Scale-Equivariant Object Perception for Autonomous Driving : Supplementary Material

Taekhyung Cho

I. Introduction

This article provides readers with detailed information that could not be comprehensively addressed in the previously published paper [1]. The original paper introduces the Separable Scale-Equivariant (SSE) backbone network, which effectively processes high-frequency components sensitive to scale variations in images, and demonstrates its superior performance through object detection and tracking evaluations on the KITTI dataset. However, after the paper was accepted in IEEE Transactions on Intelligent Vehicles (T-IV), several implementation specifics and deviations from the core theoretical principles of previous benchmarks, including scale-equivariant methods and backbone network architecture, were identified. These issues involve the basis function implementations and the use of additional activation functions, which were either briefly mentioned or inadvertently omitted from the original manuscript. To ensure transparency and reproducibility, this document presents comprehensive experimental evaluations examining the effects of these implementation choices. The remainder of this document is structured as follows: Section II details the modified components, analyzes their technical implications, and discusses their effect on Scale-Equivariance (SE) error. Section III presents comparative experimental results focusing on the most significant modification. Section IV provides concluding remarks.

II. Modifications & Deviations

The original scale-equivariant basis function proposed in Scale-Equivariant Steerable Networks [3] and DISCO [4] has a shape of $[N_b,\ S,\ K,\ K]$, where N_b denotes the number of basis functions, S represents the number of scales, and K indicates the kernel size. This configuration ensures that each scale S maintains a distinct set of N_b basis functions, which is referred to as the scale-isolated basis function in this article. However, during reimplementation of the baseline networks [3] and [4], as well as the proposed SSE for evaluation, several unintended modifications were introduced that deviate from the original mathematical framework of scale-equivariance. Additionally, the underlying ResNet [2] backbone architecture was inadvertently modified during the implementation process.

a) ***Scale-Equivariant Basis Function****:* Following the publication of [1], a detailed review identified deviations from the theoretical frameworks of SESN and DISCO in the evaluated implementations, compromising the intended properties of scale-equivariance. First, SESN and the proposed SSE were implemented with a fixed kernel size across all scales to maintain ResNet compatibility. However, the steerable filter property $\psi_\sigma(\mathbf{x}) = \sigma^{-1}\psi(\sigma^{-1}\mathbf{x})$ requires the size of the kernel to scale proportionally with the scale σ to preserve the spatial coverage of the filter. For example, a filter at $\sigma = 1.0$ is spatially expanded to properly represent corresponding functions at $\sigma \in \{1.4, 2.0\}$. As a result, the fixed kernel size truncates this expansion, causing a breakdown of the scale-equivariance property for larger scales, particularly at $\sigma = 2.0$ in evaluations. Table I demonstrates that the corrected implementation (denoted by *), which restores scale-adaptive kernel sizing, achieves lower SE error in most cases, affirming the validity of this theoretical requirement.

TABLE I
Comparative analysis of SE error across implementation modification

	Scale-combined					
Backbone	**w/ additional activation**			**w/o additional activation**		
	P2	**P3**	**P4**	**P2**	**P3**	**P4**
Vanilla [2]	-	-	-	0.7059	0.7469	0.8677
SESN [3]	0.1399	0.1569	0.2609	0.2589	0.3514	0.4418
SESN* [3]	0.1411	0.1625	0.2717	0.2734	0.3841	0.4893
DISCO [4]	0.1117	0.1349	0.2354	0.2254	0.3096	0.3983
SSE [1]	0.0515	0.0573	0.1065	0.1043	0.1440	0.1968
SSE* [1]	**0.0496**	**0.0514**	**0.0934**	**0.0958**	**0.1253**	**0.1699**

	Scale-isolated					
Backbone	**w/ additional activation**			**w/o additional activation**		
	P2	**P3**	**P4**	**P2**	**P3**	**P4**
Vanilla [2]	-	-	-	0.7059	0.7469	0.8677
SESN [3]	0.2191	0.2313	0.3596	0.3773	0.4667	0.5573
SESN* [3]	0.1945	0.2078	0.3312	0.3471	0.4209	0.5148
DISCO [4]	0.1775	0.1621	0.2537	0.3423	0.3597	0.4177
SSE [1]	0.0952	0.0956	0.1586	0.1893	0.2325	0.2882
SSE* [1]	**0.0821**	**0.0797**	**0.1364**	**0.1614**	**0.1901**	**0.2387**

Table Note: * denotes with scale-adaptive spatial coverage kernel. **Red bold** indicates the lowest SE error across both with and without additional activation functions. **Black bold** represents the lowest SE error for each version, except the red bold.

