IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 25 April 2025, accepted 12 May 2025, date of publication 30 May 2025, date of current version 13 June 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3575066

RESEARCH ARTICLE

Global Structural Knowledge Distillation for Semantic Segmentation

HYEJIN PARK, KEONHEE AHN, HYESONG CHOI, (Student Member, IEEE), AND DONGBO MIN, (Senior Member, IEEE)

Department of Computer Science and Engineering, Ewha Womans University, Seoul 03760, Republic of Korea

Corresponding author: Dongbo Min (dbmin@ewha.ac.kr)

This work was supported in part by the Basic Research Laboratory Program through the National Research Foundation (NRF) of Korea under Grant RS-2023-00222385; and in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) through the Korean Government of the Ministry of Science and Information and Communication Technology (MSIT), Artificial Intelligence Innovation Hub under Grant RS-2021-II212068.

ABSTRACT Knowledge distillation (KD) has become a cornerstone for compressing deep neural networks, allowing a smaller student model to learn from a larger teacher model. In the context of semantic segmentation, traditional KD methods primarily focus on pixel-level feature alignment, where the student model is trained to match the teacher's features at each pixel. Despite performance improvements, the pixel-level alignment can introduce noise and redundant information, particularly in complex scenes, and often overlook the global structural context that is crucial for robust segmentation. To overcome these limitations, we propose Global Structural Knowledge Distillation (GSKD), a novel approach that moves beyond dense pixel-level alignment. Instead of aligning features pixel-by-pixel, we focus on capturing and transferring global structural information within an image. Our method begins with Class-Balanced Sampling (CBS), which ensures that representative features from various classes are sampled evenly from the teacher's feature maps. This helps the model better represent both common and rare classes, addressing class imbalance. Next, we construct a Global Structural Similarity Map (GSSM) for both the teacher and student models. This map encodes the key structural patterns of the image by calculating pairwise similarities between the sampled points, providing the structural information of the scene. To enhance the knowledge transfer process, we generate Sub-Image Descriptors (SID) through row-wise shuffling and column-wise grouping of the GSSM. These descriptors allow the student model to capture high-level semantic relationships and structural patterns, overcoming the limitations of traditional pixel-level feature alignment. The proposed method is designed to be flexible; It can be used both as a standalone method and as a plug-and-play module for integration with existing KD techniques. Our extensive experiments demonstrate that GSKD consistently outperforms or matches recent KD methods in standalone settings and significantly enhances the performance of state-of-the-art KD methods when incorporated as a plug-in-play module.

INDEX TERMS Knowledge distillation, model compression, semantic segmentation, image classification, global structural similarity, class-balanced sampling, sub-image descriptors.

I. INTRODUCTION

Semantic segmentation, a dense prediction task focused on segmenting objects at the pixel level, has made significant strides due to the advancements in deep neural networks (DNNs) [1], [2], [3], [4], [5]. However, deploying these complex models in resource-constrained environments remains a significant challenge due to their high computational and memory demands. To address this, various model compression techniques, such as network pruning [6], [7], network quantization [8], [9], and knowledge distillation (KD) [10], [11], [12], have been explored to reduce the model

The associate editor coordinating the review of this manuscript and approving it for publication was Tai Fei.

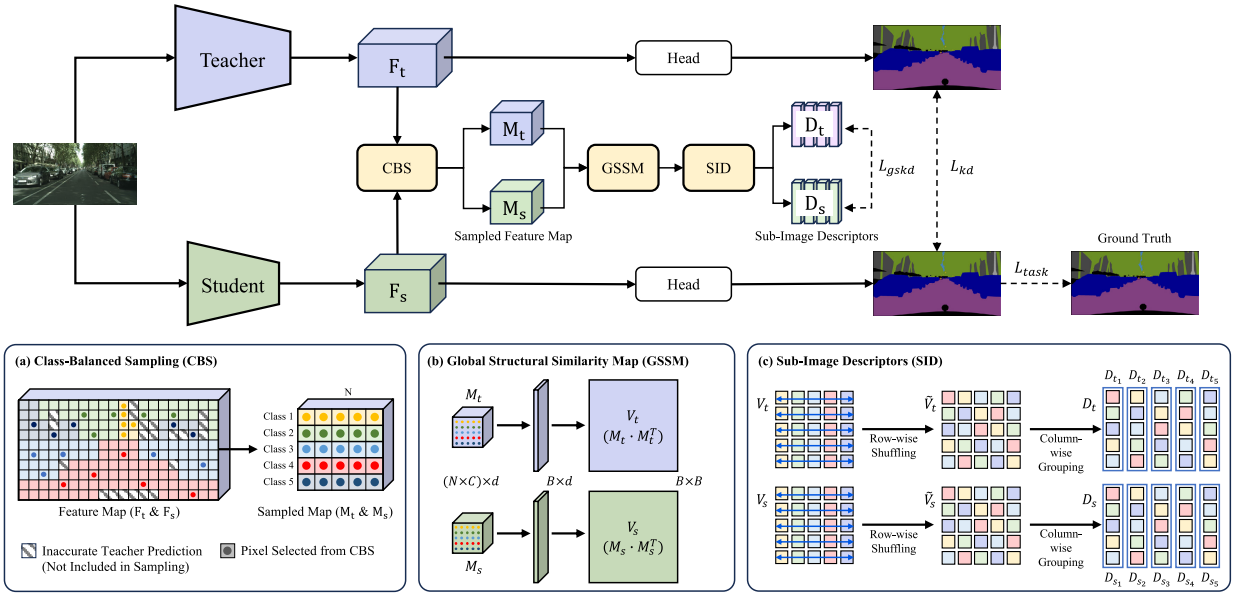

FIGURE 1. Overview of the proposed method. Global Structural Knowledge Distillation (GSKD) begins by sampling $\mathcal{B}$ (= NC) pixels from the teacher feature map in a class-balanced manner only on accurately predicted locations (highlighted in solid colors), as shown in (a). N and C represents the number of sampled points per class and the number of classes, respectively. Pixels at the same location are also sampled from the student feature map. The sampled feature maps M_t and M_S are then used to compute the global self-similarity maps V_t and V_S with dimensions $\mathcal{B} \times \mathcal{B}$, as depicted in (b). They are divided into sub-image descriptors through row-wise shuffling and column-wise grouping (c), which capture the global structural knowledge to be transferred from the teacher to the student. Finally, the knowledge is distilled using the loss functions L_{gskd} and L_{kd}.

size while preserving accuracy. Among these, KD stands out as a particularly effective method, wherein a large pre-trained teacher model transfers its knowledge to a smaller, more efficient student model.

In semantic segmentation, precise pixel-level predictions are crucial, and improving the performance of a student model requires aligning not only the prediction distributions but also the richer information within the feature maps of the teacher and student models. Structured Knowledge Distillation (SKD) [13] has proven to be an effective method for semantic segmentation, where it aligns pairwise similarities at the pixel level. By transferring these fine-grained pixel-level similarities from the teacher to the student, SKD captures structural information, which is essential for segmenting fine details in an image. However, SKD has its limitations. While it focuses on pixel-level alignments, it may propagate noise and redundant information, which hinders the student model's ability to understand the broader, global structure of the image. This arises because SKD attempts to propagate all pairwise pixel relationships, which, although useful for fine details, makes it difficult to capture the global structural patterns that are necessary for effective segmentation, especially in complex scenes with multiple objects and regions.

Accurate segmentation in complex scenes requires not only precise boundary detection but also a comprehensive understanding of the overall image context. Relying solely on pixel-level alignment risks neglecting essential relationships between different regions or classes, which are necessary to distinguish between multiple objects or classes. This highlights the need to move beyond fine-grained pixel similarity alignment and capture global structural relationships that provide a broader contextual understanding of the scene.

Despite recent improvements, such as CIRKD [14] and Af-DCD [15], which extend traditional KD by incorporating inter-image relationships and local structural knowledge. CIRKD introduces cross-image relationships to better capture long range dependencies across images, and Af-DCD improves the distillation process by using augmentation techniques. While these methods improve relational understanding within and across images, they still focus primarily on pixel-level or local structural alignments within individual images, which limits their ability to provide the necessary global structural context.

To address these limitations, we propose **Global Structural Knowledge Distillation (GSKD)**, a novel approach that moves beyond traditional fine-grained pixel-level alignment. Unlike conventional KD methods that align pixel-level similarities, GSKD focuses on capturing global structural relationships within an image. By transferring global structural knowledge instead of pixel-level details, GSKD enables the student model to learn more abstract, semantically rich representations, which are crucial for improving segmentation performance in complex spacial configurations.

To achieve this, GSKD employs **Class-Balanced Sampling (CBS)** to ensure a balanced representation of both common and rare class features. This sampling strategy reduce noise and addresses class imbalance, ultimately

improving the quality of the distilled knowledge. From the sampled feature maps, we construct a **Global Structural Similarity Map (GSSM)**, which encodes global structural patterns by calculating pairwise similarities between the sampled pixels. To extract more meaningful knowledge, we then generate **Sub-Image Descriptors (SID)** by row-wise shuffling and column-wise grouping the GSSM. This process allows the student model to focus on high-level semantic relationships, rather than fine-grained pixel alignments, and distill a more abstract, semantically rich representation of the image. Figure 1 illustrates the overall framework of GSKD, showing how CBS, GSSM, and SID are integrated.

Our method is flexible and can be both a standalone method and a plug-and-play module that can be easily integrated with existing KD methods. Extensive experiments show that GSKD significantly improves the performance of semantic segmentation tasks, especially when combined with state-of-the-art methods as a plug-and-play module, by leveraging a more effective transfer of global structural knowledge.

The main contributions of this paper are summarized as follows:

- We introduce **Class-Balanced Sampling (CBS)**, which reduces noise and redundancy by performing class-balanced sampling based on the teacher's accurate predictions. This ultimately improves the quality of the distilled knowledge.
- We generate **Global Structural Similarity Maps (GSSM)** to capture higher-level structural patterns by calculating pairwise similarities among class-balanced sampled points, providing a more holistic view of the image's structural context.
- We propose a novel technique for creating **Sub-Image Descriptors (SID)** from the GSSM, diversifying global structural patterns through row-wise shuffling and column-wise grouping for more effective knowledge distillation.
- GSKD offers **versatile integration**, fitting seamlessly into existing KD frameworks and consistently delivering significant performance improvements across various semantic segmentation benchmarks.

