 3\,\text{m}]$ for $Z$, with voxel dimensions of $(0.1\,\text{m}, 0.1\,\text{m}, 0.2\,\text{m})$.

| Methods | HA-Net | TKD | NDS↑ | mAP↑ |
|---|---|---|---|---|
| Baseline | | | 55.0 | 47.1 |
| RCTDistill | ✓ | | 55.7 | 48.0 |
| | ✓ | ✓ | 57.0 | 48.5 |

Table 2. Ablation study of HA-Net.

## 2. Details of HA-Net

We introduce Temporal Knowledge Distillation (TKD) to mitigate feature misalignment along the trajectory direction caused by the motion of dynamic objects. In this process, we apply HA-Net to align and merge historical BEV features across consecutive frames, effectively facilitating TKD. The detailed structure of HA-Net is shown in Figure 1. The HA-Net first integrates historical BEV features through concatenation and a $1 \times 1$ convolution layer, followed by two down-sampling blocks. The first block comprises a $3 \times 3$ convolution with layer normalization and two UniRepLK [3] blocks, each employing $3 \times 3$ depth-wise convolutions. The second block follows a similar structure, starting with a $3 \times 3$ convolution and layer normalization, but includes only a single UniRepLK block with a larger $13 \times 13$ convolution kernel to expand the receptive field. Finally, two up blocks with residual connections are utilized to produce the aligned historical features. The large receptive field of HA-Net integrates feature variations, clearly capturing the motion of dynamic objects and effectively resolving temporal inconsistencies in historical BEV features.

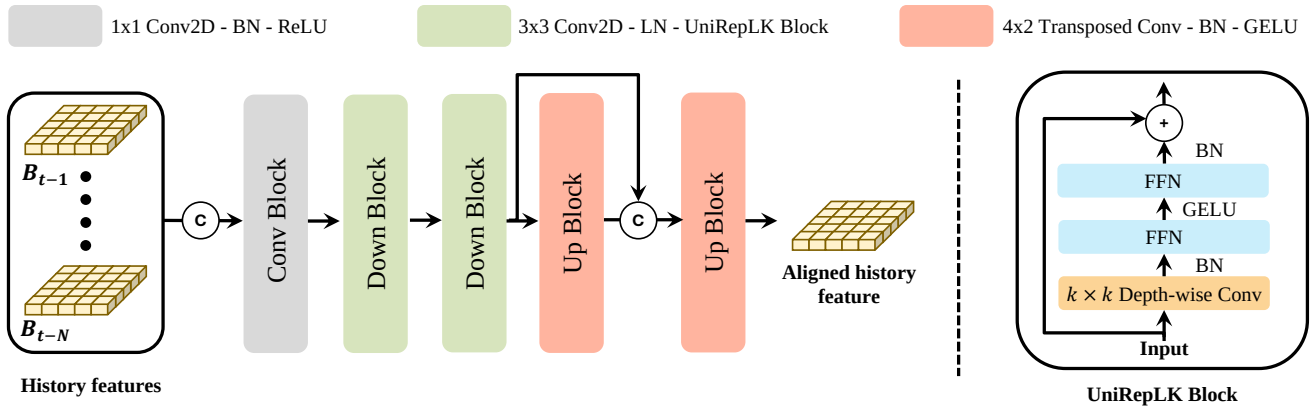



Figure 1. **Detailed structure of HA-Net.**

| $\alpha_l$ | $\alpha_w$ | NDS↑ | mAP↑ |
|---|---|---|---|
| 15 | 3 | 57.6 | 50.3 |
| 25 | 5 | 58.4 | 51.0 |
| 50 | 10 | 57.8 | 50.6 |

(a) $\alpha$: The scaling factor in RAKD.

| $\tau$ | NDS↑ | mAP↑ |
|---|---|---|
| 0.05 | 57.4 | 49.9 |
| 0.1 | **58.4** | **51.0** |
| 0.2 | 57.2 | 49.8 |

(b) $\tau$: Mask threshold in RAKD, TKD and RDKD.

| $t_s$ | NDS↑ | mAP↑ |
|---|---|---|
| 2 | 57.6 | 50.8 |
| 3 | **58.4** | **51.0** |
| 4 | 57.3 | 50.4 |

(c) $t_s$: History reference time in TKD.