Second, while the original authors of SESN and DISCO utilized scale-isolated basis functions, which are designed to maintain independent scale parameters across each dimension, the SSE implementation inadvertently employed scale-combined basis functions. Although the basis functions were initially computed following the scale-isolated principle, the deviation occurred during the final convolution kernel construction phase, where a tensor permutation that swaps the N_b and S dimensions was inadvertently applied. This modification results in a convolution kernel where basis functions from different scales are mixed within each scale dimension,

TABLE II
PERFORMANCE COMPARISON OF SCALE-COMBINED AND SCALE-ISOLATED METHODS FOR OBJECT DETECTION ON THE KITTI DATASET. E, M, AND H DENOTE *Easy*, *Moderate*, AND *Hard* RESPECTIVELY.

Model	Method	Backbone	Car (IoU $\geq$ 0.7)			Pedestrian (IoU $\geq$ 0.5)			Cyclist (IoU $\geq$ 0.5)			mAP	Input size
			E	M	H	E	M	H	E	M	H		
Scale-combined methods													
Deformable DETR	Vanilla [5]	ResNet-50	85.78	81.78	76.70	64.09	60.03	55.71	60.38	58.93	56.73	66.68	188 x 621
	SESN [3]	SE_ResNet-50	87.83	86.55	80.91	72.81	69.49	65.30	74.73	70.16	68.48	75.14	188 x 621
	DISCO [4]	SE_ResNet-50	87.75	86.52	80.82	72.02	68.37	64.90	73.47	69.39	68.29	74.61	188 x 621
	SSE [1]	SSE_ResNet-50	**88.03**	**87.52**	**83.10**	**75.06**	**70.82**	**66.60**	**80.23**	**73.61**	**71.22**	**77.35**	188 x 621
Deformable DAB-DETR	Vanilla [6]	ResNet-50	86.05	76.98	71.76	67.16	63.29	58.43	64.15	58.11	56.34	66.92	188 x 621
	SESN [3]	SE_ResNet-50	**88.83**	**82.22**	**78.27**	76.41	72.55	67.85	77.93	70.22	68.96	75.92	188 x 621
	DISCO [4]	SE_ResNet-50	88.72	82.05	76.61	75.62	70.61	66.27	74.99	69.87	68.09	74.76	188 x 621
	SSE [1]	SSE_ResNet-50	88.65	81.31	77.77	**77.41**	**74.37**	**69.36**	**79.66**	**76.8**	**74.47**	**77.76**	188 x 621
DINO	Vanilla [7]	ResNet-50	89.47	86.00	79.71	72.83	67.70	63.20	67.10	62.24	60.63	72.10	188 x 621
	SESN [3]	SE_ResNet-50	90.55	88.26	83.70	75.77	71.91	68.09	**76.19**	**70.63**	**68.00**	77.01	188 x 621
	DISCO [4]	SE_ResNet-50	90.55	88.36	83.43	75.65	70.93	66.96	73.37	67.46	64.66	75.71	188 x 621
	SSE [1]	SSE_ResNet-50	**90.62**	**89.07**	**84.63**	**77.07**	**73.15**	**69.66**	75.41	68.53	65.79	**77.10**	188 x 621
Scale-isolated methods													
Deformable DETR	Vanilla [5]	ResNet-50	85.78	81.78	76.70	64.09	60.03	55.71	60.38	58.93	56.73	66.68	188 x 621
	SESN [3]	SE_ResNet-50	87.48	85.55	79.40	70.84	66.12	63.08	71.02	68.59	66.08	73.13	188 x 621
	DISCO [4]	SE_ResNet-50	88.16	87.42	82.48	73.35	69.80	66.13	76.74	70.44	68.32	75.87	188 x 621
	SSE [1]	SSE_ResNet-50	**88.79**	**88.08**	**82.60**	**77.28**	**73.41**	**69.35**	**80.00**	**74.00**	**70.84**	**78.26**	188 x 621
Deformable DAB-DETR	Vanilla [6]	ResNet-50	86.05	76.98	71.76	67.16	63.29	58.43	64.15	58.11	56.34	66.92	188 x 621
	SESN [3]	SE_ResNet-50	**89.77**	**85.38**	**78.56**	73.75	69.24	65.67	72.23	65.36	63.65	73.74	188 x 621
	DISCO [4]	SE_ResNet-50	88.67	85.00	78.44	**79.16**	**73.82**	**69.92**	76.46	72.06	69.51	77.01	188 x 621
	SSE [1]	SSE_ResNet-50	88.60	84.49	78.54	78.87	73.30	69.33	**79.03**	**75.10**	**71.80**	**77.67**	188 x 621
DINO	Vanilla [7]	ResNet-50	89.47	86.00	79.71	72.83	67.70	63.20	67.10	62.24	60.63	72.10	188 x 621
	SESN [3]	SE_ResNet-50	90.10	88.10	83.14	75.65	71.39	67.81	75.23	69.45	66.93	76.42	188 x 621
	DISCO [4]	SE_ResNet-50	**90.42**	88.69	**84.27**	76.07	71.88	68.52	79.34	69.89	67.98	77.45	188 x 621
	SSE [1]	SSE_ResNet-50	90.26	**88.96**	84.21	**78.04**	**73.92**	**70.88**	**80.33**	**72.75**	**69.84**	**78.80**	188 x 621