II. RELATED WORK

A. KNOWLEDGE DISTILLATION

Knowledge distillation [16], [17], [18], [19], [20], [21] is a a well-established technique where a large, pre-trained teacher model transfers its knowledge to a smaller student model, improving the student's performance. Knowledge distillation methods can generally be classified into three main categories: response-based, feature-based, and relation-based methods. Response-based methods [10] focus on matching the teacher and student model's output distributions (or logits) using Kullback-Leibler (KL) divergence, allowing the student to learn from the soft probabilities produced by the teacher. Feature-based methods [11], [12] transfer intermediate feature maps from the teacher to the student, aligning feature activations to guide the student's learning. Relation-based methods [19], [22] involve transferring knowledge about the relational information between features or layers, helping the student learn structural and spatial patterns from the teacher's representation. Recent works have also explored ensemble-based strategies such as multi-teacher distillation [23], [24], where knowledge is aggregated from multiple teachers to improve generalization. While these approaches leverage the diversity of multiple teacher models, they often require additional training complexity or architectural coordination.

B. KNOWLEDGE DISTILLATION FOR SEMANTIC SEGMENTATION

While KD is widely used in image classification [25], [26], [27], [28], its application to semantic segmentation presents unique challenges due to the dense, pixel-wise nature of the task. To address these challenges, specialized KD methods have been developed, focusing on aligning spatial visual patterns between teacher and student models. Reference [29] proposed distilling class probabilities for each pixel by leveraging local similarity maps incorporating boundary information from the teacher model. Building on this, SKD [13] introduces a method that captures global patterns by computing pairwise similarities [30] across the entire image. CWD [31] takes a different approach by normalizing channel activation maps and minimizing KL divergence between the teacher's and student's channel-wise probability distributions. Meanwhile, CIRKD [14] captures relationships between pixels across different training images using memory banks, facilitating pixel-to-pixel and pixel-to-region distillation. DIST [32] innovates by replacing KL divergence with a correlation-based loss that emphasizes inter- and intra-class relationships within batches.

More recent methods have explored additional strategies to enhance KD for semantic segmentation. MasKD [33] selects distillation regions using pixel-wise masks, while Af-DCD [15] employs contrastive learning [30], [34], [35], [36] techniques, focusing on spatial and channel contrasts to refine fine-grained feature representations. DiffKD [37] introduces denoising techniques to eliminate noise, ensuring the transfer of only valuable information from teacher to student. FreeKD [38] addresses the issue of information loss due to downsampling using semantic frequency prompts, which help preserve critical features during distillation.

Despite these advancements, most methods remain focused on fine-grained spatial relationships, often neglecting broader global structures crucial for segmentation. While approaches like FreeKD aim to retain high-level information, they still primarily emphasize localized features, potentially overlooking essential global patterns.

III. PRELIMINARY

We first provide a brief overview of semantic segmentation and knowledge distillation (KD) to establish the context

for our proposed method. Semantic segmentation assigns a label or class to each pixel of an image from C categories. The segmentation network takes an image I of dimensions $H \times W \times 3$ as input and extracts a feature map F of size $h \times w \times d$, where d is the number of channels. The classifier then transforms the feature map F into a categorical logit map X of size $h \times w \times C$. Finally, the segmentation map q is generated by selecting the class with the highest probability in X. The typical segmentation loss, $\mathcal{L}_{task}$, is given by:

$$\mathcal{L}_{task} = \frac{1}{hw} \sum_{p \in \mathcal{R}} \mathcal{CE}(q_p, y_p). \tag{1}$$

where $\mathcal{CE}$ is the cross-entropy loss, $\mathcal{R}$ represents all the pixels in the feature map F, and y_p is the ground truth label at a pixel p. To transfer knowledge from the teacher model to the student model, traditional KD methods [10] align the soft probability distributions between the two models using Kullback-Leibler (KL) divergence:

$$\mathcal{L}_{kd} = \frac{1}{hw} \sum_{p \in \mathcal{R}} \mathcal{KL}(X_t/\tau || X_s/\tau). \tag{2}$$

where τ is a temperature for KL divergence.

In semantic segmentation, a pairwise self-similarity map [13] is often used as structural knowledge to be transferred to the student model. This approach aligns the pairwise self-similarity of all pixels between the teacher and student models by minimizing the following loss:

$$\mathcal{L}_{pair} = \frac{1}{(hw)^2} \sum_{p \in \mathcal{R}} (A_{t,p}/\tau - A_{s,p}/\tau)^2. \tag{3}$$

$A_{t,p}$ (or $A_{s,p}$) represents the p^{th} index of the self-similarity map A_t (or A_s) computed from the teacher (or student) feature map:

$$A_l = \bar{F}_l \cdot \bar{F}_l^T \in \mathbb{R}^{hw \times hw}, \quad l = s, t. \tag{4}$$

where $\bar{F} \in R^{hw \times d}$ is reshaped from F, followed by a channel-wise normalization.

IV. PROPOSED METHOD

In this section, we introduce Global Structural Knowledge Distillation (GSKD), detailing how we define and transfer global structural similarity from the teacher model to the student model during the KD process. Pairwise self-similarity maps, as defined in Eq. (4), are commonly employed as a supervisory signal in knowledge distillation for semantic segmentation [13], [14], [39], [40]. However, these maps tend to focus heavily on fine-grained, pixel-level relationships, which are often noisy and fail to capture the broader, more abstract patterns necessary for understanding overall image structure. As noted in [41], this focus on pixel-level similarities can often lead to noise propagation, diminishing the accuracy and reliability of the distillation process.

To address these limitations, we propose **Global Structural Knowledge Distillation (GSKD)**, a novel approach that shifts the emphasis from pixel-level feature alignment to capturing and transferring global structural information within an input image (Figure 1). Our method starts with **Class-Balanced Sampling (CBS)** from the teacher model's feature maps, specifically focusing on accurately predicted locations, as shown in Figure 1 (a). This ensures that the sampled features represent the entire image more effectively, reducing the impact of noise and class imbalance.

We then construct a **Global Structural Similarity Map (GSSM)** by calculating pairwise similarities between the sampled points (Figure 1 (b)). Although the GSSM provides an abstract and holistic view of the image's structural context by capturing higher-level semantic relationships across different regions, relying solely on this map may be insufficient for effective knowledge transfer. While the abstraction of the GSSM is rich in global structural information, it may limit its practical utility for the student model.

To enhance the distillation process, we further decompose the GSSM into **Sub-Image Descriptors (SID)** using a novel technique that involves row-wise shuffling and column-wise grouping (Figure 1 (c)). This process diversifies the structural patterns captured, enabling the student model to learn broader, high-level semantic relationships. By doing so, SID overcomes the limitations of dense pixel-level alignment, such as overfitting to local details and difficulty in adapting to various scene contexts.

Note that our approach shares conceptual similarities with classical hand-crafted descriptors [42], [43], [44], [45], calculating self-similarities between pixel pairs within an image patch, capturing structural information at a more abstract level.

A. CLASS-BALANCED SAMPLING (CBS)

To accurately capture global structural patterns within an image, we begin by selecting $\mathcal{B} = N \times C$ samples from the teacher (or student) feature map F_t (or F_s), where N represents the number of samples per class and C is the total number of classes present in the image. This results in a set of sampled features M_t (or M_s) $\in \mathbb{R}^{\mathcal{B} \times d}$, where d denotes the number of channels in the feature map. M_t and M_s are L2-normalized along the channel dimension.

The sampling process is guided by a class-balanced strategy to ensure that pixels from all classes are represented evenly. To further reduce noise and enhance reliability of sampled features, we focus on regions where the teacher model has made accurate predictions. This combined approach ensures that the sampled features are both diverse and representative across all classes, as illustrated in Figure 1 (a).

For clarity, we denote the concatenation of these sampled features as $M = [m_1, m_2, \ldots, m_{\mathcal{B}}]^T$, where each m_i is a d-dimensional feature vector corresponding to a sampled pixel. It is critical that the sampling positions are identical for both the teacher and student models to allow accurate comparison of their structural similarities. This consistency in spatial locations between M_t and M_s ensures meaningful alignment and comparison of their respective structural patterns.

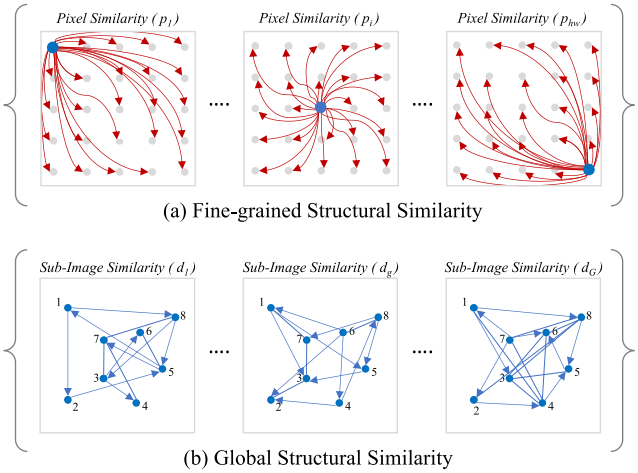

FIGURE 2. Distinction between Fine-grained and Global Structural Similarity: (a) The fine-grained structural similarity approach [13], [14], visualized through a dense network of arrows, represents pixel-level pairwise comparisons across the entire feature map. Each arrow denotes the similarity measurement from a reference pixel (blue dot) to all other pixels (gray dots), across the spatial resolution of the feature map (indicated by *hw*). (b) In contrast, the proposed global structural similarity approach is characterized by sparser connections, showcasing sub-image descriptors that capture relationships among a selected set of sample pixels. The number of sub-image descriptors (denoted by *G*) is determined by the strategic sampling of points, rather than an exhaustive pairwise comparison, leading to an abstract representation of structural relationships.