Table 3. Ablation studies of hyperparameters.

| # of H.F. | Mem (GB) | GFLOPs | FPS↑ | NDS↑ | mAP↑ |
|---|---|---|---|---|---|
| 2 | 2.865 | 168.3 | 26.5 | 55.8 | 47.6 |
| 4 | 2.871 | 168.5 | 26.5 | 56.9 | 49.4 |
| 8 | 2.945 | 168.9 | 26.2 | 58.4 | 51.0 |

Table 4. Evaluation model efficiency. H.F.: history frames, Mem: GPU memory usage.

Table 2 presents the experimental results analyzing the contributions of HA-Net and TKD to performance improvement. When HA-Net was applied to the baseline model, NDS increased by 0.7% and mAP improved by 0.9%. Subsequently, implementing TKD yielded an additional enhancement of 1.3% in NDS and 0.5% in mAP. These findings suggest that the performance gains are not solely due to HA-Net's enhanced capacity but also stem from TKD's ability to address challenges posed by the motion of dynamic objects, playing a vital role in further boosting accuracy. Furthermore, the results demonstrate that HA-Net and TKD contribute to performance improvements in a synergistic manner.

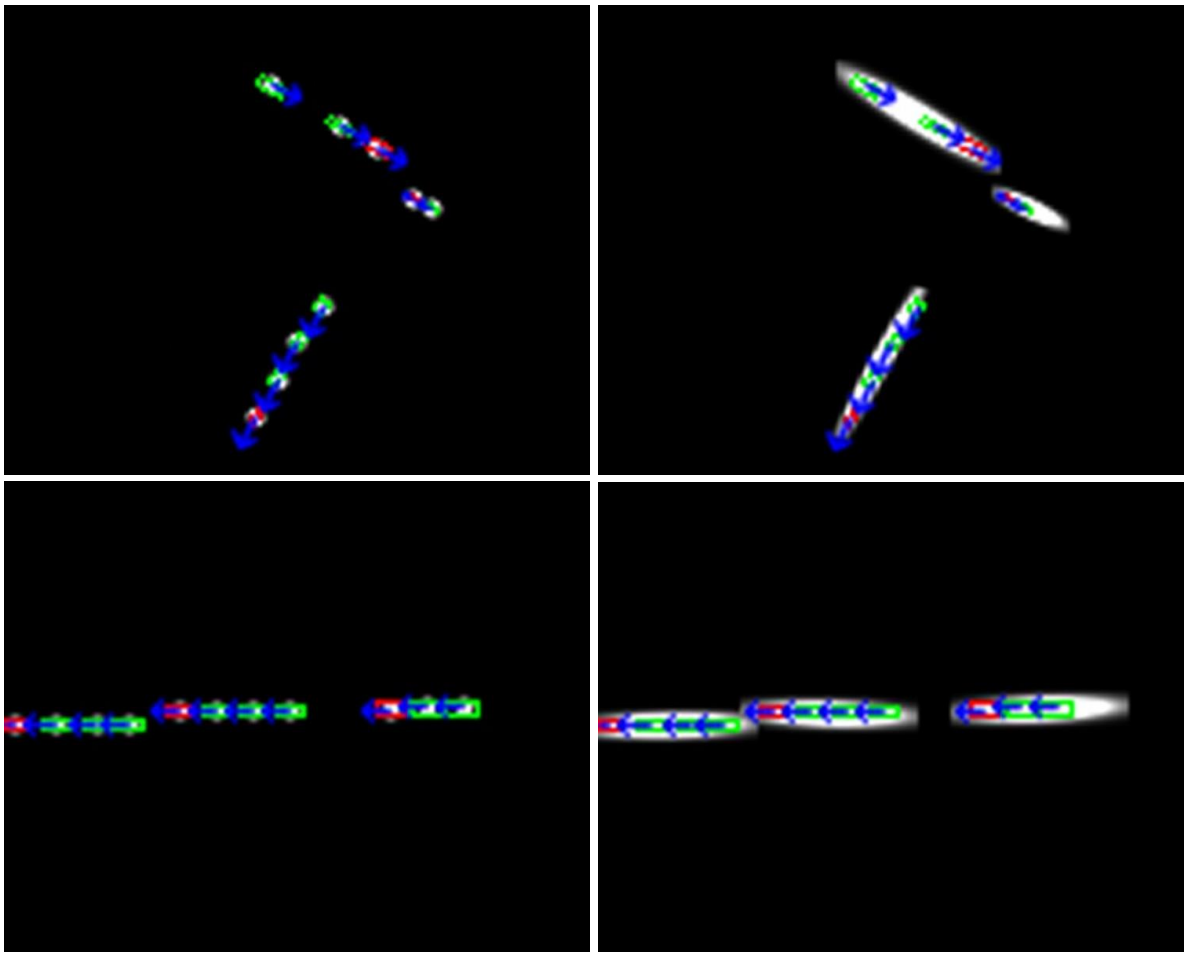
## 3. Additional Experiments

**Hyperparameters Analysis.** Table 3 presents the ablation studies for the hyperparameters used in RCTDistill. Table 3a explores the scaling factor $\alpha_l$, $\alpha_w$, which determines the lengths of the major and minor axes radius $r_1$ and $r_2$ in the elliptical region in RAKD. For both of $r_1$ and $r_2$, the best performance is achieved when $\alpha_l$ is set to 25 and $\alpha_w$ is set to 5. Table 3b analyzes the threshold $\tau$, which is used to construct the mask region in the three proposed knowledge distillation methods. As the $\tau$ value decreases, the number of background BEV grids rises, reducing the proportion of crucial BEV grids, which may hinder the model's ability to focus on essential locations. Conversely, if the $\tau$ value is too high, crucial BEV grids may be missed, potentially leading to decreased performance. Table 3c investigates the historical reference time $t_s$, which is used to determine the center position of the temporal mask for each object in TKD. When $t_s$ exceeds 3, the mask region becomes excessively large, including irrelevant information. Conversely, when $t_s$ is less than 3, it restricts the range of the temporal mask, which prevents adequate consideration of the uncertainties associated with dynamic objects. Experimental results show that optimal performance is achieved when t is set to 3.

