

TIE-KD: Teacher-Independent and Explainable Knowledge Distillation for Monocular Depth Estimation

Sangwon Choi, Daejune Choi, Duksu Kim

Korea University of Technology and Education (KOREATECH)

Abstract—Monocular depth estimation (MDE) is essential for numerous applications yet is impeded by the substantial computational demands of accurate deep learning models. To mitigate this, we introduce a novel Teacher-Independent Explainable Knowledge Distillation (TIE-KD) framework that streamlines the knowledge transfer from complex teacher models to compact student networks, eliminating the need for architectural similarity. The cornerstone of TIE-KD is the Depth Probability Map (DPM), an explainable feature map that interprets the teacher’s output, enabling feature-based knowledge distillation solely from the teacher’s response. This approach allows for efficient student learning, leveraging the strengths of feature-based distillation. Extensive evaluation of the KITTI dataset indicates that TIE-KD not only outperforms conventional response-based KD methods but also demonstrates consistent efficacy across diverse teacher and student architectures. The robustness and adaptability of TIE-KD underscore its potential for applications requiring efficient and interpretable models, affirming its practicality for real-world deployment. The code and pre-trained models associated with this research are available at [here](#).

Index Terms—lightweight deep learning, knowledge distillation, explainable feature map, depth estimation

1 INTRODUCTION

Monocular depth estimation (MDE) is pivotal in computer vision, with applications ranging from autonomous vehicles [2] to robotics [3] and 3D modeling [4]. The integration of deep learning has notably enhanced MDE accuracy and efficiency [5], [6]. However, state-of-the-art models like SQLdepth [7], with 242 million parameters, pose challenges for real-time applications due to their computational demands. Methods like parameter pruning [8], [9], low-rank factorization [10], and compact convolution filters [11] aim to alleviate this.

Knowledge Distillation (KD) is another strategy that efficiently condenses the knowledge from larger models into more compact ones [12]. Initially prevalent in classification tasks [12], [13], [14], [15], [16], [17], [18], KD has expanded into other domains such as object detection [19], visual odometry [20], and so on. KD approaches are typically divided into response-based [12], [18], leveraging teacher outputs, and feature-based [13], [14], [15], [16], [17], where the student mimics the teacher’s feature maps, often resulting in superior performance. However, feature-based KD presents alignment challenges and typically requires similar network architectures between the teacher and student, unlike the more flexible response-based KD.

KD has been adapted for depth estimation [21], [22], [23], [24], effectively transferring knowledge from teacher to student models. Despite their successes, these approaches are constrained by feature-based KD limitations, requiring knowledge of the teacher’s architecture and meticulous

feature map matching between teacher and student.

The fundamental question driving our research asks if it is possible to harness the advantages of feature-based KD using only the teacher’s response (Sec. 3.1). In response, we propose a novel knowledge distillation framework for monocular depth estimation called Teacher-Independent Explainable KD (TIE-KD). Our method affords freedom from architectural constraints between teacher and student models by introducing an explainable feature map, the Depth Probability Map (DPM), generated directly from the teacher’s depth map output (Sec. 3.2). Furthermore, we elaborate on a teacher-independent KD process that capitalizes on the DPM, incorporating two specially designed loss functions to ensure efficient knowledge transfer (Sec. 3.3).

We validate our TIE-KD framework using three architecturally diverse teacher models, underscoring its robustness and adaptability (Sec. 4). Utilizing the KITTI dataset, we demonstrate that TIE-KD consistently outperforms traditional response-based KD methods (Sec. 4.3). Moreover, TIE-KD shows remarkable flexibility in accommodating various backbone architectures within student models (Sec. 4.5.2). An examination of the similarity between teacher and student outputs (Sec. 4.4) reveals a closer alignment for TIE-KD-trained pairs, confirming the method’s effectiveness in distilling knowledge (Fig. 1). These findings collectively affirm the efficacy of our TIE-KD approach in monocular depth estimation.

In summary, our main contribution are the following:

- Introduction of an innovative KD framework that operates independently of the teacher model’s architecture.
- Utilization of explainable Depth Probability Map

• Sangwon Choi, Daejune Choi, Duksu Kim are with the Department of Computer Engineering, Korea University of Technology and Education (KOREATECH), Cheonan, Korea, 31253.

E-mail: see <http://hpc.koreatech.ac.kr>

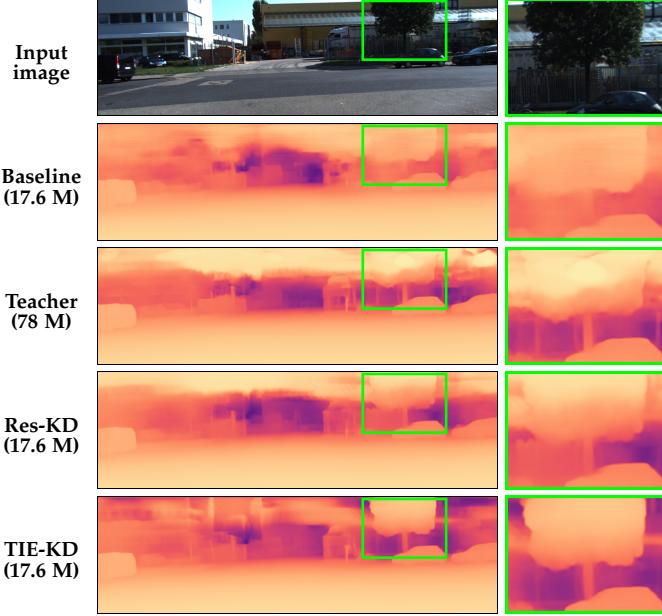


Fig. 1: Comparative visualization of depth estimation results showcasing the effectiveness of the proposed TIE-KD framework. The first row displays input images; the second row depicts outcomes from a baseline small model trained on ground truth. The third row shows results from the high-capacity teacher model, AdaBins [1]. The fourth and fifth rows illustrate depth maps from students trained via a response-based KD and our TIE-KD, respectively. Our TIE-KD demonstrates a more effective knowledge distillation performance than prior response-based KD methods, achieving greater similarity to the teacher model, especially in preserving edge definition and depth accuracy.

(DPM) derived from the teacher’s output, enhancing the interpretability and efficiency of KD.

- Superior performance over traditional response-based KD methods, as confirmed by rigorous testing on the KITTI dataset.
- Proven adaptability and effectiveness across varying student model backbones, demonstrating the framework’s versatility.

2 RELATED WORK

Knowledge Distillation enables lightweight models to enhance their performance by emulating more complex models. It was first demonstrated by Ba and Caruana [25] and formally defined by Hinton et al. [12], where the latter introduced a temperature-scaled softmax to transfer soft target probabilities via cross-entropy loss. Beyond logits, feature map matching was pioneered by Romero et al. [13], while Zagoruyko and Komodakis [14] introduced attention maps for distillation, and Heo et al. [16] focused on the activation boundaries within neuron activations. Recently, Zaho et al. [18] presented an innovative method for decoupling logits in classification tasks, further diversifying the applicability of knowledge distillation. This technique has been extended to various domains, including object detection [19], visual odometry [20], and so on.

Monocular Depth Estimation (MDE) is the process of predicting the depth for each pixel in an image. It is naturally a per-pixel regression problem, regression-based models have been extensively explored [5], [26], [27], [28], [29], [30]. Despite their effectiveness, these approaches can suffer from slow convergence and sub-optimal solutions [31], [32]. Addressing these limitations, Fu et al. [31] introduced DORN, a per-pixel classification-based MDE framework, assigning depth ranges to classes to improve efficiency. Dias and Marathe [33] further refined this approach by introducing soft targets during training, enhancing the model’s ability to generalize. This classification concept has been widely adopted in subsequent research [34], [35], [36], but it often results in decreased visual quality due to quantization effects. The AdaBins model [1] presents a solution to this challenge by utilizing adaptive bins for depth intervals, leading to more accurate depth predictions through a weighted sum of bin centers and bin’s probabilities. Li et al. [32] built upon this hybrid classification-regression approach by integrating a Transformer decoder for bin generation, pushing the boundaries of classification-regression MDE.

Our proposed KD framework aligns with this classification-regression paradigm while maintaining the versatility to work with teacher models outside this framework.