Algorithm 1 Global Structural Knowledge Distillation

N: sampling size per class
C: the number of classes in an image
S: the stride of grouping columns
$\mathcal{B} = NC$, $G = \frac{\mathcal{B}}{S}$ should be an integer.
$\bar{f}_t = L2Norm(f_t), \bar{f}_s = L2Norm(f_s)$
Class-balanced Sampling
Initialize sampled feature $M_t = \varnothing, M_s = \varnothing$
for $i = 1, 2, \ldots C$ **do**
  $m_{t_i} \in \mathbb{R}^{N\times 1}$: N samples with a class i from $\bar{f}_t$
  $m_{s_i} \in \mathbb{R}^{N\times 1}$: N samples with a class i from $\bar{f}_s$
  $M_t \leftarrow M_t \bigcup m_t$
  $M_s \leftarrow M_s \bigcup m_s$
end for
Global self-similarity map $V_t, V_s \in \mathbb{R}^{\mathcal{B}\times\mathcal{B}}$
$V_t, V_s = M_t M_t^T, M_s M_s^T$
$\tilde{V}_t, \tilde{V}_s$ = Row-wise shuffling(V_t, V_s)
Sub-image Descriptor Loss
$L_{gskd} = 0$
for $i = 0, 1, \ldots, G-1$ **do**
  $D_{t_i} = [\tilde{V}_{t:i\times S}, \ldots, \tilde{V}_{t:(i+1)\times S-1}]$
  $D_{s_i} = [\tilde{V}_{t:i\times S}, \ldots, \tilde{V}_{t:(i+1)\times S-1}]$
  $\mathcal{L}_{gskd} = \mathcal{L}_{gskd} + \mathcal{KL}(\sigma(D_{s_i}/\tau), \sigma(D_{t_i}/\tau))$
end for
$\mathcal{L}_{gskd} = \mathcal{L}_{gskd}/G$

B. GLOBAL STRUCTURAL SIMILARITY MAP (GSSM)

To capture global relationships across different regions of the image, we compute the GSSM, represented in $\mathbb{R}^{\mathcal{B}\times\mathcal{B}}$, as illustrated in Figure 1 (b). This GSSM captures pairwise similarities between the sampled features in M as follows:

$$V_l = M_l M_l^T \in \mathbb{R}^{\mathcal{B}\times\mathcal{B}}, \quad l = s, t. \tag{5}$$

where V_t and V_s represent the global structural similarity maps for the teacher and student models, respectively.

These GSSMs, V_t and V_s, are conceptually similar to the self-similarity maps A_t and A_s of Eq. (4), as both are computed using the self-similarity metric. However, our GSSM is constructed using only the subset of features generated through CBS, making it more focused and efficient. While GSSM effectively encapsulates global structural patterns, we further enhance knowledge transfer by introducing a sub-image descriptor based on GSSM, which will be detailed in the next section.

C. SUB-IMAGE DESCRIPTORS FOR KD

As shown by the comparison between the results in Table 11 (d) and (f), directly using the GSSM for knowledge transfer can be too abstract, limiting its effectiveness, especially in dense prediction tasks. To address this, we propose dividing the GSSM into multiple sub-image descriptors (SID).

As illustrated in Figure 1 (c), we first apply row-wise random shuffling to V_t and V_s to diversify the structural patterns within each descriptor, preventing columns from being tied to a single reference point. Without shuffling, each column would represent similarities centered around a specific sampled location as illustrated in Figure 2 (a), limiting the diversity of captured patterns. After shuffling, the GSSMs $\tilde{V}_t$ and $\tilde{V}_s$ are then transformed into a vector format within $\mathbb{R}^{\mathcal{B}^2}$. These vectors are then grouped to form sub-image descriptors, enhancing the representational capacity of the distilled knowledge by capturing diverse structural patterns in a global context (see Figure 2 (b)).

The SIDs D_t and D_s are formally defined as:

$$\begin{aligned} D_{t_i} &= [\tilde{V}_{t:i\times S}, \ldots, \tilde{V}_{t:(i+1)\times S-1}], \quad i \in [0, G-1] \\ D_{s_i} &= [\tilde{V}_{s:i\times S}, \ldots, \tilde{V}_{s:(i+1)\times S-1}], \quad i \in [0, G-1]. \end{aligned} \tag{6}$$

where $\tilde{V}$: i indicates the i^{th} column of matrix $\tilde{V}$, and $\tilde{V}t_{:i\times S}$ and $\tilde{V}s$: $i \times S$ represent the i^{th} sub-image descriptors (SIDs) generated from the teacher and student feature maps, respectively. Here, S denotes the grouping size of columns in the GSSM $\tilde{V}$, and is chosen such that $G = \mathcal{B}/S$ is an integer. G represents the total number of SIDs.

The SIDs D_{t_i} and D_{s_i} capture a more refined and structured representation compared to using the entire GSSM as a single descriptor. By dividing the GSSM into smaller, localized sub-image descriptors, we focus on a more diverse set of global structural patterns, which provides more nuanced supervision for the student model. This approach allows us to better capture and transfer global structural information, addressing limitations in pixel-level alignment. The effectiveness of this approach is validated in our ablation studies, which

demonstrate significant performance improvements with the use of SIDs (see Table 11 (d), (e), and (f)).

D. LOSS FUNCTION FOR GSKD

This global structural similarity loss $\mathcal{L}_{gskd}$ is computed as follows:

$$\mathcal{L}_{gskd} = \frac{1}{G}\sum_{i=0}^{G-1} \mathcal{KL}(\sigma(D_{s_i}/\tau), \sigma(D_{t_i}/\tau)). \quad (7)$$

where D_{t_i} and D_{s_i} are the sub-image descriptors for the teacher and student models, respectively. The $\mathcal{KL}$ loss measures the discrepancy between these descriptors. We use KL-divergence to capture differences in the structural patterns of the descriptors, with σ as the softmax function and τ as the temperature parameter, which is fixed to 1 in our experiments.

1) TOTAL LOSS

The total loss function for GSKD, when used as a standalone framework, is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \mathcal{L}_{kd} + \lambda \cdot \mathcal{L}_{gskd}. \quad (8)$$

where the hyper-parameter λ controls the weighting of L_{gskd}, balancing the contribution of structural similarity during training. Our method is depicted in Algorithm 1.

2) PLUG-AND-PLAY INTEGRATION

When our global structural similarity loss is integrated into existing KD frameworks, the total loss can be expressed as:

$$\mathcal{L}_{plugin} = \mathcal{L}_{other} + \lambda \cdot \mathcal{L}_{gskd}. \quad (9)$$

where L_{other} represents the loss used in the existing methods. This formulation allows our method to be seamlessly combined with other KD techniques, enhancing their ability to transfer more abstract and global structural knowledge.

V. EXPERIMENTS

A. EXPERIMENTAL SETUPS

1) DATASETS

We conducted the evaluation on the following benchmarks:

- The Cityscapes [46] is a semantic urban scene understanding dataset that contains 2975/500/1525 images for training, validation, and testing with 19 semantic classes for segmentation performance.
- The CamVid [47] is a road/driving scene understanding database that contains 367/101/233 images for training, validation, and testing with 11 semantic classes.
- The PascalVOC [48] is a visual object recognition dataset that contains 1464/1499/1456 images for training, validation, and testing with 20 foreground object classes and one background class.
- The ADE20K [49] is a scene understanding dataset featuring 150 diverse scene classes, divided into 20,000 training, 2,000 validation, and 3,000 testing images. This dataset provides a wide array of scene types and objects, making it a valuable resource for tasks like semantic segmentation and scene parsing. Each image is annotated with detailed object labels, aiding in advanced computer vision research.

2) NETWORK ARCHITECTURE

We used DeepLabV3 [50] with a ResNet101 [51] backbone as the teacher network. We adopt various segmentation architectures and backbones for the student networks to verify the effectiveness of distillation methods. We use the different segmentation framework DeepLabV3 with ResNet18 and MobileNetV2 [52] backbones, as well as PSPNet [5] with a ResNet18 backbone. We additionally evaluated our method using Segformer [53], employing MiT-B4 as the teacher network and MiT-B0 as the student network.

3) TRAINING DETAILS

For training our GSKD model as a standalone method, we utilized the sampling size N of 16, the grouping stride S of 8, and set the loss weight λ for $\mathcal{L}_{gskd}$ to 10. For plug-in with other KD methods, the sampling size N was adjusted to 8, and the grouping stride S was set to the number of classes (C) present in the image. λ for $\mathcal{L}_{gskd}$ in plug-ins was set to 2 for CWD [31], 5 for SKD [13], CIRKD [14], and Af-DCD [15], and 10 for DIST [32]. All models were trained using SGD with a momentum of 0.9 and an initial learning rate of 0.02. The crop sizes for the training datasets were set to 512×1024 for Cityscapes, 360×360 for CamVid, and 512×512 for both Pascal VOC and ADE20K. Models were trained for 40K iterations with a batch size of 16. All other training configurations adhered to the settings established in the CIRKD codebase. Experiments were run on a server equipped with NVIDIA RTX 3090 and H100 GPUs, and the code was implemented using PyTorch. To assess computational efficiency, we also measured the runtime overhead of loss computation. Despite the descriptor-based operations, GSKD loss takes on average ∼0.011s per image, which is still faster than SKD with adversarial training (∼0.019s).

B. SEMANTIC SEGMENTATION

We evaluated GSKD both as a standalone method and in combination with various KD methods across multiple models and datasets, using mIoU metric. Experiments marked with * were re-run under consistent conditions due to reproducibility issues; all other results are from the original papers.

1) PERFORMANCE ON VARIOUS DATASETS

We evaluate our proposed Global Structural Knowledge Distillation (GSKD) across four semantic segmentation benchmarks—Cityscapes, CamVid, PascalVOC, and ADE20K—using various student-teacher architectures. Overall, GSKD consistently outperforms baseline students and shows strong compatibility with existing knowledge distillation (KD) methods. In complex urban scenes like Cityscapes and CamVid, GSKD delivers particularly large

TABLE 1. Quantitative result comparison with state-of-the-art KD methods on Cityscape. * denotes that we reproduce the method using the code that the authors released to the public. T and S are the abbreviations for Teacher and Student.