**Efficiency analysis.** Table 4 summarizes the efficiency and accuracy changes in RCTDistill as the number of history frames increases. Our model utilizes a memory bank mechanism to store and reuse precomputed BEV features, which ensures that adding more history frames leads to only a minimal increase in GPU memory usage, GFLOPs, and the number of parameters. As a result, our model leverages additional frames with minimal impact on computational

| Methods | C.B. | R.B. | Encoder | Decoder | HEAD | Total |
|---|---|---|---|---|---|---|
| Baseline | 12.01 | 3.27 | 11.11 | 3.46 | 6.12 | 35.97 |
| RCTDistill | 12.01 | 3.27 | 13.32 | 3.46 | 6.12 | 38.25 |

Table 5. Ablation study of inference time.



(a) VCD (b) RCTDistill

Figure 2. **Comparison of temporal knowledge distillation region in RCTDistill and VCD [5].** Pink arrows represent ground truth (GT) velocityW vector at each time step. White region denotes temporal knowledge distillation region. Red boxes denotes dynamic GT at the current timestamp, while green boxes represent dynamic GT at previous timestamps. The green boxes that are continuously connected to the red boxes indicate the previous positional information of the same object. RCTDistill considers the entire trajectory of fast-moving objects, while VCD does not completely capture it.

resources, significantly boosting performance.

Table 5 provides each component's latency in milliseconds for Baseline and RCTDistill. RCTDistill was designed using three distinct distillation losses based on a baseline model. Distillation methods are typically employed only during the training phase, ensuring no additional latency is introduced during inference relative to the baseline model. However, in the Temporal Knowledge Distillation (TKD) process proposed in RCTDistill, we introduced an additional module called HA-Net to capture the temporal variation of historical BEV feature maps between consecutive frames. This integration may slightly increase the encoder's latency during inference relative to the baseline model; however, this difference is minimal and does not significantly impact overall efficiency. As a result, RCTDistill is a model designed to maximize performance through diverse distillation losses

during training while maintaining high efficiency during inference, thereby demonstrating its overall effectiveness.