Knowledge Distillation for Depth Estimation has been addressed in various studies [21], [22], [23], [24]. Pilzer et al. [21] pioneered knowledge distillation in unsupervised monocular depth estimation, employing a self-distillation structure within a single model. Wang et al. [22] evolved from pixel-wise to pair-wise distillation, inspired by segmentation techniques [37], enabling student models to emulate teacher feature maps more effectively. Song et al. [23] leveraged stereo-based teacher models and introduced selective distillation for multi-scale feature maps from student encoders. Hu et al. [24] tackled the challenge of capacity disparities between student and teacher models by integrating auxiliary unlabeled data into the distillation process, a departure from traditional methods that focus solely on depth map features.

Our KD approach distinguishes itself from the aforementioned methods by generating an interpretable feature map directly from the teacher model’s depth map output. This not only simplifies the distillation process but also provides an explicit representation of the distilled knowledge, paving the way for a more transparent and potentially more generalizable learning paradigm for depth estimation.

3 METHODOLOGY

3.1 Motivation

Knowledge distillation techniques fall into two primary categories: response-based and feature-based KD.

In response-based KD, the student model is trained to mimic the teacher model’s output (i.e., response). This method permits varying architectures between the teacher and student models, as the alignment is exclusively based on the teacher’s output. The output could be a scalar value like pixel depth (hard label) or probability logits in classification tasks (soft label). When hard labels approximate the

ground truth (GT), the process resembles conventional training using the GT. Soft labels, however, allow for knowledge transfer and induce regularization [12].

In feature-based KD, the emphasis shifts to replicating the teacher’s feature maps, veering away from the response-based approach. This method inherently provides a regularization effect and often boosts the student model’s performance [13], [14], [16]. However, it usually benefits from similar architectures between the teacher and student models and faces challenges in aligning layers or feature maps effectively [38]. This is because it is difficult to define what specific knowledge a given feature map contains.

Depth estimation models output hard labels, in the form of depth maps, complicating the use of response-based KD in this domain. While feature-based KD can be applied, it is still encumbered by the limitations inherent to feature-based KD. Motivated by these challenges, we formulate the central question of our study: ‘*Can the advantages of feature-based KD be replicated using only the teacher’s response?*’ To address this, we propose a novel method to generate feature maps from the teacher’s response, circumventing the need for extracting feature maps from the teacher model. We also propose a knowledge distillation process that utilizes these generated feature maps.

3.2 Explainable Depth Probability Map

The conversion of depth regression tasks into classification problems was first pioneered by Fu et al. [31]. This methodology was later refined by AdaBins [1], which introduced adaptive binning. Such classification-based strategies yield logits for each depth bin, thereby generating soft labels that are particularly beneficial for knowledge distillation. Consistent with these developments, our work leverages this classification-based approach to maximize the advantages of using soft labels in the context of knowledge distillation.

Previous classification-regression-based techniques [1], [32] decode the final depth value by a specific decoding function like a weighted sum of logits along with the center values of the corresponding bins. However, we observe that the logit value for a specific bin does not unambiguously indicate the likelihood of that bin representing the true depth value. Various combinations of weights and center values can yield identical results, leading to a misalignment with the common expectation that the bin with the highest logit should naturally correspond to the actual depth. This method of representation is both model-dependent and lacks intuitive interpretability, making it less suitable for response-based KD methods.

To address these issues, we introduce an easily interpretable feature map known as the *depth probability map*. This map is filled with probabilities that directly quantify the likelihood of each bin’s contribution to the final depth value, offering a more transparent and model-independent representation.

Yuan et al. [39] proposed a label-smoothing technique that distributes probabilities uniformly across all classes while maintaining a higher probability for the class representing the GT. Although effective in general classification problems, this method loses its efficacy in depth estimation tasks. For depth estimation, the probability values in adjacent bins are not isolated but have a strong relationship

owing to the continuous nature of depth. We address this shortcoming by allocating higher probabilities to bins that are more proximate to the GT bin, thereby taking advantage of the inherent continuity in depth values.

To realize this idea, we uniformly divide the depth range into k bins and apply a Gaussian-like function, expressed in Eq. 1, to softly distribute probabilities around the GT depth value.

$$f(x, GT) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-GT}{\sigma})^2} \quad (1)$$

The probability assigned to the i -th bin B_i is computed using integration, as presented in Eq. 2:

$$p(B_i) = \int_{B_i} f(x, GT) dx \quad (2)$$

To mitigate the dilution of probabilities in the core region due to extremely low values, we employ a cut-off threshold, such as 10^{-16} . The probabilities of the remaining bins are then normalized using a softmax function.

Lastly, we apply this formulation to each pixel in the teacher’s response, generating a depth probability map. The resulting feature map’s dimension is $[H, W, B]$, where H and W refer to the height and width of the input image, and B denotes the number of bins.

3.3 Teacher-Independent Knowledge Distillation

Leveraging our explainable depth probability map, we propose a novel knowledge distillation method that is agnostic to the teacher model architecture while also benefiting from feature-based KD techniques. The overview of our approach is illustrated in Fig. 2.

Student network The student model features a flexible architecture, combining a backbone network with an encoder-decoder configuration. Its final output is a Depth Probability Map (DPM), tailored to capture depth information efficiently. The backbone is adaptable, allowing for various architectures and sizes, and the DPM is dimensioned as $[H, W, B]$. The depth estimation for each pixel is computed from the DPM, where the final depth value is derived through a weighted sum of the probabilities and the corresponding bin center depths. This process is mathematically represented as:

$$d = \sum_i center(B_i) \times p(B_i) \quad (3)$$

3.3.1 KD process

Both the teacher and student networks receive the same input image and produce depth maps as outputs. The teacher network’s output is a depth map, denoted as $Depth_T$, which is then converted into a depth probability map (DPM_T) using Eq. 2. On the other hand, the student network’s last layer outputs a depth probability map (DPM_S), which is subsequently decoded into a depth map ($Depth_S$). Thus, we obtain depth probability maps and depth maps from both models.

To optimize the student model, we employ two distinct loss functions: L_{DPM} for the depth probability map and L_{depth} for the depth map. The overall loss, L , is a composite of these two functions, weighted and scaled as expressed in Eq. 4. Here, α represents the weight assigned to L_{DPM} , and

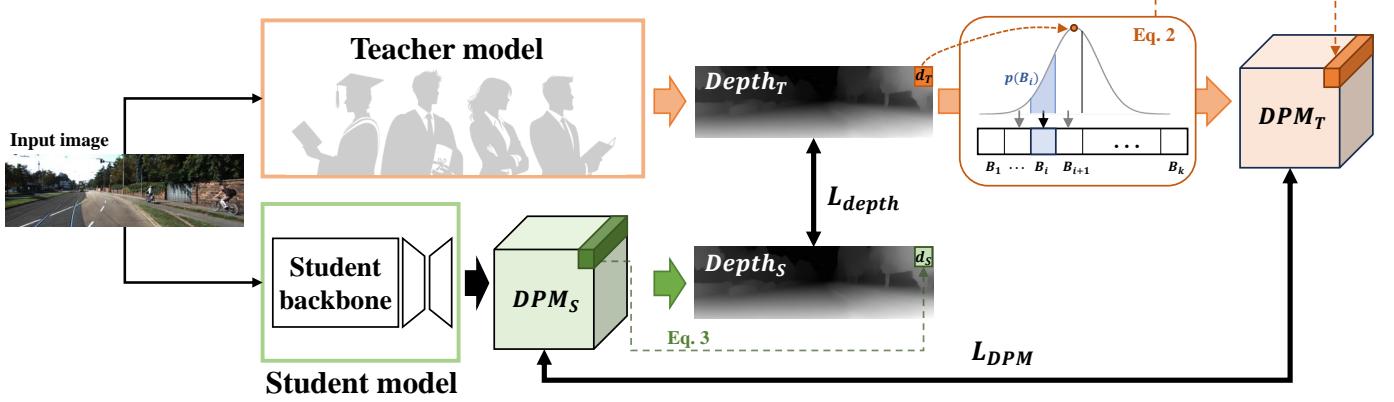


Fig. 2: Overview of our teacher-independent and explainable KD process for single image depth estimation

β serves as a scaling factor to adjust the overall magnitude of the loss.

$$L = \beta(\alpha L_{DPM} + (1 - \alpha)L_{depth}) \quad (4)$$

3.3.2 Depth Probability Map Loss

The L_{DPM} term measures the divergence between DPM_T and DPM_S using Kullback-Leibler (KL) divergence at the pixel level. The overall loss is an average of these pixel-wise divergences, as delineated in Eq. 5.

$$L_{DPM} = \frac{1}{M} \sum_{m \in M} P_S(m) \log \left(\frac{P_S(m)}{P_T(m)} \right) \quad (5)$$

Here, M represents the set of all pixels, and $P_S(m)$ and $P_T(m)$ are the probabilities from the student's and teacher's depth probability maps for pixel m , respectively.