Model	Params(M)	FLOPs(G)	Val mIoU(%)
T: DeepLabV3-ResNet101	61.1M	2371.7G	78.07
S: DeepLabV3-ResNet18	13.6M	572.0G	74.21
+ GSKD			76.34 (+2.22)
+ SKD [13]			75.42
+ SKD + GSKD			75.99 (+0.57)
+ CWD [31]			75.55
+ CWD + GSKD			76.93 (+1.38)
+ CIRKD [14]			76.38
+ CIRKD + GSKD			76.80 (+0.42)
+ DIST [32]			77.10
+ DIST + GSKD			78.17 (+1.07)
+ Af-DCD* [15]			76.05
+ Af-DCD + GSKD			76.54 (+0.49)
S: DeepLabV3-MobileNetV2	3.2M	128.9G	65.17
+ GSKD			75.07 (+9.9)
+ SKD [13]			73.82
+ SKD + GSKD			74.98 (+1.16)
+ CWD [31]			74.66
+ CWD + GSKD			75.14 (+0.48)
+ CIRKD [14]			75.42
+ CIRKD + GSKD			75.66 (+0.24)
+ DIST [32]			74.67
+ DIST + GSKD			75.87 (+1.20)
+ Af-DCD* [15]			75.07
+ Af-DCD + GSKD			75.64 (+0.57)
S: PSPNet-ResNet18	12.9M	507.4G	72.55
+ GSKD			73.66 (+1.11)
+ SKD [13]			73.29
+ SKD + GSKD			73.87 (+0.58)
+ CWD [31]			74.36
+ CWD + GSKD			75.86 (+1.50)
+ CIRKD [14]			74.73
+ CIRKD + GSKD			75.52 (+0.79)
+ DIST [32]			75.79
+ DIST + GSKD			76.21 (+0.42)
+ Af-DCD* [15]			73.53
+ Af-DCD + GSKD			74.03 (+0.50)

gains, occasionally exceeding teacher-level performance, while PascalVOC sees moderate yet consistent improvements. Furthermore, our analysis of per-class IoU highlights that GSKD excels on rare or structurally complex classes, and qualitative results confirm that GSKD sharply refines object boundaries. We also demonstrate the extensibility of GSKD on ADE20K and a ViT-based segmentation framework, underscoring the method's robustness and broad applicability.

a: CITYSCAPES

In Table 1, the teacher model (DeepLabV3-ResNet101) achieves 78.07% mIoU. By contrast, the baseline student (DeepLabV3-ResNet18) attains 74.21%, which GSKD alone raises to 76.34% (+2.13%). We also see steady gains when combining GSKD with other distillation methods (SKD, CWD, CIRKD, Af-DCD), each improving over its respective baseline. Notably, DIST+GSKD achieves 78.17%, even surpassing the teacher by +0.10%. A similar trend is observed for DeepLabV3-MobileNetV2, where GSKD provides a large gain (+9.90%) over the baseline and consistently improves performance when integrated with other KD methods.

TABLE 2. Comparison of qualitative results with state-of-the-art KD methods on CamVid. * denotes that we reproduce the method using the code that the authors released to the public. T and S are the abbreviations for Teacher and Student.

Model	Params(M)	FLOPs(G)	Test mIoU(%)
T: DeepLabV3-ResNet101	61.1M	280.2G	69.84
S: DeepLabV3-ResNet18	13.6M	61.0G	66.92
+ GSKD			68.15 (+1.23)
+ SKD [13]			67.46
+ SKD + GSKD			67.74 (+0.28)
+ CWD [31]			67.71
+ CWD + GSKD			68.59 (+0.98)
+ CIRKD [14]			68.21
+ CIRKD + GSKD			68.91 (+0.70)
+ DIST [32]			69.23
+ DIST + GSKD			69.86 (+0.63)
+ Af-DCD* [15]			67.49
+ Af-DCD + GSKD			68.17 (+0.68)
S: PSPNet-ResNet18	12.9M	45.6G	66.73
+ GSKD			67.18 (+0.45)
+ SKD* [13]			67.02
+ SKD + GSKD			67.73 (+0.71)
+ CWD [31]			67.92
+ CWD + GSKD			68.98 (+1.06)
+ CIRKD [14]			68.65
+ CIRKD + GSKD			69.48 (+0.83)
+ DIST [32]			67.70
+ DIST + GSKD			69.96 (+2.26)
+ Af-DCD* [15]			66.62
+ Af-DCD + GSKD			67.38 (+0.76)

TABLE 3. Comparison of quantitative results with state-of-the-art KD methods on PascalVOC. * denotes that we reproduce the method using the code that the authors released to the public. T and S are the abbreviations for Teacher and Student.

Model	Params(M)	FLOPs(G)	Val mIoU(%)
T: DeepLabV3-ResNet101	61.1M	1294.6G	77.67
S: DeepLabV3-ResNet18	13.6M	503.0G	73.21
+ GSKD			73.57 (+0.36)
+ SKD [13]			73.51
+ SKD + GSKD			73.82 (+0.31)
+ CWD [31]			74.02
+ CWD + GSKD			74.85 (+0.83)
+ CIRKD* [14]			73.53
+ CIRKD + GSKD			73.92 (+0.39)
+ DIST [32]			73.74
+ DIST + GSKD			74.05 (+0.29)
+ Af-DCD* [15]			74.84
+ Af-DCD + GSKD			74.95(+0.11)
S: PSPNet-ResNet18	12.9M	260.0G	73.33
+ GSKD			74.34 (+1.01)
+ SKD* [13]			73.30
+ SKD + GSKD			73.69 (+0.39)
+ CWD [31]			73.99
+ CWD + GSKD			74.09 (+0.10)
+ CIRKD* [14]			73.97
+ CIRKD + GSKD			74.53 (+0.56)
+ DIST [32]			74.12
+ DIST + GSKD			74.28 (+0.16)
+ Af-DCD* [15]			75.88
+ Af-DCD + GSKD			76.41(+0.53)

b: CAMVID

Table 2 shows that the teacher (DeepLabV3-ResNet101) records 69.84% mIoU, while the student (DeepLabV3-ResNet18) starts at 66.92%. Adopting GSKD alone lifts it to 68.15%, and DIST+GSKD achieves 69.86%, slightly

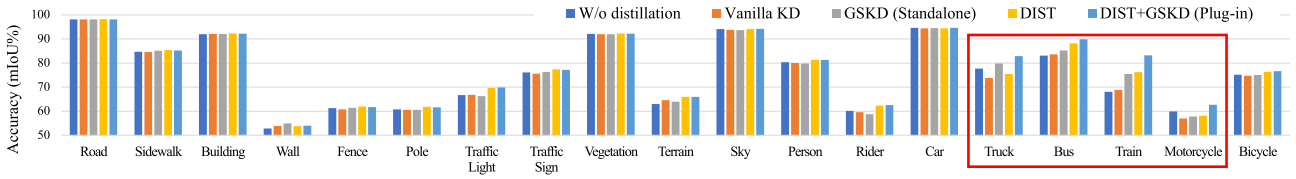

FIGURE 3. Illustration of individual class IoU scores on Cityscapes over original student netowrk (w/o distillation), Vanilla KD, GSKD (standalone), DIST [32], and DIST+GSKD (plug-in). DeepLabV3 with ResNet18 is used as the student network. The proposed GSKD improves the segmentation performance. More significant improvement is shown for Truck, Bus, Train, and Motorcycle.

TABLE 4. Class-level IoU on Cityscapes, grouped by pixel frequency (common vs. rare) as illustrated in Figure 3. We compare GSKD, DIST, and DIST+GSKD to show how GSKD mitigates class imbalance, especially in rare classes such as Truck and Train. This table complements the bar chart in Figure 3 by providing detailed per-class metrics, highlighting notable gains for structurally complex objects.

	Common Classes							Rare Classes											
Class	Road	Bldg	Veg	Car	Sidewalk	Sky	Person	Terrain	Bicycle	T.Light	Wall	Fence	Pole	Rider	Bus	T.Sign	Moto	Train	Truck
Pixel Freq. (%)	15.0	12.5	11.0	7.0	5.0	5.0	1.2	0.9	0.8	0.5	0.5	0.4	0.3	0.3	0.3	0.3	0.2	0.2	0.2
W/o distill	98.11	91.94	92.05	94.62	84.71	94.05	80.33	62.96	75.09	66.63	52.79	61.30	60.70	60.11	83.10	76.08	59.94	68.09	77.67
GSKD	98.14	92.03	91.99	94.50	85.09	93.63	79.79	63.88	75.08	66.22	54.92	61.38	60.51	58.77	85.21	76.31	57.76	75.47	79.80
DIST	**98.22**	**92.25**	**92.23**	94.44	**85.39**	94.07	**81.31**	**65.96**	76.34	69.64	53.71	**61.90**	**61.76**	62.35	88.17	**77.33**	58.12	76.21	75.42
DIST+GSKD	98.07	92.21	92.15	**94.57**	85.20	**94.15**	81.23	65.92	**76.57**	**69.84**	**54.02**	61.70	61.64	**62.58**	**89.88**	77.11	**62.63**	**83.15**	**82.85**

exceeding the teacher's performance. PSPNet-ResNet18 follows a similar pattern, jumping from 66.73% to 69.96% with DIST+GSKD (+3.23%), again surpassing the teacher. These results highlight how GSKD's global structural guidance can enhance lightweight students, even in challenging driving scenes.

c: PASCALVOC

In PascalVOC (Table 3), where images often contain only 1-2 classes, the gains from GSKD are relatively modest. For example, as a plug-in with Af-DCD, GSKD improved performance by 0.11% for DeepLabV3-ResNet18 and 0.53% for PSPNet-ResNet18. This less boost is likely due to the limited class diversity in PascalVOC, which reduces the effectiveness of CBS in generating diverse SIDs, leading to less significant improvements compared to datasets with more varied class distribution.

d: STANDALONE EFFECTIVENESS OF GSKD

When used as a standalone method, GSKD achieves an improvement of up to 1.25% mIoU over SKD in various models and datasets, effectively boosting student performance by focusing on global structural knowledge rather than exhaustive pixel-level comparisons. Both SKD and GSKD leverage pairwise self-similarity maps; however, GSKD distinguishes itself by utilizing a targeted set of sampled features to generate GSSM and SIDs. This approach not only reduces computational overhead but also improves feature alignment, demonstrating that distilling meaningful structural information can be more effective than relying on full pixel-wise similarity calculations.

e: SYNERGY WITH DIST

Notably, the significant performance gains observed when integrating GSKD with DIST can be attributed to their complementary characteristics. DIST is specifically designed to distill distribution-level knowledge by leveraging correlation-based metrics in the logit space, thereby modeling both inter-class and intra-class relationships. In contrast, GSKD focuses on capturing global structural patterns within the intermediate feature representations. By jointly leveraging distribution-aware guidance from DIST and structure-aware supervision from GSKD, the student model learns richer relational patterns across both the output and feature spaces, ultimately resulting in more discriminative and generalizable representations. This synergy proves particularly advantageous in lightweight student models and structurally complex scenes, occasionally enabling the student to even outperform the teacher. For instance, in Tables 1 and 2, the combination of DIST+GSKD surpasses the teacher's performance by up to +0.10% on Cityscapes and +0.12% on CamVid. We attribute these gains to the complementary nature of DIST and GSKD: while DIST sharpens prediction-level distributions, GSKD provides global structural guidance that the teacher model may not fully exploit. By merging precise prediction cues from DIST with GSKD's holistic feature alignment, the student inherits the teacher's strengths and gains additional context—thus forming a more robust representation than the teacher in certain cases.

f: COMPLEMENTARY WITH OTHER KD METHODS

Moreover, GSKD exhibits a complementary effect when combined with other methods that focus on local structural knowledge, such as SKD, CWD, CIRKD, and Af-DCD. By enriching these approaches with global structural information, GSKD significantly improves their ability to transfer meaningful knowledge from teacher to student models. This synergy not only reinforces fine-grained structural details but also enables the student model to develop a more holistic understanding of the scene, leading to more robust segmentation performance across diverse datasets.