Table Note: **Red bold** indicates the best performance across both scale-combined and scale-isolated configurations. **Black bold** represents the best for each version, except the red bold.

creating a scale-combined representation. This specific change deviates from original intentions, particularly the spatial isolation intended by DISCO. This scale-combined approach was consistently applied across all comparison networks [3] and [4] as well as the proposed SSE throughout the experiments. However, Table I demonstrates that this deviation interestingly improved performance regarding SE error. The scale-combined basis implementation consistently outperforms the original scale-isolated formulation, achieving lower SE error across most cases, suggesting potential benefits warranting further investigation.

*b) **Additional activation functions:*** Additionally, ReLU activation functions were incorporated before residual connections throughout all scale-equivariant backbone architectures evaluated in the experiments. This modification was applied universally, except for the identity down-sampling layers; however, for the SSE method, the additional ReLU was also included in the high-frequency down-sampling layers. Empirical evaluation indicated that this modification contributed positively to reducing the overall SE error, as shown in Table I.

The comparative analysis in Table I reveals the distinction between scale-combined and scale-isolated basis functions as the most critical implementation modification, exhibiting substantial performance variations across different configurations. In contrast, other architectural modifications demonstrated relatively minor or universally beneficial effects. This unexpected improvement suggests that the scale-combined representation may better capture cross-scale feature relationships in practical applications, despite deviating from the theoretical framework. Consequently, the following section provides a comprehensive experimental evaluation focusing specifically on this key distinction, examining its effects on object detection and tracking capabilities within autonomous driving scenarios.

III. EXPERIMENTS

Comparative analyses of scale-combined and scale-isolated basis functions for object detection and tracking capabilities are presented in Tables II and III. The detection task follows the previously established setup in [1] for comparison. However, unlike the detection task, novel tracking experiments were conducted. This approach was necessary as prior tracking results relied on the performance of the integrated detection component; the interdependence between detection capability and tracking accuracy made it challenging to confirm the isolated tracking capability of the underlying methods. To specifically verify the detection-independent Re-Identification

TABLE III

Performance Comparison of Re-Identification Capabilities for Object Tracking on the KITTI Dataset.