3.3.3 Depth Map Loss

The L_{depth} term quantifies the dissimilarity between depth maps $Depth_T$ and $Depth_S$, utilizing the Structural Similarity Index (SSIM) for this comparison. The loss is calculated as specified in Eq. 6.

$$L_{depth} = 1 - \text{SSIM}(Depth_T, Depth_S) \quad (6)$$

This loss serves to minimize the discrepancy between the teacher's and the student's depth maps, thereby facilitating effective learning for the student model.

4 EXPERIMENTS

Our comprehensive experiments, conducted using three unique teacher models, validate the efficacy of the TIE-KD framework. In the following sections, we describe the datasets, performance metrics, and experimental details (Sec. 4.1 and Sec. 4.2). We then compare TIE-KD's effectiveness against response-based KD methods, highlighting its robustness across different architectures (Sec. 4.3 and Sec. 4.4). Ablation studies on loss functions, student backbones, and hyper-parameters (Sec. 4.5) further illustrate TIE-KD's flexibility and robustness.

4.1 Dataset and evaluation metrics

4.1.0.1 Dataset: For our experiments, we utilized the well-established KITTI dataset, which provides stereo images and corresponding 3D LiDAR point clouds of street scenes [2]. The RGB images are of high resolution (1241×376 pixels), and the depth maps capture distances up to 80 meters. Following the protocol established by Eigen et al. [5], we trained our models on a set of approximately 26K left camera images and validated on a separate test set comprising 697 images. During training, we augmented the data by randomly cropping the images to a resolution of 704×352 pixels.

4.1.0.2 Evaluation metrics: Our assessment adheres to the established evaluation protocol for depth estimation, following the precedent set by prior work [5], [21], [23]. Specifically, we measure performance using the Absolute Relative Error (AbsRel), Squared Relative Error (SqRel), Root Mean Squared Error (RMSE), and Root Mean Squared Logarithmic Error (RMSE_{log}). Additionally, we evaluate the accuracy under threshold, defined by the metric $\delta_i < 1.25^i$ for $i = 1, 2, 3$, to capture the proportion of depth estimates within specified error bounds.

4.2 Implementation Details

Our TIE-KD framework and five distinct response-based KD methodologies were developed using PyTorch. For all methods, including our own and comparative approaches, we utilized the Adam optimizer with $\beta_1 = 0.95$ and $\beta_2 = 0.99$, implementing a one-cycle policy with a peak learning rate of 1e-3. For TIE-KD, we assign values of 0.1 and 10 to the weight parameters α and β , respectively, in Eq. 4. Additionally, the σ in Eq. 1 is set to 0.8. Models were trained for 24 epochs with a batch size of 32, selecting the weights from the epoch that yielded the best performance.

- Student Network:** Our student model utilizes a UNet-inspired encoder-decoder structure [40], with MobileNetV2 [41] serving as the backbone. The final layer outputs a DPM with dimensions of $[704, 352, 257]$, thereby segmenting the depth range from 0 to 80 meters into 257 distinct bins. The total model size depends on the chosen backbone, adjusted accordingly within the encoder and decoder

segments. Overall, the student network comprises roughly 17.6 million parameters.

- **Teacher models:** For our teacher models, we utilized three well-known monocular depth estimation architectures: AdaBins [1], BTS [27], and DepthFormer [30], with parameter counts of 78 million, 48 million, and 273 million, respectively. We utilized pre-trained models obtained from a public repository¹.
- **Baseline:** The baseline model shares the same architecture as the student network but replaces the DPM layer with a regression layer, resulting in a similar parameter count of around 17.6 million. The baseline was trained from scratch on the KITTI dataset using the scale invariant loss (SI) introduced by Eigen et al. [5].
- **Response-based KD (Res-KD):** We implemented general response-based KD methods suitable for use with any teacher model architecture, in contrast to feature-based approaches. These methods entailed comparing depth maps generated by the teacher and student models, utilizing a range of loss functions including SSIM, MSE (Mean Square Error), and SI. Combinations of SSIM with SI, and SSIM with MSE, were also evaluated.

We noted that teacher models often show less consistent performance in the upper image regions, likely due to the sparsity of LiDAR data in the KITTI dataset, particularly in distant areas like the sky or mountains. To address this and improve the reliability of our KD process, we excluded the top 110 pixels from the height of the images for loss computation. This strategy focuses our training on regions with denser and more reliable depth data, ensuring a more consistent and dependable dataset.

4.3 Comparison with Response-Based KD Methods

Table 1 presents the performance of baseline models, a range of teacher models, and student models that have been trained using different KD methods. Students trained under our TIE-KD framework consistently surpassed the performance of both the baseline and other response-based KD methods across all metrics, independent of the teacher model’s architecture. This achievement is noteworthy considering that TIE-KD students were not trained with GT data from the KITTI dataset, in contrast to the baseline model which was trained directly on the target dataset.

Among the response-based KD methods, those utilizing SSIM loss, or a blend of SSIM and SI loss, were typically the most effective. Nevertheless, these methods often did not outperform the baseline, even when the students were instructed by highly capable teachers. Our TIE-KD approach, in comparison, consistently exhibited superior performance, demonstrating its robust capacity to effectively harness the knowledge conveyed by teacher models.

Among the student models trained with three distinct teachers, the student instructed by AdaBins generally outperformed its counterparts. It achieved up to a 7.4% improvement (e.g., in SqRel) and an average enhancement of

4.5% across the four performance metrics (AbsRel, SeRel, RMSE, and RMSE_{\log}) when compared to the baseline. Interestingly, the student trained under the guidance of DepthFormer, despite its larger model size and superior performance, did not exceed the performance of those trained by AdaBins. This result aligns with the findings of Mirzadeh et al. [42], which suggest that a large disparity in parameters between teacher and student models can negatively impact performance.

4.3.0.1 Qualitative Comparison: Fig. 3 presents a visual comparison of depth maps generated by five models: the baseline, the teacher, and students trained with various KD methods. The GT from the KITTI dataset is not depicted as it consists of sparse LiDAR points, which differ significantly from the continuous depth map representation.

The depth map from the baseline model is significantly blurrier, particularly around object edges, than that produced by the teacher model. In contrast, the student model utilizing our TIE-KD framework presents depth estimations remarkably similar to the teacher’s, with improved edge definition. This improvement is exemplified in the middle image of Fig. 3, where the TIE-KD_{AdaBins} model delineates traffic sign boundaries with greater clarity than the teacher, as evidenced by visual comparisons.

Depth maps from students trained via response-based KD methods are closer to the teacher’s output than the baseline but still fall short of the fidelity achieved by TIE-KD, particularly in capturing the fine details at object boundaries and transitions.

4.4 Similarity to the Teacher Model

A crucial goal of knowledge distillation is the accurate transfer of knowledge from the teacher to the student model. To evaluate how well our TIE-KD framework preserves the teacher’s knowledge, we compared the similarity between the outputs of the teacher and the student models using three metrics: AbsRel, RMSE, and δ_1 . For this evaluation, we excluded the top 110 pixels, consistent with the rationale provided in Sec. 4.2. The results, presented in Table 2, show that students trained with TIE-KD more closely mirror their respective teacher models than those trained by other teachers not involved in their training process. While Res-KD demonstrates some degree of correlation with the teachers’ outputs, the TIE-KD students generally exhibit a higher level of similarity. These results underscore the efficacy of the TIE-KD framework in faithfully transferring the teacher’s knowledge to the student model.