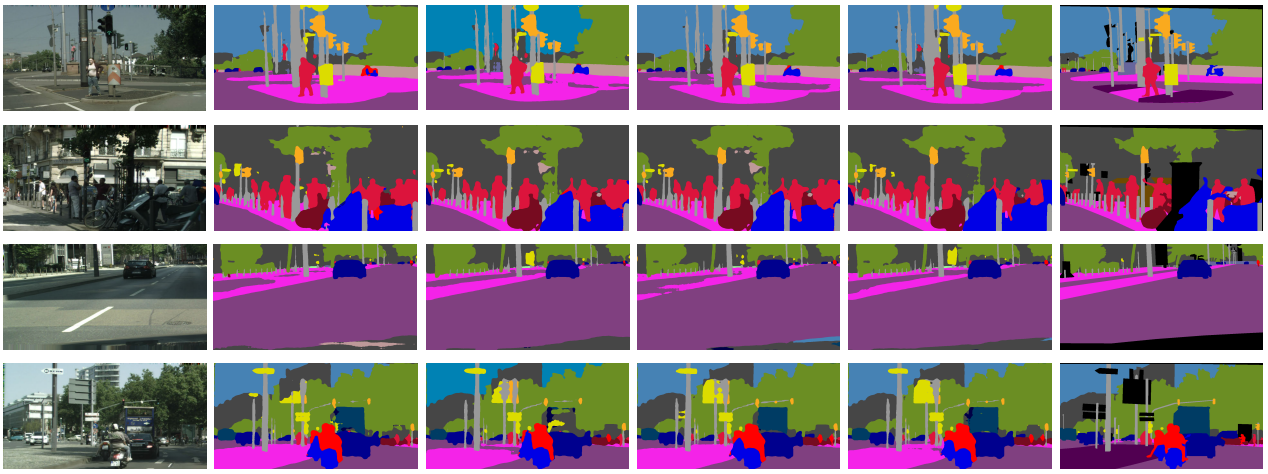

FIGURE 4. Qualitative segmentation results using DeepLabV3 with ResNet18 network on the validation set of Cityscapes. Images in each column are: (a) input image, (b) original student network (w/o distillation), (c) GSKD (standalone), (d) DIST [32], (e) DISG+GSKD (plug-in) and (f) Ground Truth (GT).

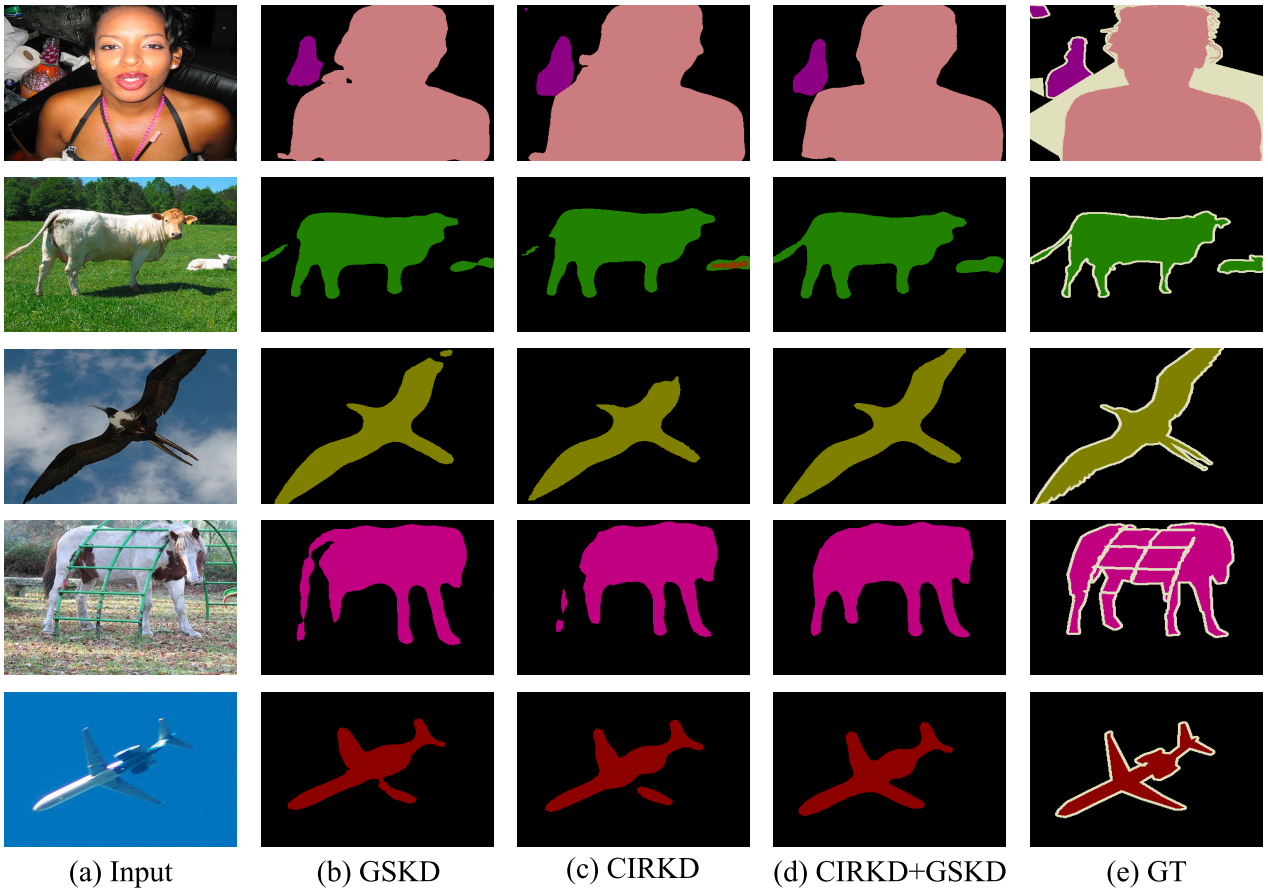

FIGURE 5. Qualitative segmentation results using DeepLabV3 with ResNet18 network on the validation set of PascalVOC. Images in each column are: (a) input image, (b) GSKD (standalone), (c) CIRKD, (d) CIRKD+GSKD (plug-in) and (e) Ground Truth (GT).

2) PER-CLASS IoU ANALYSIS

Figure 3 illustrates how our method improves segmentation accuracy across various Cityscapes classes, with notable gains in challenging categories such as *Wall*, *Rider*, and *Motorcycle*. To further investigate this, we categorize each class as either common or rare based on pixel frequency (Table 4). GSKD shows particularly large improvements for rare and structurally complex objects, including *Truck*, *Bus*,

Train, and *Bicycle*—classes that are typically underrepresented and more difficult to segment accurately.

When used as a standalone method, GSKD not only improves overall mIoU but also delivers substantial class-level gains, especially for *Truck* and *Train*. Furthermore, combining GSKD with DIST (DIST+GSKD) results in even stronger improvements in these rare categories, underscoring the complementary nature of distribution-level and global structural guidance. Collectively, these findings demonstrate that GSKD's focus on holistic, global features helps alleviate class imbalance and improve segmentation performance for underrepresented categories—ultimately leading to more robust overall results. This per-class analysis confirms that GSKD significantly enhances performance on rare or infrequent classes, contributing to the overall mIoU gain observed across datasets.

3) QUALITATIVE ANALYSIS

Figure 4 presents segmentation results on the Cityscapes test set, highlighting the advantages of GSKD and its plug-in variant DIST+GSKD. GSKD produces more accurate and coherent segmentation maps for challenging classes such as *Wall* and *Tree*. For instance, in the second row, GSKD yields sharper and more consistent predictions for *Pedestrian* and *Pole* compared to DIST [32]. In the first row, the *Motorcycle* class is segmented with cleaner boundaries and fewer artifacts, reflecting GSKD's strength in preserving spatial structure–particularly for medium-to-large objects where long-range consistency is beneficial.

When combined with DIST, the DIST+GSKD configuration further enhances clarity and object definition, particularly in complex scenes with densely packed elements or intricate boundaries. For example, the segmentation of the *Traffic Sign* in the bottom row is significantly cleaner, with fewer cluttered pixels and no holes inside the segmented area. This demonstrates how global structure from GSKD complements pixel-level guidance from DIST, leading to more robust and accurate segmentation results.

In Figure 5, we present segmentation results on PascalVOC, where GSKD (standalone) is compared against CIRKD and their plug-in combination (CIRKD+GSKD). Notably, GSKD alone already refines object boundaries for challenging shapes, such as the *Cow* and *Airplane*, by capturing more coherent silhouettes and reducing internal noise. When further integrated into CIRKD, GSKD yields even sharper segmentation masks with fewer misclassified pixels—highlighting the synergy between local (CIRKD) and global (GSKD) structural guidance. For instance, the *Cow*'s legs in the second row and the *Bird*'s wings in the third row are segmented more precisely with CIRKD+GSKD than with CIRKD alone. These qualitative improvements illustrate that adding global structural knowledge to local relational cues can provide a more holistic representation of the scene, ultimately resulting in cleaner and more accurate predictions.

4) VISUALIZATION AND INTERPRETABILITY

We further analyze the effectiveness of GSKD through qualitative visualizations. Figure 6 presents Grad-CAM heatmaps for selected classes (*Sky, Tree, Pole*) on the CamVid test set. The student model is PSPNet with a ResNet18 backbone, and the teacher is DeepLabV3 with ResNet101. Compared to SKD, GSKD highlights semantically meaningful regions with better spatial alignment to object boundaries. The plug-in version (SKD+GSKD) additionally captures both fine-grained local and holistic global features, resulting in more focused and consistent activation maps.