| | TKD | NDS↑ | mAP↑ | mATE↓ | mAVE↓ |
|---|---|---|---|---|---|
| Static | | 44.4 | 25.7 | 0.532 | 0.054 |
| | ✓ | 45.7$_{+1.3}$ | 25.8$_{+0.1}$ | 0.524 | 0.044 |
| Dynamic | | 38.5 | 16.4 | 0.537 | 0.536 |
| | ✓ | 41.4$_{+2.9}$ | 18.0$_{+1.6}$ | 0.504 | 0.505 |

Table 6. TKD performance comparison for static and dynamic objects.

| Methods | False Positive | False Negative | mAP↑ |
|---|---|---|---|
| Baseline | 100.0 | 100.0 | 47.1 |
| Baseline+RDKD | 66.5 | 84.7 | 49.0 |

Table 7. Comparison of FP and FN before and after RDKD.

**Analysis of the TKD method.** The proposed TKD method leverages the velocity of dynamic objects at the current time step to generate an elliptical Gaussian mask along the object's trajectory. This mask is used as the knowledge distillation region, effectively mitigating errors caused by the motion of dynamic objects during the temporal fusion process. In contrast, VCD [5] proposes a distillation method that leverages the central region of the ground truth at all time steps.

As illustrated in Figure 2, the TKD approach captures the entire trajectory area even for fast-moving objects, whereas the VCD [5] method has limitations in capturing this region. These results demonstrate that our proposed method better reflects the trajectory area of dynamic objects compared to conventional method.

Table 6 presents the performance of TKD when applied to static objects ($|\mathbf{v}| \leq \tau_v$) and dynamic objects ($|\mathbf{v}| > \tau_v$), with $\tau_v$ set to 0.1 m/s. The experimental results reveal a significant performance gain in dynamic objects relative to static objects after TKD is applied. This indicates that TKD not only enhances overall detection accuracy but also provides superior performance specifically for dynamic objects.

**Analysis of the RDKD method.** In RDKD, the student model selects BEV feature positions based on a low threshold score, resulting in most areas being selected during the early stages of training. As training progresses, knowledge transfer from the teacher model helps refine these initially inaccurate selections, reducing FP (false positive) and FN (false negative).

Table 7 illustrates the performance variation before and after applying RDKD. With RDKD, the FP is reduced by 33.5%, the FN by 15.3%, and the mAP is improved by 1.9%. This indicates that RDKD contributes to overall performance enhancement by boosting foreground features and suppressing background features.

## 4. Qualitative Results

Figure 3 presents a qualitative analysis of our RCTDistill approach on the nuScenes validation dataset. We compare the object detection results between RCTDistill and a baseline model by visualizing the predictions in BEV space. Both models utilize a ResNet-50 backbone to extract features from camera inputs. The left image highlights the effectiveness of TKD, showing that RCTDistill significantly reduces false positive detections along the trajectory of dynamic objects compared to the baseline. Similarly, the middle and right images emphasize the benefits of RAKD, as RCTDistill achieves higher detection accuracy in the range-azimuth direction and demonstrates superior performance in detecting small objects compared to the baseline.

## References


[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020. 1

[2] MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d, 2020. 1

[3] Xiaohan Ding, Yiyuan Zhang, Yixiao Ge, Sijie Zhao, Lin Song, Xiangyu Yue, and Ying Shan. Unireplknet: A universal perception large-kernel convnet for audio video point cloud time-series and image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5513–5524, 2024. 1

[4] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 1

[5] Linyan Huang, Zhiqi Li, Chonghao Sima, Wenhai Wang, Jingdong Wang, Yu Qiao, and Hongyang Li. Leveraging vision-centric multi-modal expertise for 3d object detection. *Advances in Neural Information Processing Systems*, 36:38504–38519, 2023. 3, 4

[6] Youngseok Kim, Juyeb Shin, Sanmin Kim, In-Jae Lee, Jun Won Choi, and Dongsuk Kum. Crn: Camera radar net for accurate, robust, efficient 3d perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17615–17626, 2023. 1