Notably, TIE-KD_{DepthFormer} and Res-KD_{DepthFormer} students showed less alignment with their respective teacher model. This divergence is likely attributed to the substantial differences in parameter counts, consistent with the findings of Mirzadeh et al. [42]. However, when applying the TIE-KD framework to a student model equipped with a larger ResNet50 backbone, which encompasses approximately 78M parameters, there is a notable increase in the similarity of the student model’s output to that of the DepthFormer. This observation, as presented in Table 3, suggests that minimizing the disparity in parameters between teacher and student models can potentially improve the efficiency of knowledge transfer in the TIE-KD framework.

1. <https://github.com/zhyever/Monocular-Depth-Estimation-Toolbox>

TABLE 1: Comparative evaluation of depth estimation performance across various models. Bold values indicate the best performance among student models, and underlined values denote the second-best methods.

Teacher model	Method (# of parameters)	Loss function(s)	Lower is better (\downarrow)				Higher is better (\uparrow)			
			AbsRel	SqRel	RMSE	RMSE _{log}	δ_1	δ_2	δ_3	
Baseline (17.6 M)			0.0663	0.2340	2.5625	0.1017	0.9501	0.9926	0.9984	
AdaBins [1]	Teacher (78 M)		0.0593	0.1941	2.3309	0.0901	0.9631	0.9946	0.9990	
	Res-KD	SSIM	0.0697	0.2407	2.5639	0.1041	0.9457	0.9933	0.9985	
		MSE	0.0786	0.2793	2.6964	0.1155	0.9319	0.9911	0.9981	
		SI	0.0739	0.2747	2.7371	0.1112	0.9382	0.9916	0.9980	
		SSIM,SI	0.0701	0.2445	2.5833	0.1047	0.9458	0.9932	0.9985	
	TIE-KD		L_{DPM}, L_{depth}	0.0654	0.2179	2.4315	0.0980	0.9540	0.9939	
BTS [27]	Teacher (47M)		0.0586	0.2060	2.4798	0.0916	0.9602	0.9940	0.9986	
	Res-KD	SSIM	0.0697	0.2460	2.6357	0.1050	0.9434	0.9930	0.9985	
		MSE	0.0820	0.2977	2.7440	0.1203	0.9263	0.9895	0.9977	
		SI	0.0782	0.2931	2.8106	0.1157	0.9346	0.9903	0.9977	
		SSIM,SI	0.0690	0.2462	2.6168	0.1044	0.9467	0.9928	0.9983	
	TIE-KD		L_{DPM}, L_{depth}	0.0656	0.2247	2.4984	0.0995	0.9523	0.9985	
DepthFormer [30]	Teacher (273 M)		0.0513	0.1511	2.1038	0.0783	0.9752	0.9970	0.9993	
	Res-KD	SSIM	0.0692	0.2337	2.5009	0.1019	0.9493	0.9937	0.9987	
		MSE	0.0805	0.2692	2.6029	0.1134	0.9361	0.9915	0.9983	
		SI	0.0724	0.2648	2.6717	0.1078	0.9411	0.9921	0.9984	
		SSIM,SI	0.0682	0.2382	2.5709	0.1018	0.9488	0.9937	0.9987	
	TIE-KD		L_{DPM}, L_{depth}	0.0657	0.2208	2.4402	0.0980	0.9534	0.9940	

TABLE 2: Comparative analysis of the similarity between teacher and student models based on their outputs. Student models are identified by the subscript indicating their respective teacher models used in the KD process.“Res-KD” denotes the response-based KD method using SSIM and SI loss functions. The best performances, marked in bold, signify the closest alignment with the teacher’s output.

Evaluation target	AdaBins’s output			BTS’s ouput			DepthFormer’s output		
Metric	AbsRel \downarrow	RMSE \downarrow	$\delta_1\uparrow$	AbsRel \downarrow	RMSE \downarrow	$\delta_1\uparrow$	AbsRel \downarrow	RMSE \downarrow	$\delta_1\uparrow$
Res-KD _{AdaBins}	0.0640	2.2512	0.9598	0.0675	2.3932	0.9564	0.0757	2.4273	0.9495
TIE-KD _{AdaBins}	0.0612	2.0905	0.9650	0.0669	2.3572	0.9573	0.0733	2.2698	0.9533
Res-KD _{BTS}	0.0673	2.5704	0.9528	0.0635	2.2174	0.9619	0.0757	2.6581	0.9482
TIE-KD _{BTS}	0.0627	2.4227	0.9588	0.0613	2.1385	0.9642	0.0726	2.5311	0.9523
Res-KD _{DepthFormer}	0.0667	2.4252	0.9569	0.0682	2.6015	0.9576	0.0710	2.2278	0.9580
TIE-KD _{DepthFormer}	0.0646	2.3254	0.9604	0.0683	2.5579	0.9565	0.0704	2.1605	0.9582

TABLE 3: Comparison of TIE-KD student models with ResNet50 backbone (78.3M) against various teacher model outputs

Evaluation target	AbsRel \downarrow	RMSE \downarrow	$\delta_1\uparrow$
AdaBins’s output	0.0561	2.1732	0.9675
BTS’s output	0.0559	2.3654	0.9718
DepthFormer’s output	0.0597	1.9085	0.9738

4.5 Ablation Study

4.5.1 Impact of Loss Functions

Table 4 demonstrates the impact of utilizing our two proposed loss functions within the TIE-KD framework. When applied independently, both L_{DPM} and L_{depth} achieved performances comparable to the baseline, with L_{depth} marginally outperforming L_{DPM} . However, the combined application of L_{DPM} and L_{depth} led to the most substantial performance gains. We also explored the influence of the weighting factor (α) between these two loss functions, varying α from 0.1 to 0.9. As a result, we found around 0.1 typically yields the best performance.

TABLE 4: Performance impact of different loss function configurations within the TIE-KD framework. Each row compares the outcomes when employing L_{DPM} , L_{depth} , or their combination. The best results are highlighted in bold.

L_{DPM}	L_{depth}	AbsRel \downarrow	SqRel \downarrow	RMSE \downarrow	$\delta_1\uparrow$
AdaBins					
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	0.0718	0.2485	2.5433	0.9398
		0.0696	0.2268	2.4646	0.9468
	<input checked="" type="checkbox"/>	0.0654	0.2179	2.4315	0.9540
BTS					
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	0.0722	0.2568	2.6459	0.9385
		0.0679	0.2315	2.5694	0.9477
	<input checked="" type="checkbox"/>	0.0656	0.2247	2.4984	0.9523
DepthFormer					
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	0.0713	0.2438	2.5241	0.9411
		0.0698	0.2299	2.4805	0.9458
	<input checked="" type="checkbox"/>	0.0657	0.2208	2.4402	0.9534

These results affirm our proposition that the depth probability map not only conveys the teacher’s knowledge more effectively to the student but also acts as a beneficial regularizer, significantly enhancing the student’s performance.