To investigate the representation structure learned by different methods, we visualize the final-layer features using t-SNE, as shown in Figure 7. We use the same student-teacher configuration and test set (CamVid). GSKD yields more compact intra-class clusters and greater inter-class separation than SKD, indicating improved feature discriminability. For instance, the *Bicyclist* class forms a clearly separated cluster under GSKD and SKD+GSKD, closely resembling the teacher's structure, while it appears more entangled with *Car* and *Pedestrian* under SKD. Similarly, *Tree* and *Road* clusters remain better disentangled in GSKD, preserving the teacher's inter-class boundaries. These examples highlight how GSKD enables the student to learn more semantically coherent representations. Furthermore, the SKD+GSKD plug-in variant demonstrates further alignment with the teacher's structure, suggesting the complementary benefits of combining global and local guidance in the distillation process.

5) PERFORMANCE ON LARGE-SCALE DATASET

Table 5 demonstrates the effectiveness of GSKD on the complex ADE20K dataset. GSKD achieves a 0.84% improvement over the baseline. When combined with SKD and Af-DCD, GSKD consistently improves the accuracy of student model, highlighting its robustness and ability to boost KD performance in challenging, large-scale environments. Figure 8 further illustrates the benefits of GSKD and its plug-in usage with Af-DCD. For instance, in the first row, GSKD improves the boundary definition of *Lamp* and *Pillow* regions, while in the second row, the segmentation of *people* is more consistent and less fragmented compared to Af-DCD. In the third row, the standalone GSKD recovers fine-grained object details, such as the *Lamp* and the *Side Table* next to the bed. The plug-in variant (Af-DCD+GSKD) further enhances both spatial consistency and class-level distinction—especially in crowded scenes like the fourth row, where *Roadlines*, *Trees*, and *Pedestrians* are better preserved. These results show that GSKD not only complements relational distillation methods like Af-DCD but also offers structural guidance that improves prediction coherence across complex scenes.

6) PERFORMANCE ON DIFFERENT BACKBONE

To evaluate the generality of GSKD, we conduct experiments using Segformer [53], a transformer-based segmentation

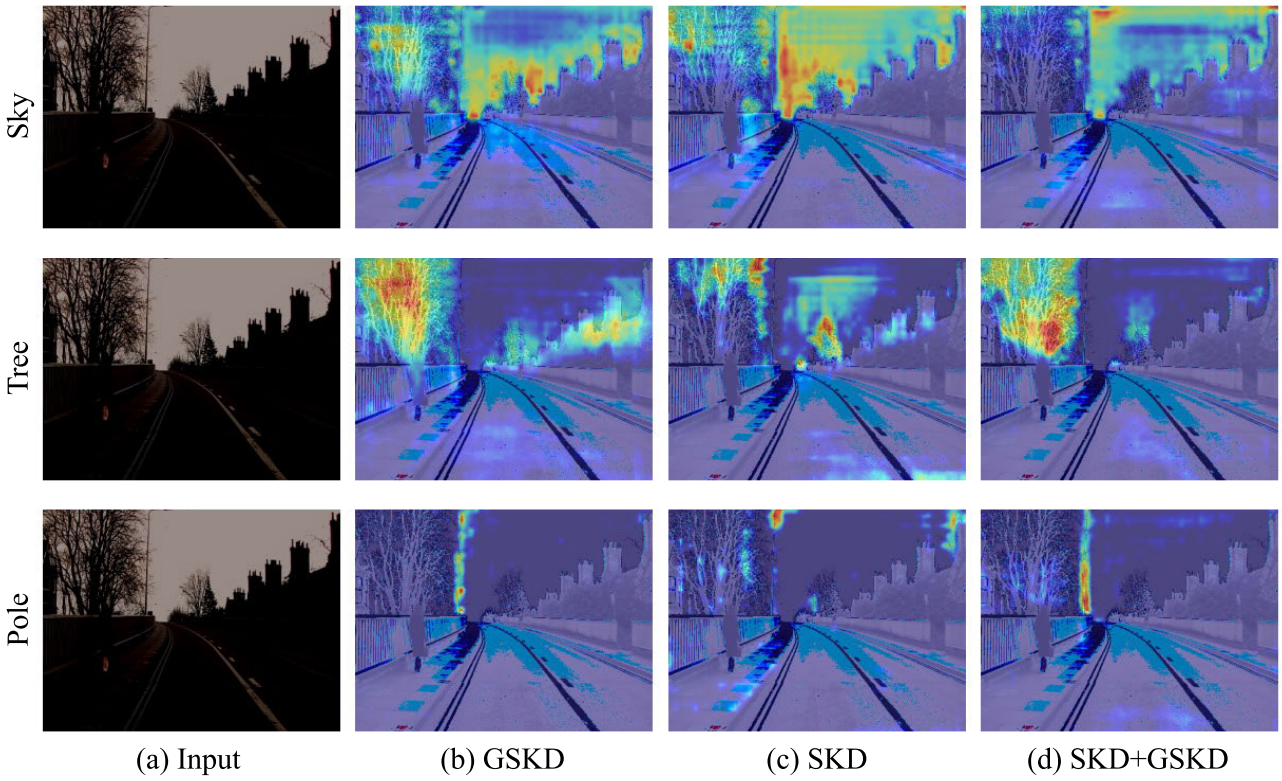

FIGURE 6. Grad-CAM visualizations on the CamVid test set using PSPNet-ResNet18 as the student and DeepLabV3-ResNet101 as the teacher. Each row corresponds to a target class (Sky, Tree, and Pole), and columns compare different distillation methods. GSKD highlights more semantically relevant regions compared to SKD, while SKD+GSKD further enhances spatial alignment by integrating both local and global structural cues.

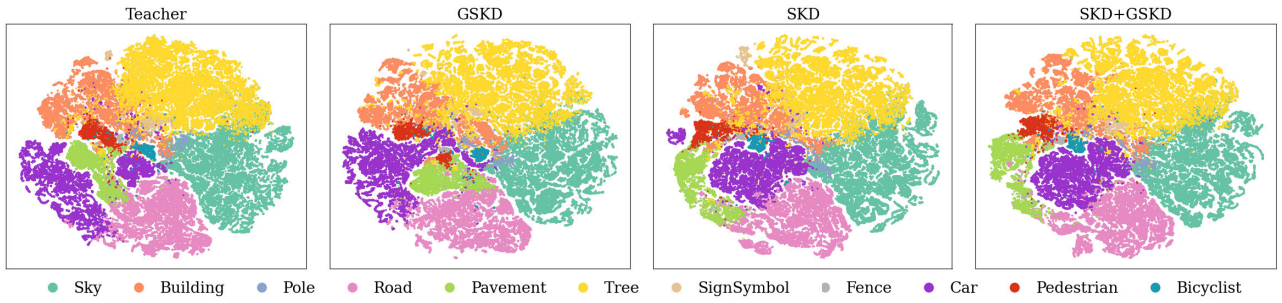

FIGURE 7. t-SNE visualizations of the student's final-layer features (PSPNet-ResNet18) on the CamVid test set. The model trained with GSKD shows tighter intra-class clustering and clearer inter-class separation than SKD. When combined (SKD+GSKD), the plug-in variant further improves the feature structure, closely matching the teacher (DeepLabV3-ResNet101) distribution.

framework. Table 6 shows results on the Cityscapes validation set using MiT-B4 as the teacher and MiT-B0 as the student. We perform feature distillation directly on the fused encoder features—immediately before the decoder input. To better leverage the spatial richness of the high-resolution feature maps in Segformer's MiT-B0 backbone, we increase the number of sampled points for CBS to $N = 32$. This enables the GSSM to capture global structural relations more effectively using diverse and distributed samples.

Despite the teacher's strong global modeling capacity, GSKD achieves a significant improvement of +0.85 mIoU over the student baseline (75.58 → 76.43). Furthermore, when combined with other KD methods such as SKD and DIST, adding gains are observed. These results suggest that while transformers inherently capture global dependencies through self-attention, GSKD introduces explicit structural supervision that enhances feature consistency across semantically related regions-leading to further performance improvements even on transformer-based architectures.

C. ABLATION STUDY

We conducted ablation studies on the Cityscapes using DeepLabV3-ResNet101 as the teacher and DeepLabV3-ResNet18 as the student. The study evaluates the impact of

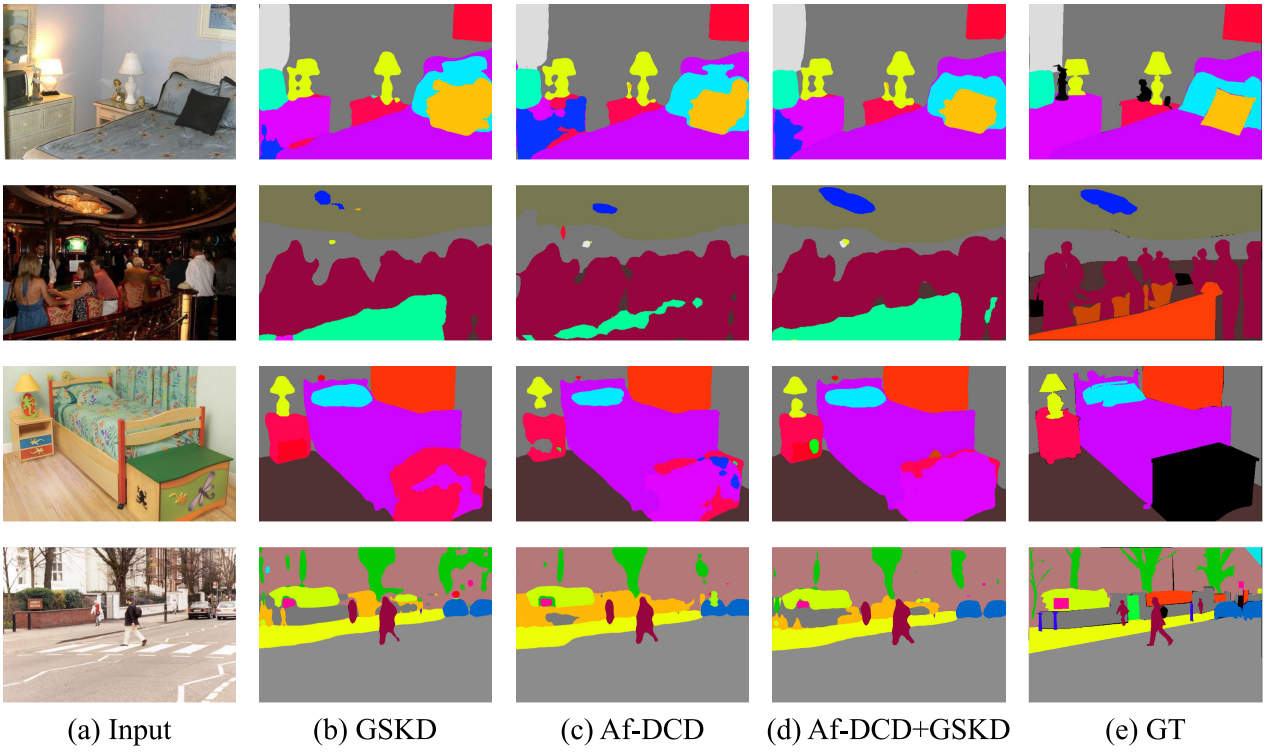

FIGURE 8. Qualitative segmentation results using DeepLabV3 with ResNet18 network on the validation set of ADE20K. Images in each column are: (a) input image, (b) GSKD (standalone), (c) Af-DCD, (d) Af-DCD+GSKD (plug-in) and (e) Ground Truth (GT).