	Re-ID Tracking Metrics				
	IDF1(↑)	IDR(↑)	IDP(↑)	IDsw(↓)	Frag(↓)
Scale-combined					
Vanilla [2]	83.26	80.98	85.67	123	108
SESN [3]	85.75	83.71	87.90	97	86
DISCO [4]	86.44	84.41	88.58	91	83
SSE [1]	**87.07**	**85.09**	**89.14**	**90**	**79**
Scale-isolated					
Vanilla [2]	83.26	80.98	85.67	123	108
SESN [3]	84.79	82.74	86.94	104	91
DISCO [4]	86.65	84.62	88.79	94	84
SSE [1]	**87.35**	**85.34**	**89.47**	**89**	**81**

Table Note: **Red bold** indicates the best performance across both scale-combined and scale-isolated configurations. **Black bold** represents the best for each version, except the red bold.

(ReID) capability of scale-equivariant models, a simple ReID model integrated with pre-trained backbone networks was employed. Backbones from methods including [5], [6], and [7] were utilized under the same detection task configurations and kept frozen during ReID training. Features extracted from these frozen backbones were normalized and passed through a Feature Pyramid Network (FPN). These were subsequently processed by a simple ReID branch, adapted from [8], to generate tracking features for each bounding box region. To ensure detection independence, ReID features for target objects were generated using ground truth bounding boxes. These features were then integrated with the DeepSORT [9] tracker to measure the final track re-identification performance. The ReID branch was trained for 20 epochs with a batch size of 8, using a learning rate initialized at 1×10^{-3} and annealed to 1×10^{-5} via a cosine annealing schedule. This methodology provides a cleaner assessment of the backbone's intrinsic feature quality for the re-identification task, independent of detection capability. The results of this assessment are presented in Table III, which details the average Top-5 ReID tracking performance for each detection model [5]–[7] backbone network comparing their vanilla versions against different scale-equivariant methods.

As demonstrated in Tables II and III, the scale-isolated basis function implementation, which aligns with the mathematical design intent, generally achieved superior performance across most object detection and tracking ReID tasks compared to the scale-combined basis function. While the scale-combined basis function demonstrated a lower SE error in the preliminary analysis, the superior performance of the scale-isolated approach indicates its ability to generate more discriminative features for object perception with scale-equivariant methods. Nevertheless, given the scale-combined basis function's comparable performance and its demonstrated lower SE Error, this implementation demonstrates merit. Therefore, exploring methods to effectively leverage this novel approach remains a valuable direction for future research.

IV. Conclusion

Presented a comprehensive supplementary analysis detailing implementation deviations in the scale-combined scale-equivariant backbone network from the core theoretical principles of previous scale-equivariant benchmarks. These issues, including the inadvertent use of scale-combined basis functions and fixed spatial coverage kernel sizing, arose from implementation specifics and a potential lack of full adherence to the original mathematical frameworks. While this unintended scale-combined approach interestingly achieved a lower SE error in preliminary analysis, the original scale-isolated formulation generally yielded superior performance across practical object detection and tracking Re-ID metrics. Despite its lower practical performance, the scale-combined implementation's ability to minimize SE error demonstrates merit, positioning the effective exploration of this novel approach as a valuable direction for future research.

References

[1] T. Cho, H. Nam, and J. Choi, "Scale-equivariant object perception for autonomous driving," *IEEE Transactions on Intelligent Vehicles*, vol. 10, no. 9, pp. 4361–4370, 2025.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[3] I. Sosnovik, M. Szmaja, and A. Smeulders, "Scale-equivariant steerable networks," in *International Conference on Learning Representations*, 2020.

[4] I. Sosnovik, A. Moskalev, and A. W. M. Smeulders, "DISCO: accurate discrete scale convolutions," in *32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021*. BMVA Press, 2021, p. 334.

[5] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: deformable transformers for end-to-end object detection," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

[6] S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, and L. Zhang, "DAB-DETR: Dynamic anchor boxes are better queries for DETR," in *International Conference on Learning Representations*, 2022.

[7] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. Ni, and H.-Y. Shum, "DINO: DETR with improved denoising anchor boxes for end-to-end object detection," in *The Eleventh International Conference on Learning Representations*, 2023.

[8] Y. Zhang, C. Wang, X. Wang, W. Zeng, and W. Liu, "Fairmot: On the fairness of detection and re-identification in multiple object tracking," *International Journal of Computer Vision*, vol. 129, pp. 3069–3087, 2021.

[9] N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 3645–3649.