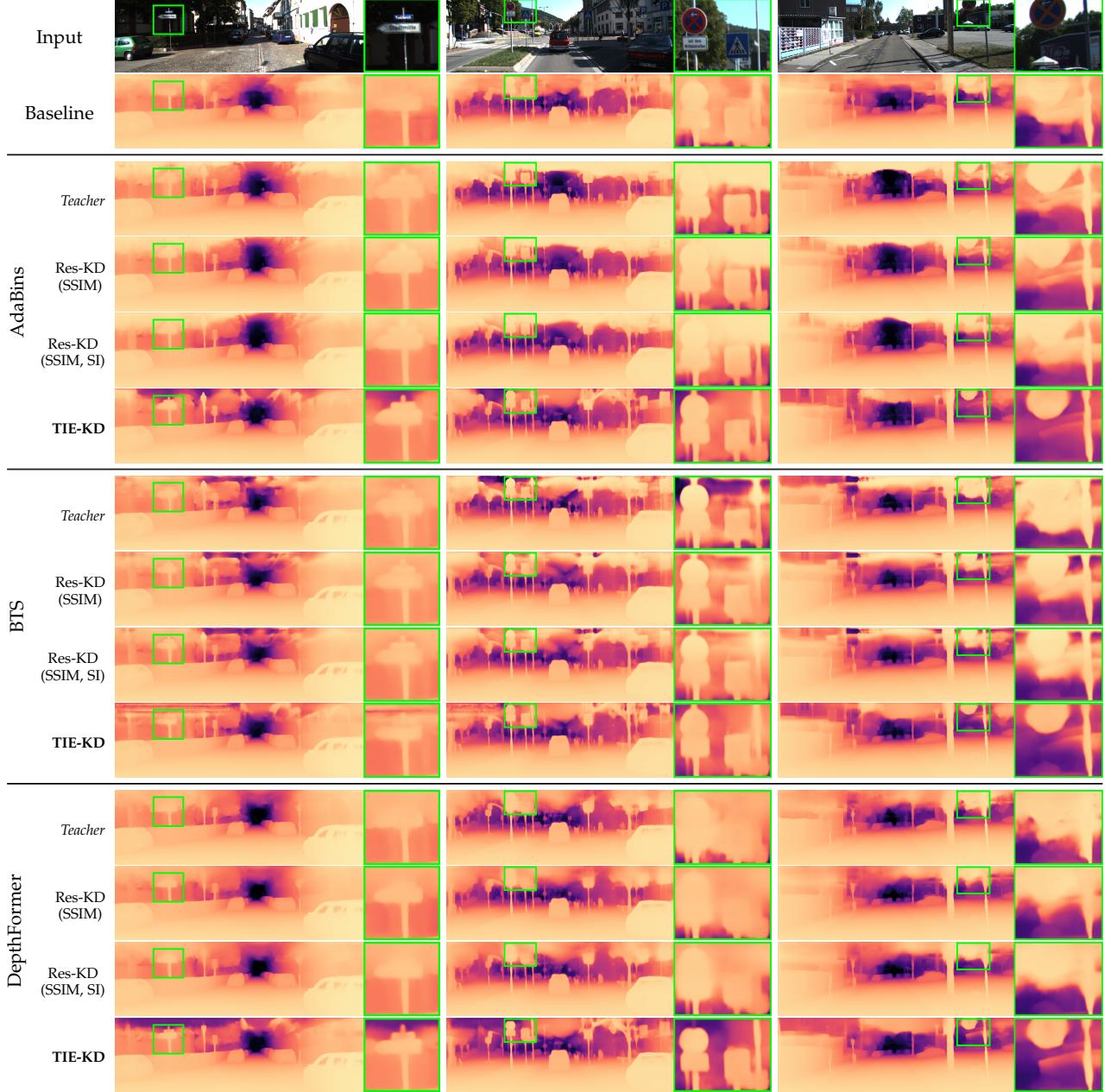


Fig. 3: Visual comparison of depth maps across various models for three different scenes, highlighting detailed variances within the regions enclosed by green boxes.

4.5.2 Flexibility Across Different Student Backbones

The adaptability of our TIE-KD framework to diverse network architectures was evaluated by employing alternative backbones for the student model. We incorporated ResNet architectures with varying capacities, specifically ResNet18 and ResNet50 [43], into our student models. The ResNet18-based student model was comparable in capacity to our original student model, whereas the ResNet50-based student model was the most capacious among those tested.

Performance outcomes, as delineated in Table 5, confirm the anticipated trend that increased backbone capacity correlates with improved baseline performance. Remarkably, the TIE-KD students, particularly those with the ResNet50 backbone, approached the performance of the teacher model (AdaBins). Moreover, across all backbone architectures, TIE-

KD-trained models consistently outperformed their baseline equivalents, attesting to the TIE-KD framework's effectiveness and its robustness with various student architectures.

4.5.3 Effect of the Weight (α) for the Loss Function

The performance of our TIE-KD framework with varying weight α is detailed in Table 6 and Fig. 4, as specified in Eq. 4 (Sec. 3.3). By systematically adjusting α between 0.05 to 0.7 in increments of 0.05 using AdaBins [1] as the teacher, we discovered that an α value of 0.1 generally yielded optimal results across most metrics.

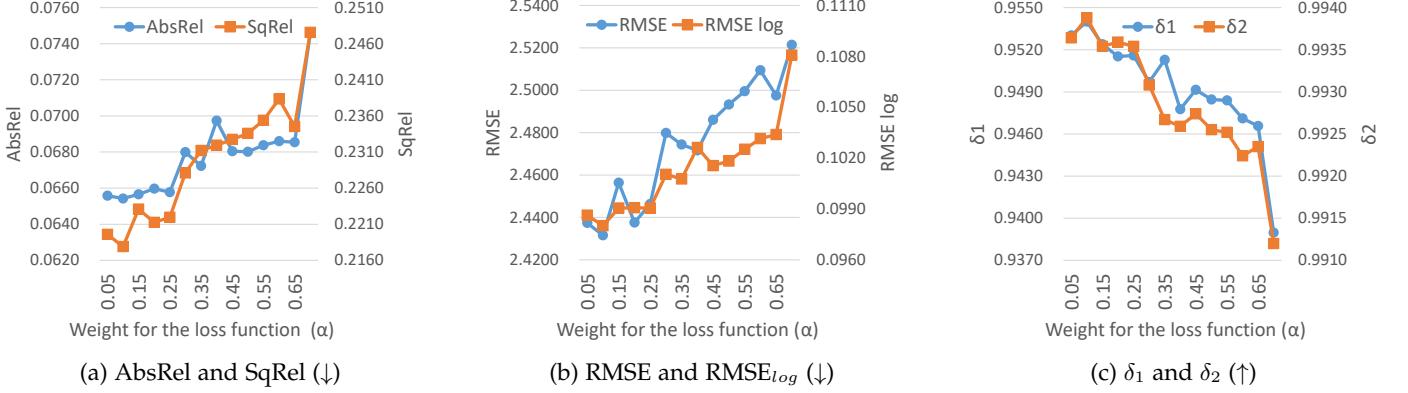


Fig. 4: Impact of loss function weight (α) on TIE-KD performance, with each subfigure representing a different metric.

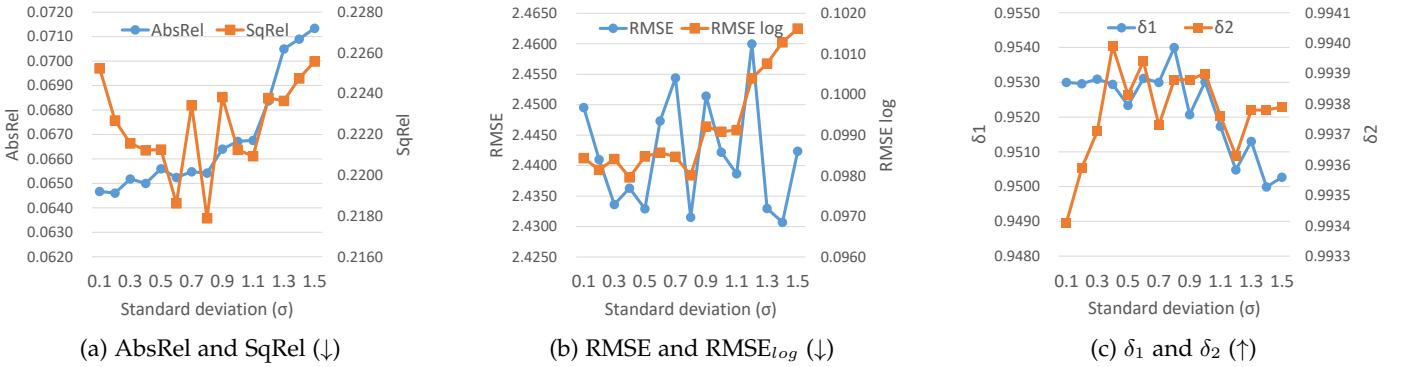


Fig. 5: Performance variation of the TIE-KD framework with respect to the standard deviation (σ) used in depth probability map generation. Each subfigure presents a different metric.

TABLE 5: Comparison of student network performance across different backbone architectures. The baseline is trained on the KITTI dataset, while the student is trained using the TIE-KD framework with AdaBins as the teacher. The term ‘model size’ refers to the total number of parameters of the student model including the backbone network.

Backbone (model size)	Method	AbsRel \downarrow	RMSE \downarrow	$\delta_1\uparrow$
MobileNetV2 (17.6M)	baseline	0.0663	2.5625	0.9501
MobileNetV2 (17.6M)	TIE-KD	0.0654	2.4315	0.9540
ResNet18 (16.7M)	baseline	0.0634	2.5311	0.9531
ResNet18 (16.7M)	TIE-KD	0.0628	2.4029	0.9559
ResNet50 (78.3M)	baseline	0.0605	2.4159	0.9576
ResNet50 (78.3M)	TIE-KD	0.0596	2.3060	0.9605
AdaBins (78M)	Teacher	0.0593	2.3309	0.9631

4.5.4 Effect of the Standard Deviation (σ) for Depth Probability Map Generation

Table 7 and Fig. 5 illustrate the performance variations in our TIE-KD framework relative to different standard deviation (σ) values used in Eq. 1 for generating the depth probability map (refer to Section 3.2). While optimal σ values varied across different evaluation metrics, a σ value of 0.1 generally yielded good overall performance.