TABLE 5. Comparison of quantitative results with state-of-the-art KD methods on ADE20K. * denotes that we reproduce the method using the code that the authors released to the public. T and S are the abbreviations for Teacher and Student.

Model	Params(M)	FLOPs(G)	Val mIoU(%)
T: DeepLabV3-ResNet101	61.1M	1294.6G	42.70
S: DeepLabV3-ResNet18	13.6M	61.0G	33.91
+ GSKD			34.75 (+0.84)
+ SKD [13]			35.17
+ SKD + GSKD			35.37 (+0.20)
+ Af-DCD* [15]			35.82
+ Af-DCD + GSKD			36.08 (+0.26)

TABLE 6. Comparison of quantitative results with state-of-the-art KD methods using Segformer on Cityscapes. T and S are the abbreviations for Teacher and Student.

Model	Params(M)	FLOPs(G)	Val mIoU(%)
T: Segformer-MiT-B4	64.1M	121.7G	81.23
S: Segformer-MiT-B0	3.8M	8.4G	75.58
+ GSKD			76.43 (+0.85)
+ SKD [13]			76.43
+ SKD + GSKD			76.77(+0.34)
+ DIST [32]			76.47
+ DIST + GSKD			76.62(+0.15)

sampling size (N), column size (S), and the effect of each component in GSKD. We also provide further ablation studies on GSKD when used as a plug-in module.

1) SAMPLING SIZE N FOR CBS

As shown in Table 7, a sampling size of $N = 16$ for Class-Balanced Sampling (CBS) produced the best performance for generating the GSSM. Larger samples introduced redundancy, while smaller sizes failed to capture sufficient global structural knowledge. While GSKD performs optimally with a sampling size of 16 when used independently, a smaller size of 8 proved to be more effective when used as a plug-in module as shown in Table 9. This suggests that when GSKD is integrated with other components, a smaller sampling size may be more effective in capturing the essential global structural information during CBS.

2) GROUPING STRIDE S IN SID

Table 8 illustrates that a column stride S of 8 yields optimal performance when generating sub-image descriptors (SID) for knowledge transfer, with the sampling size N fixed at 16. Increasing the stride reduced performance, indicating that a smaller stride is more effective in capturing essential structural patterns. We investigated the effect of stride size S as in a plug-in module as well in Table 10. The optimal performance is achieved when the stride S is set to the number of classes C present in an image. This suggests that aligning the stride with the class distribution allows the SIDs to more effectively capture and represent the structural patterns necessary for successful KD, resulting in improved model performance.

TABLE 7. Impact of sampling size N for Class-Balanced Sampling (CBS) on Cityscapes. Grouping stride S is fixed at 8.

Sampling Size (N)	8	16	32	64
Val mIoU (%)	75.19	**76.34**	75.91	75.18

TABLE 8. Impact of grouping stride S in Sub-Image Descriptors (SID) on Cityscapes. Sampling size N is fixed at 16.

Grouping Stride (S)	8	16	32	64
Val mIoU (%)	**76.34**	76.09	75.87	75.58

TABLE 9. Impact of Sampling Size N in Class-Balanced Sampling (CBS) for DIST+GSKD on Cityscapes. The optimal number of samples is $N = 8$.

Sampling Size (N)	4	8	16	32
Val mIoU (%)	77.97	**78.17**	78.00	78.01

TABLE 10. Impact of Grouping Stride S in Sub-Image Descriptors (SID) for DIST+GSKD on Cityscapes with a fixed sampling size $N = 8$. The optimal stride corresponds to the number of classes C presented in an image.

Grouping Size (S)	4	8	C
Val mIoU (%)	77.87	77.81	**78.17**

TABLE 11. Effect of each component on Cityscapes. RS: Random Sampling, CBS: Class-Balanced Sampling, CW: Column-Wise Grouping, RW: Row-Wise Shuffling.

	Configuration	RS	CBS	CW	RW	Val mIoU
	Vanilla KD					75.49
(a)	+ RS	✓				75.75
(b)	+ RS + CW	✓		✓		76.19
(c)	+ RS + CW + RW	✓		✓	✓	75.60
(d)	+ CBS		✓			76.09
(e)	+ CBS + CW		✓	✓		75.87
(f)	+ CBS + CW + RW		✓	✓	✓	**76.34**

3) EFFECT OF EACH COMPONENT

Table 11 highlights the contributions of each component in our framework. The results demonstrate that the combination of Class-Balanced Sampling (CBS), Column-Wise Grouping (CW), and Row-Wise Shuffling (RW) deliver the best performance, with each component playing a key role in improving segmentation results.

We first examine the impact of (a) random sampling (RS) versus (d) class-balanced sampling (CBS). CBS ensures a balanced representation across all classes, capturing more meaningful and relevant structural information from the teacher model's feature maps. In contrast, RS focuses on random pixels, which can introduce more noise and fail to capture important class-specific patterns. In short, CBS generates a more reliable Global Structural Similarity Map (GSSM), leading to better knowledge transfer from the teacher to the student model.

Next, we compare (b) RS+CW and (e) CBS+CW. Column-Wise Grouping (CW) generates Sub-Image Descriptors (SIDs) by grouping the similarity scores between sampled pixels through column-wise partitioning, capturing the structural relationships within subsets of sampled scores. When combined with RS, CW benefits from the diversity introduced by random sampling, leading to improved performance than (e). However, this process is highly dependent on the sampled pixels, as random sampling may introduce noise or irrelevant information, potentially degrading the quality of the descriptors. In contrast, when paired with CBS, CW groups similarity scores around fixed reference points based on column strides. This grouping strategy focuses on a limited set of reference points, reducing the diversity of the descriptors. As a result, this approach hampers the model's ability to capture the full range of global structural patterns, leading to lower performance compared to using the GSSM directly as transferred knowledge by applying only CBS in (d).

Finally, the combination of (f) CBS+CW+RW achieves the best performance, underscoring the critical role of Row-Wise Shuffling (RW). RW enhances the diversity of the Sub-Image Descriptors (SIDs) by preventing the model from focusing on a single reference point. It ensures that the similarity scores within each column are no longer concentrated around a fixed reference point, but instead become more diverse, capturing a wider range of relationships between different pixels. This increased diversity enables the student model to capture richer global structural patterns, leading to a more comprehensive, high-level understanding of the image's context. RW, therefore, plays a pivotal role in overcoming the limitations of CBS+CW by broadening the diversity of the relationships captured and facilitating a more effective transfer of global structural knowledge.

D. IMAGE CLASSIFICATION

In addition to semantic segmentation, we further investigated whether leveraging abstract, high-level global structural knowledge could improve the performance of the image classification task. Our experiments aim to verify that GSKD's focus on abstract, semantically rich representations can boost classification accuracy, particularly in cases where understanding the broader structure is more crucial than fine-grained pixel-level details. Since image classification assigns labels to entire images rather than individual pixels, we hypothesized that transferring global structural knowledge to capture broader patterns across the image would improve performance.

1) DATASET

We used the ImageNet-1K [57] dataset, a widely used benchmark for image classification tasks. The dataset consists of approximately 1.2 million training images, 50,000 validation images, and 100,000 test images, categorized into 1,000 classes.

TABLE 12. Image Classification Evaluation Results on ImageNet-1K using ResNet34 as a teacher with ResNet18 as a student.

Model	Top-1 Acc. (%)
T: ResNet34	73.31
S: ResNet18	69.76
+ KD [10]	70.66
+ OFD [54]	71.08
+ CRD [55]	71.17
+ SRRL [56]	71.73
+ Review [22]	71.61
+ DIST [32]	72.07
+ GSKD	74.08 (+3.42)
+ DIST [32] + GSKD	74.53 (+2.46)

2) TRAINING DETAILS

We followed the training procedures for ImageNet-1K classification as outlined in DIST [32], specifically adhering to the B1 strategy to ensure a fair comparison of our framework.

3) HYPER-PARAMETERS SETUP

For the Sub-Image Descriptors (SIDs) construction in image classification tasks, we generally maintained the same hyper-parameters as used in the semantic segmentation, where S corresponds to the number of classes per image. However, since image classification does not involve class distributions within a single image, we used random sampling instead of Class-Balanced Sampling (CBS), and S was set to a fixed value. In all subsequent experiments, the hyper-parameters for GSKD were set as follows: the sampling size for random sampling was set to $N = 16$, the grouping stride for SIDs was set to $S = 10$, and the weight value for L_{gskd} was set to $\lambda = 5$.

4) PERFORMANCE ON ImageNet-1K

The effectiveness of our KD method on the ImageNet-1K dataset, a crucial benchmark for image classification, is highlighted in Table 12. We employed ResNet34 as teacher model and observed significant performance improvements in ResNet18 as student model. Our method, when applied to existing KD methods, consistently demonstrates significant improvements in the classification task, with an average performance increase ranging from 1.9% to 3.4%. These results validate the efficacy of our approach in enhancing abstract information processing across a variety of student models and distillation strategies.

In the distillation from ResNet34 to ResNet18, our method not only improved the Top-1 accuracy by 3.42% compared to the standard KD method [10] but also achieved a 2.46% improvement when added to DIST [32]. These results demonstrate that the inclusion of our global structural similarity as an additional knowledge source effectively capitalizes on the nuances of abstract data representations, significantly boosting classification performance in complex datasets.