[7] Zhaoqi Leng, Guowang Li, Chenxi Liu, Ekin Dogus Cubuk, Pei Sun, Tong He, Dragomir Anguelov, and Mingxing Tan. Lidar augment: Searching for scalable 3d lidar data augmentations. In *IEEE International Conference on Robotics and Automation*, pages 7039–7045. IEEE, 2023. 1

[8] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detec-

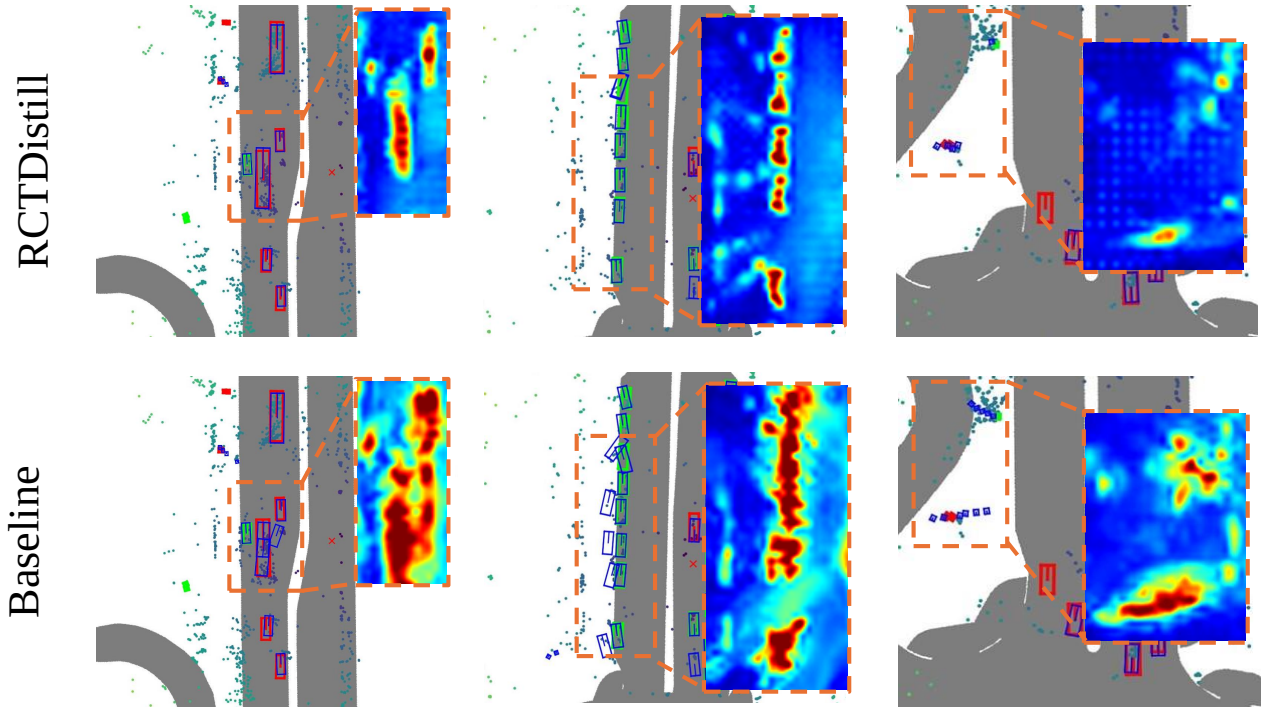



Figure 3. **Qualitative results comparing RCTDistill and Baseline.** Blue, green and red boxes denotes prediction, static and dynamic ground truth (GT), respectively. Highlighted regions with dashed orange boxes emphasize areas where RCTDistill effectively handles modality-specific uncertainties and the motion of dynamic object, demonstrating improved object localization and reduced false positives compared to the Baseline.

tion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1477–1485, 2023. 1

[9] Jinhyung Park, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris Kitani, Masayoshi Tomizuka, and Wei Zhan. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. In *International Conference on Learning Representations*, 2023. 1

[10] Lingjun Zhao, Jingyu Song, and Katherine A Skinner. Crkd: Enhanced camera-radar object detection with cross-modality knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15470–15480, 2024. 1