5 CONCLUSION

In this work, we introduced the Teacher-Independent Explainable Knowledge Distillation (TIE-KD) framework, a

novel approach to monocular depth estimation. Central to TIE-KD is the Depth Probability Map (DPM), which enables an efficient distillation process that is not constrained by the architectural compatibility between teacher and student models. Rigorous testing on the KITTI dataset has shown that TIE-KD surpasses traditional response-based KD methods, demonstrating not only a closer alignment with the teacher model’s performance but also adaptability to various student model architectures.

Our aspirations for the TIE-KD framework are to pave the way for more effective and scalable solutions in monocular depth estimation, contributing significantly to the practical deployment of deep learning models, especially in settings where computational resources are limited.

REFERENCES

- [1] S. F. Bhat, I. Alhashim, and P. Wonka, “Adabins: Depth estimation using adaptive bins,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4009–4018.
- [2] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision meets robotics: The kitti dataset,” *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.
- [3] G. Dudek and M. Jenkin, *Computational principles of mobile robotics*. Cambridge university press, 2010.
- [4] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molnyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon, “Kinectfusion: Real-time dense surface mapping and tracking,” in *2011 10th IEEE international symposium on mixed and augmented reality*. Ieee, 2011, pp. 127–136.

TABLE 6: Effect of the weight (α) of Eq. 4 in the main paper (Sec. 3.3), with AdaBins [1] as the teacher model. Bold values indicate the best performance, and underlined values denote the second-best methods.

α	AbsRel(\downarrow)	SqRel(\downarrow)	RMSE(\downarrow)	RMSE _{log} (\downarrow)	$\delta_1(\uparrow)$	$\delta_2(\uparrow)$	$\delta_3(\uparrow)$
0.05	0.0656	<u>0.2196</u>	<u>2.4374</u>	0.0986	0.9530	0.9936	0.9985
0.1	0.0654	0.2179	2.4315	0.0980	0.9540	0.9939	0.9985
0.15	0.0657	0.2231	2.4564	0.0990	0.9524	0.9935	0.9984
0.2	0.0660	0.2213	2.4376	0.0991	0.9515	0.9936	0.9986
0.25	0.0658	0.2220	2.4464	0.0991	0.9516	0.9935	0.9986
0.3	0.0680	0.2281	2.4798	0.1010	0.9497	0.9931	0.9985
0.35	0.0672	0.2312	2.4744	0.1008	0.9513	0.9927	0.9983
0.4	0.0697	0.2319	2.4716	0.1026	0.9478	0.9926	0.9985
0.45	0.0680	0.2327	2.4861	0.1016	0.9491	0.9927	0.9985
0.5	0.0680	0.2336	2.4932	0.1018	0.9485	0.9926	0.9984

TABLE 7: Effect of the standard deviation (σ) of Eq. 1 in the main paper (Sec. 3.2), with AdaBins [1] as the teacher model. Bold values indicate the best performance, and underlined values denote the second-best methods.

σ	AbsRel(\downarrow)	SqRel(\downarrow)	RMSE(\downarrow)	RMSE _{log} (\downarrow)	$\delta_1(\uparrow)$	$\delta_2(\uparrow)$	$\delta_3(\uparrow)$
0.1	0.0647	0.2252	2.4495	0.0984	0.9530	0.9934	0.9984
0.2	0.0646	0.2227	2.4410	0.0981	0.9530	0.9936	0.9984
0.3	0.0652	0.2216	2.4336	0.0984	0.9531	0.9937	0.9984
0.4	0.0650	0.2212	2.4363	0.0980	0.9529	0.9940	0.9986
0.5	0.0656	0.2213	2.4329	0.0985	0.9523	0.9938	0.9987
0.6	0.0652	<u>0.2186</u>	2.4473	0.0986	<u>0.9531</u>	0.9939	0.9985
0.7	0.0655	0.2234	2.4544	0.0985	0.9530	0.9937	0.9985
0.8	0.0654	0.2179	<u>2.4315</u>	0.0980	0.9540	0.9939	0.9985
0.9	0.0664	0.2238	2.4514	0.0992	0.9521	0.9939	0.9985
1	0.0667	0.2213	2.4422	0.0991	0.9530	0.9939	0.9985
1.1	0.0668	0.2209	2.4387	0.0991	0.9517	0.9938	0.9986
1.2	0.0684	0.2238	2.4600	0.1004	0.9505	0.9936	0.9985
1.3	0.0705	0.2237	2.4330	0.1008	0.9513	0.9938	0.9985
1.4	0.0709	0.2248	2.4307	0.1013	0.9500	0.9938	0.9987
1.5	0.0713	0.2256	2.4424	0.1016	0.9503	0.9938	0.9986

- [5] D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” *Advances in neural information processing systems*, vol. 27, 2014.
- [6] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, “Deeper depth prediction with fully convolutional residual networks,” in *2016 Fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 239–248.
- [7] Y. Wang, Y. Liang, H. Xu, S. Jiao, and H. Yu, “Sqldepth: Generalizable self-supervised fine-structured monocular depth estimation,” *arXiv preprint arXiv:2309.00526*, 2023.
- [8] J. Wu, C. Leng, Y. Wang, Q. Hu, and J. Cheng, “Quantized convolutional neural networks for mobile devices,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4820–4828.
- [9] S. Han, J. Pool, J. Tran, and W. Dally, “Learning both weights and connections for efficient neural network,” *Advances in neural information processing systems*, vol. 28, 2015.
- [10] X. Yu, T. Liu, X. Wang, and D. Tao, “On compressing deep models by low rank and sparse decomposition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7370–7379.
- [11] S. Zhai, Y. Cheng, Z. M. Zhang, and W. Lu, “Doubly convolutional neural networks,” *Advances in neural information processing systems*, vol. 29, 2016.
- [12] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015.
- [13] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, “Fitnets: Hints for thin deep nets,” *arXiv preprint arXiv:1412.6550*, 2014.
- [14] S. Zagoruyko and N. Komodakis, “Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer,” *arXiv preprint arXiv:1612.03928*, 2016.
- [15] J. Kim, S. Park, and N. Kwak, “Paraphrasing complex network: Network compression via factor transfer,” *Advances in neural information processing systems*, vol. 31, 2018.
- [16] B. Heo, M. Lee, S. Yun, and J. Y. Choi, “Knowledge transfer via distillation of activation boundaries formed by hidden neurons,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 3779–3787.
- [17] D. Chen, J.-P. Mei, Y. Zhang, C. Wang, Z. Wang, Y. Feng, and C. Chen, “Cross-layer distillation with semantic calibration,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 8, 2021, pp. 7028–7036.
- [18] B. Zhao, Q. Cui, R. Song, Y. Qiu, and J. Liang, “Decoupled knowledge distillation,” in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2022, pp. 11 953–11 962.
- [19] Q. Li, S. Jin, and J. Yan, “Mimicking very efficient network for object detection,” in *Proceedings of the ieee conference on computer vision and pattern recognition*, 2017, pp. 6356–6364.
- [20] M. R. U. Saputra, P. P. De Gusmao, Y. Almalioglu, A. Markham, and N. Trigoni, “Distilling knowledge from a deep pose regressor network,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 263–272.
- [21] A. Pilzer, S. Lathuiliere, N. Sebe, and E. Ricci, “Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9768–9777.
- [22] Y. Wang, X. Li, M. Shi, K. Xian, and Z. Cao, “Knowledge distillation for fast and accurate monocular depth estimation on mobile devices,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2457–2465.
- [23] K. Song and K.-J. Yoon, “Learning monocular depth estimation via selective distillation of stereo knowledge,” *arXiv preprint arXiv:2205.08668*, 2022.
- [24] J. Hu, C. Fan, H. Jiang, X. Guo, Y. Gao, X. Lu, and T. L. Lam, “Boosting lightweight depth estimation via knowledge distillation,” in *International Conference on Knowledge Science, Engineering and Management*. Springer, 2023, pp. 27–39.
- [25] J. Ba and R. Caruana, “Do deep nets really need to be deep?” *Advances in neural information processing systems*, vol. 27, 2014.
- [26] C. Godard, O. Mac Aodha, and G. J. Brostow, “Unsupervised monocular depth estimation with left-right consistency,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 270–279.
- [27] J. H. Lee, M.-K. Han, D. W. Ko, and I. H. Suh, “From big to small: Multi-scale local planar guidance for monocular depth estimation,” *arXiv preprint arXiv:1907.10326*, 2019.