VI. CONCLUSION

We introduced Global Structural Knowledge Distillation (GSKD) to improve semantic segmentation by effectively transferring global structural patterns from a teacher model to a student model. Our approach complements existing fine-grained KD methods by introducing sub-image descriptors that capture higher-level, abstract visual information. Experiments show that GSKD significantly improves segmentation accuracy, with each module—class-balanced sampling, global structural similarity map, and sub-image descriptors—proving essential. The method's flexibility allows for seamless integration with existing KD techniques, boosting performance across benchmarks.

While GSKD demonstrates notable improvements, its reliance on class-balanced sampling may limit its effectiveness in scenarios where images contain a limited number of classes, leading to less significant performance gains. Additionally, further exploration of methods to better integrate local and global structural information could enhance the overall effectiveness of knowledge transfer.

Furthermore, our work is empirically driven, but extending global knowledge distillation toward theoretical guarantees—especially especially in dense prediction tasks—remains a valuable direction for future research.

ACKNOWLEDGMENT

(Hyejin Park and Keonhee Ahn contributed equally to this work.)

REFERENCES

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[3] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*.

[4] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7151–7160.

[5] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[6] Y. Cheng, D. Wang, P. Zhou, and T. Zhang, "Model compression and acceleration for deep neural networks: The principles, progress, and challenges," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 126–136, Jan. 2018.

[7] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," 2018, *arXiv:1803.03635*.

[8] J. Wu, C. Leng, Y. Wang, Q. Hu, and J. Cheng, "Quantized convolutional neural networks for mobile devices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4820–4828.

[9] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, and K. Keutzer, "A survey of quantization methods for efficient neural network inference," 2021, *arXiv:2103.13630*.

[10] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[11] A. Romero, N. Ballas, S. Ebrahimi Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for thin deep nets," 2014, *arXiv:1412.6550*.
[12] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017. [Online]. Available: https://openreview.net/pdf?id=Sks9_ajex
[13] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2599–2608.
[14] C. Yang, H. Zhou, Z. An, X. Jiang, Y. Xu, and Q. Zhang, "Cross-image relational knowledge distillation for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12309–12318.
[15] J. Fan, C. Li, X. Liu, M. Song, and A. Yao, "Augmentation-free dense contrastive knowledge distillation for efficient semantic segmentation," in *Proc. 37th Conf. Neural Inf. Process. Syst.*, Jan. 2023, pp. 51359–51370.
[16] C. Shen, Y. Huang, H. Zhu, J. Fan, and G. Zhang, "Student-oriented teacher knowledge refinement for knowledge distillation," in *Proc. 32nd ACM Int. Conf. Multimedia*, Oct. 2024, pp. 4543–4552.
[17] P. Luo, Z. Zhu, Z. Liu, X. Wang, and X. Tang, "Face model compression by distilling knowledge from neurons," in *Proc. AAAI Conf. Artif. Intell.*, Mar. 2016, vol. 30, no. 1, p. 10449.
[18] B. Heo, M. Lee, S. Yun, and J. Y. Choi, "Knowledge transfer via distillation of activation boundaries formed by hidden neurons," in *Proc. AAAI Conf. Artif. Intell.*, Jul. 2019, vol. 33, no. 1, pp. 3779–3787.
[19] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3962–3971.
[20] H. Choi, H. Lee, S. Kim, S. Kim, S. Kim, K. Sohn, and D. Min, "Adaptive confidence thresholding for monocular depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12788–12798.
[21] H. Park and D. Min, "Dynamic guidance adversarial distillation with enhanced teacher knowledge," in *Proc. Eur. Conf. Comput. Vis.*, Nov. 2024, pp. 204–219.
[22] P. Chen, S. Liu, H. Zhao, and J. Jia, "Distilling knowledge via knowledge review," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5008–5017.
[23] A. Amirkhani, A. Khosravian, M. Masih-Tehrani, and H. Kashiani, "Robust semantic segmentation with multi-teacher knowledge distillation," *IEEE Access*, vol. 9, pp. 119049–119066, 2021.
[24] C. Wang, J. Zhong, Q. Dai, Q. Yu, Y. Qi, B. Fang, and X. Li, "MTED: Multiple teachers ensemble distillation for compact semantic segmentation," *Neural Comput. Appl.*, vol. 35, no. 16, pp. 11789–11806, Jun. 2023.
[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 60, May 2017, pp. 84–90.
[26] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu, "SimMIM: A simple framework for masked image modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9653–9663.
[27] H. Choi, H. Park, K. Moo Yi, S. Cha, and D. Min, "Salience-based adaptive masking: Revisiting token dynamics for enhanced pre-training," 2024, *arXiv:2404.08327*.
[28] H. Choi, H. Lee, S. Joung, H. Park, J. Kim, and D. Min, "Emerging property of masked token for effective pre-training," 2024, *arXiv:2404.08330*.
[29] J. Xie, B. Shuai, J.-F. Hu, J. Lin, and W.-S. Zheng, "Improving fast segmentation with teacher–student learning," 2018, *arXiv:1810.08476*.
[30] H. Choi, H. Lee, W. Song, S. Jeon, K. Sohn, and D. Min, "Local-guided global: Paired similarity representation for visual reinforcement learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 15072–15082.
[31] C. Shu, Y. Liu, J. Gao, Z. Yan, and C. Shen, "Channel-wise knowledge distillation for dense prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 5291–5300.
[32] T. Huang, S. You, F. Wang, C. Qian, and C. Xu, "Knowledge distillation from a stronger teacher," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2022, pp. 33716–33727.
[33] T. Huang, Y. Zhang, S. You, F. Wang, C. Qian, J. Cao, and C. Xu, "Masked distillation with receptive tokens," in *Proc. 11th Int. Conf. Learn. Represent.*, Jan. 2022, pp. 1–12. [Online]. Available: https://openreview.net/forum?id=mWRngkvIki3
[34] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. D. Guo, M. Gheshlaghi Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent: A new approach to self-supervised learning," in *Proc. 34th Conf. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, Dec. 2020, pp. 21271–21284.
[35] X. Chen and K. He, "Exploring simple Siamese representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15750–15758.
[36] H. Choi, H. Lee, S. Jeong, and D. Min, "Environment agnostic representation for visual reinforcement learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 263–273.
[37] T. Huang, Y. Zhang, M. Zheng, S. You, F. Wang, C. Qian, and C. Xu, "Knowledge diffusion for distillation," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2023, pp. 65299–65316.
[38] Y. Zhang, T. Huang, J. Liu, T. Jiang, K. Cheng, and S. Zhang, "FreeKD: Knowledge distillation via semantic frequency prompt," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 15931–15940.
[39] Y. Wang, W. Zhou, T. Jiang, X. Bai, and Y. Xu, "Intra-class feature variation distillation for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, Jan. 2020, pp. 346–362.
[40] Z. Fang, J. Wang, L. Wang, L. Zhang, Y. Yang, and Z. Liu, "SEED: Self-supervised distillation for visual representation," in *Proc. Int. Conf. Learn. Represent.*, 2021. [Online]. Available: https://openreview.net/forum?id=AHm3dbp7D1D
[41] Q. Song, J. Li, H. Guo, and R. Huang, "Denoised non-local neural network for semantic segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 5, pp. 7162–7174, May 2024.
[42] S. Kim, D. Min, B. Ham, S. Ryu, M. N. Do, and K. Sohn, "DASC: Dense adaptive self-correlation descriptor for multi-modal and multi-spectral correspondence," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2103–2112.
[43] E. Shechtman and M. Irani, "Matching local self-similarities across images and videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8.
[44] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: Binary robust independent elementary features," in *Proc. Eur. Conf. Comput. Vis.*, Jan. 2010, pp. 778–792.
[45] M. Calonder, V. Lepetit, M. Ozuysal, T. Trzcinski, C. Strecha, and P. Fua, "BRIEF: Computing a local binary descriptor very fast," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1281–1298, Jul. 2012.
[46] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.
[47] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognit. Lett.*, vol. 30, no. 2, pp. 88–97, Jan. 2009.
[48] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.
[49] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, "Semantic understanding of scenes through the ADE20K dataset," *Int. J. Comput. Vis.*, vol. 127, no. 3, pp. 302–321, Mar. 2019.
[50] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: https://openreview.net/pdf?id=ZzwDy_wiWv
[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[52] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.
[53] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2021, pp. 12077–12090.
[54] B. Heo, J. Kim, S. Yun, H. Park, N. Kwak, and J. Y. Choi, "A comprehensive overhaul of feature distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1921–1930.

[55] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," 2019, *arXiv:1910.10699*.
[56] J. Yang, B. Martinez, A. Bulat, and G. Tzimiropoulos, "Knowledge distillation via softmax regression representation learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.
[57] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

HYEJIN PARK received the B.S. degree in multimedia from Sungkyul University, Anyang, South Korea, in 2005, and the M.S. degree in computer science and engineering from Ewha Womans University, Seoul, South Korea, in 2010, where she is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering. From 2013 to 2018, she was a Research Associate with Nanyang Technological University, Singapore. Her current research interests include computer vision and deep learning, with a focus on adversarial attack/defense and model compression.

KEONHEE AHN received the B.S. and M.S. degrees in computer science and engineering from Ewha Womans University, Seoul, South Korea, in 2022 and 2024, respectively. Her current research interests include computer vision and deep learning.

HYESONG CHOI (Student Member, IEEE) received the B.S. degree from the Department of Computer Science and Engineering, Ewha Womans University, Seoul, South Korea, in 2019, where she is currently pursuing the Ph.D. degree. Her current research interests include computer vision and deep learning.

DONGBO MIN (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the School of Electrical and Electronic Engineering, Yonsei University, Seoul, South Korea, in 2003, 2005, and 2009, respectively. From 2009 to 2010, he was a Postdoctoral Researcher with the Mitsubishi Electric Research Laboratories, Cambridge, MA, USA. From 2010 to 2015, he was with the Advanced Digital Sciences Center, Singapore. From 2015 to 2018, he was an Assistant Professor with the Department of Computer Science and Engineering, Chungnam National University, Daejeon, South Korea. Since 2018, he has been with the Department of Computer Science and Engineering, Ewha Womans University, Seoul. His current research interests include computer vision, deep learning, and video processing.