- [28] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 179–12 188.
- [29] J. Zhao, K. Yan, Y. Zhao, X. Guo, F. Huang, and J. Li, "Transformer-based dual relation graph for multi-label image recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 163–172.
- [30] Z. Li, Z. Chen, X. Liu, and J. Jiang, "Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation," *Machine Intelligence Research*, pp. 1–18, 2023.
- [31] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2002–2011.
- [32] Z. Li, X. Wang, X. Liu, and J. Jiang, "Binsformer: Revisiting adaptive bins for monocular depth estimation," *arXiv preprint arXiv:2204.00987*, 2022.
- [33] R. Diaz and A. Marathe, "Soft labels for ordinal regression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4738–4747.
- [34] L. Liebel and M. Körner, "Multidepth: Single-image depth estimation via multi-task regression and classification," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 1440–1447.
- [35] A. Johnston and G. Carneiro, "Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2020, pp. 4756–4765.
- [36] M. H. Phan, S. L. Phung, and A. Bouzerdoum, "Ordinal depth classification using region-based self-attention," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 3620–3627.
- [37] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2604–2613.
- [38] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, pp. 1789–1819, 2021.
- [39] L. Yuan, F. E. Tay, G. Li, T. Wang, and J. Feng, "Revisiting knowledge distillation via label smoothing regularization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3903–3911.
- [40] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.
- [41] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [42] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 04, 2020, pp. 5191–5198.
- [43] 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.



Daejune Choi was received the Bachelor's degree with the School of Computer Engineering, Korea University of Technology and Education (KOREATECH). His research interests include lightweight deep learning, knowledge distillation, single-image super-resolution.



Duksu Kim is currently an assistant professor in the School of Computer Engineering at KOREATECH (Korea University of Technology and Education). He received his B.S. from SungKyunKwan University in 2008. He received his Ph.D. from KAIST (Korea Advanced Institute of Science and Technology) in Computer Science in 2014. He spent several years as a senior researcher at KISTI National Supercomputing Center. His research interest is designing heterogeneous parallel computing algorithms for various applications, including proximity computation, scientific visualization, and machine learning. Some of his work received the distinguished paper award at Pacific Graphics 2009, and an ACM student research competition award in 2009, and was selected as the spotlight paper for the September issue of IEEE Transactions on Visualization and Computer Graphics (TVCG) in 2013. He is a young professional member of IEEE and a professional member of ACM.



Sangwon Choi was received the M.S. degree with the School of Computer Engineering, Korea University of Technology and Education (KOREATECH). His research interests include lightweight deep learning, knowledge distillation, autonomous driving.

APPENDIX A

ADDITIONAL QUALITATIVE COMPARISONS

This section presents additional qualitative comparisons among teacher models, the baseline, response-based KD methods (Res-KD) employing various loss functions, and TIE-KD with distinct loss function configurations.

A.1 Teacher model: AdaBins [1]

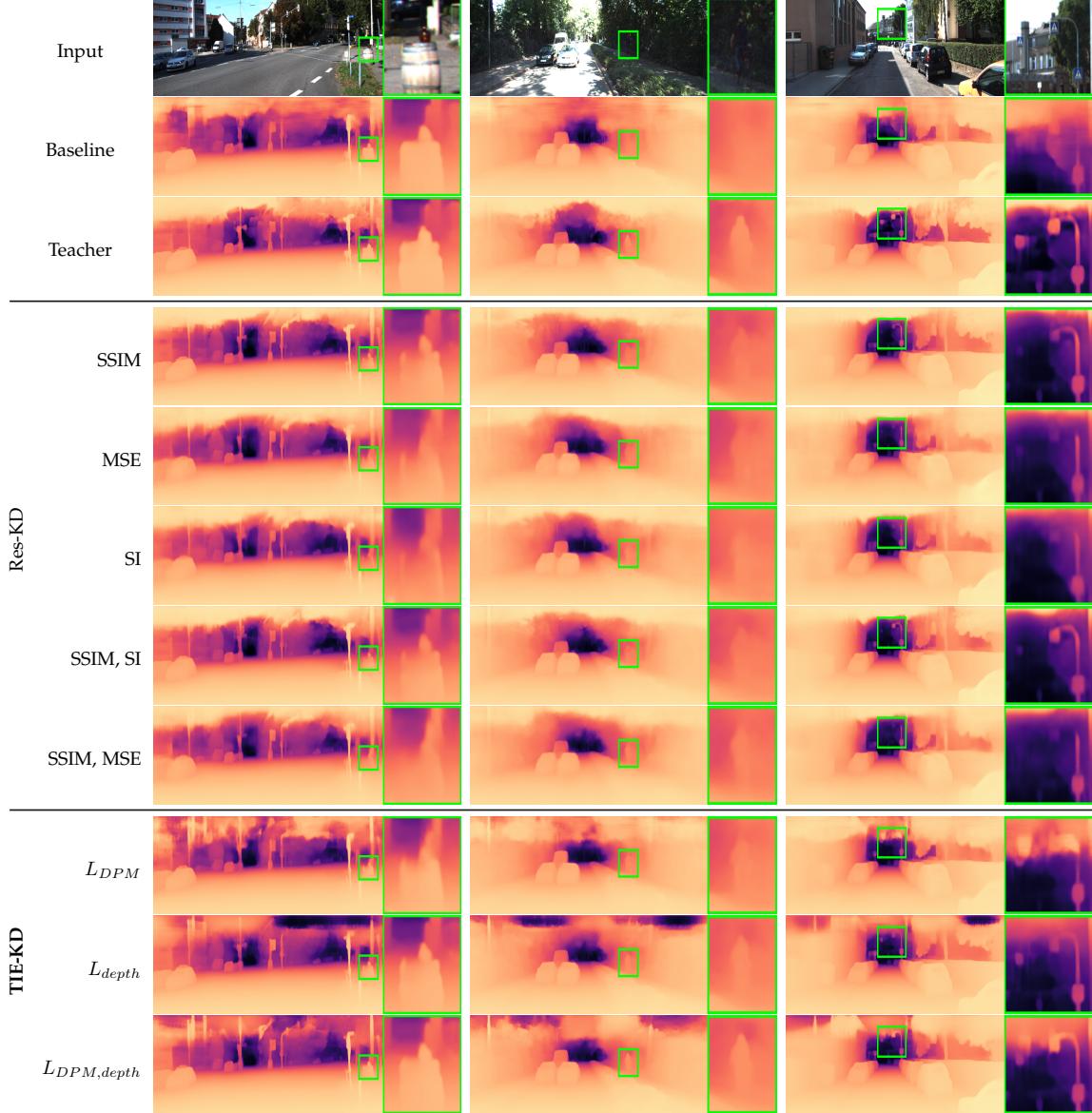


Fig. 6: Visual comparison of depth maps produced by different models for three distinct scenes, with a focus on detail variations within areas marked by green boxes. The top row presents the input images. The second row illustrates the depth maps generated by the teacher model, here adabins [1]. The third row depicts the baseline model’s output. Subsequent rows display the results of the student models trained using various response-based knowledge distillation methods (Res-KD) with different loss function combinations, and the bottom rows show the depth maps from students trained using our proposed TIE-KD framework with different loss function configurations.

A.2 Teacher model: BTS [27]

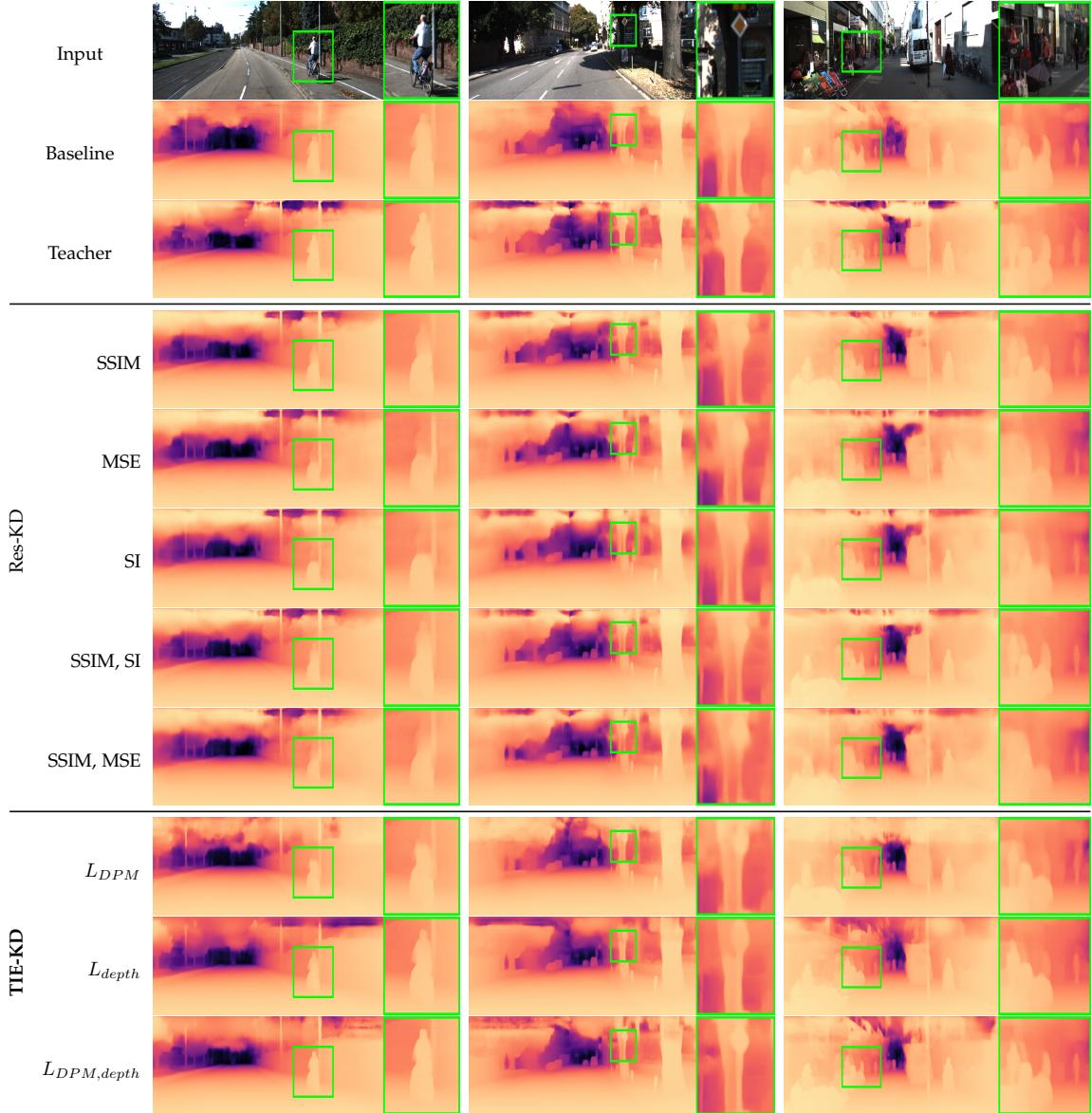


Fig. 7: Visual comparison of depth maps produced by different models for three distinct scenes, with a focus on detail variations within areas marked by green boxes. The top row presents the input images. The second row illustrates the depth maps generated by the teacher model, here BTS [27]. The third row depicts the baseline model's output. Subsequent rows display the results of the student models trained using various response-based knowledge distillation methods (Res-KD) with different loss function combinations, and the bottom rows show the depth maps from students trained using our proposed TIE-KD framework with different loss function configurations.

A.3 Teacher model: DepthFormer [30]

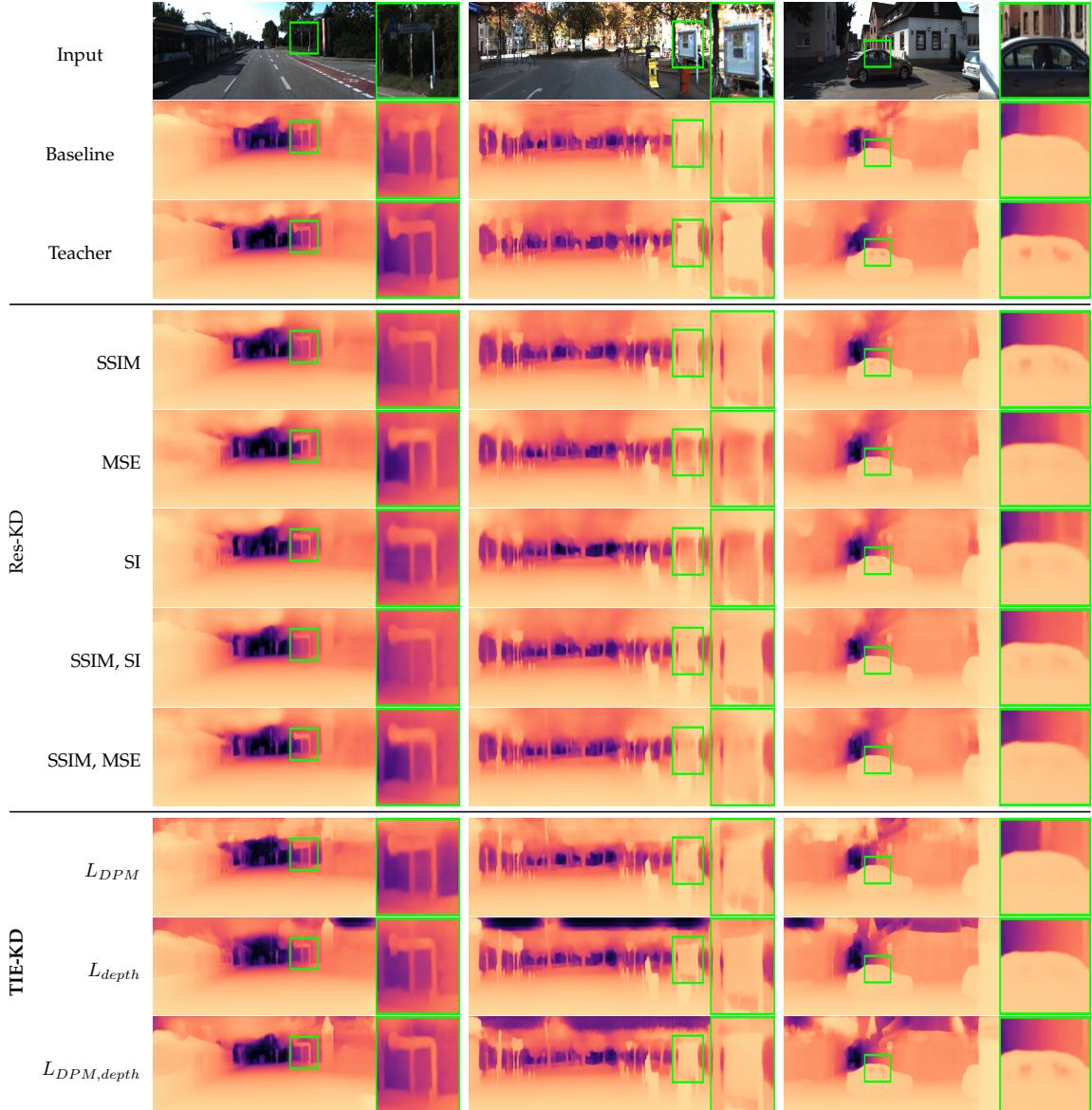


Fig. 8: Visual comparison of depth maps produced by different models for three distinct scenes, with a focus on detail variations within areas marked by green boxes. The top row presents the input images. The second row illustrates the depth maps generated by the teacher model, here DepthFormer [30]. The third row depicts the baseline model’s output. Subsequent rows display the results of the student models trained using various response-based knowledge distillation methods (Res-KD) with different loss function combinations, and the bottom rows show the depth maps from students trained using our proposed TIE-KD framework with different loss function configurations.