IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 31 July 2024, accepted 11 October 2024, date of publication 18 October 2024, date of current version 11 November 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3483434

RESEARCH ARTICLE

Dynamic Grouping With Multi-Manifold Attention for Multi-View 3D Object Reconstruction

GEORGIOS KALITSIOS, DIMITRIOS KONSTANTINIDIS, PETROS DARAS, (Senior Member, IEEE), AND KOSMAS DIMITROPOULOS

Information Technologies Institute, Centre for Research and Technology Hellas (CERTH), 57001 Thessaloniki, Greece

Corresponding author: Georgios Kalitsios (gkalitsios@iti.gr)

This work was supported by European Commission's Horizon Europe Project "DAFNEplus: Decentralized Platform for Fair Creative Content Distribution Empowering Creators and Communities through New Digital Distribution Models Based on Digital Tokens" under Grant 101061548.

ABSTRACT In a multi-view 3D reconstruction problem, the task is to infer the 3D shape of an object from various images taken from different viewpoints. Transformer-based networks have demonstrated their ability to achieve high performance in such problems, but they face challenges in identifying the optimal way to merge the different views in order to estimate with great fidelity the 3D shape of the object. This work aims to address this issue by proposing a novel approach to compute information-rich inter-view features by combining image tokens with similar distinctive characteristics among the different views dynamically. This is achieved by leveraging the self-attention mechanism of a Transformer, enhanced with a multi-manifold attention module, to estimate the importance of image tokens on-the-fly and re-arrange them among the different views in a way that improves the viewpoint merging procedure and the 3D reconstruction results. Experiments on ShapeNet and Pix3D validate the ability of the proposed method to achieve state-of-the-art performance in both multi-view and single-view 3D object reconstruction.

INDEX TERMS Dynamic grouping, multi-manifold attention, multi-view 3D reconstruction, transformer, voxel representation.

I. INTRODUCTION

The task of 3D object reconstruction is a fundamental research problem with several applications across diverse fields, including computer vision, computer graphics, robot navigation, medical imaging, virtual reality (VR), and Non-Fungible Token (NFT) creation [22]. One of the main challenges of 3D object reconstruction lies in the accurate identification and matching of distinctive features of an object across multiple images, which is crucial for determining the spatial properties of the object to reconstruct it with high fidelity [23].

With the technological breakthroughs in AI, several deep learning methods rely on Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) to solve the 3D object reconstruction task with great success [3], [8], [9], [10], [24]. Recently, Transformer networks have also been utilized for 3D object reconstruction, demonstrating outstanding performance. Transformers operate by dividing input images into several smaller patches, which are then processed in parallel. Self-attention mechanisms are also utilized to learn correlations among image patches (or tokens) and enhance the network's understanding for its task. By leveraging image tokens and attention operations, several studies have employed Transformers to explore the association of tokens within a view and across different views, leading to improvements in the final 3D reconstruction results [1], [2], [4], [5].

While early Transformer-based 3D object reconstruction methods process view-specific image tokens independently

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

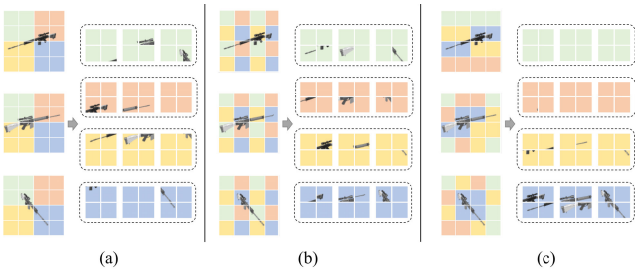

FIGURE 1. While previous approaches utilize (a) short- and (b) long-range token groupings based on fixed positions, thus mixing foreground object parts and background, the proposed method (c) performs an attention-based dynamic grouping, allowing the disentanglement of foreground object parts from the background in the subsequent token groups.

to extract descriptive intra-view representations [4], [5], recent approaches group image tokens across views either through clustering [2] or based on fixed image positions [1]. However, such approaches either model only intra-view relations or identify both intra- and inter-view relations at high computational cost or by employing rigid grouping strategies, in which the groupings are not learnable and the grouped tokens contain both foreground object parts and background regions. To overcome these issues, the proposed Transformer-based 3D reconstruction method introduces a novel dynamic grouping strategy that enables the network to learn on its own the token groupings across the different views dynamically. By leveraging the attention mechanism of Transformers, the proposed grouping strategy evaluates the importance of the tokens based on their attention values, allowing the disentanglement of foreground object parts from the background and the grouping of relevant tokens across views, as shown in Figure 1. Coupled with a multi-manifold attention mechanism that boosts the discrimination ability of the computed attention maps, the proposed method can model both intra- and inter-view associations to infer with high accuracy the 3D shape of an object in the form of voxels. In summary, the main contributions of this work are:

- A novel dynamic grouping strategy that effectively models inter-view relations by merging image tokens across views on-the-fly based on their attention values.
- The attention mechanism is enhanced by projecting tokens in different manifolds for optimal attention estimation and token grouping results.
- Experimental results on ShapeNet and Pix3D verify the state-of-the-art performance of the proposed method in both multi- and single-view 3D object reconstruction.

II. RELATED WORK

In multi-view 3D object reconstruction, traditional methods [25], [26] relied on handcrafted feature extraction and matching across views, which was time-consuming and prone to inaccuracies as highly overlapping views and similar lighting conditions were assumed. Early deep learning 3D object reconstruction techniques were based on CNN [3], [7], [9], [10], [28], RNN [8], [27], Generative Adversarial Networks (GANs) [31], [32] and Variational Autoencoders (VAEs) [33], [34], [35] that automatically processed and fused multiple view-specific images without searching for discriminative common features across views. Transformers have been met with great research interest when they were initially introduced for natural language processing [30] and later modified for computer vision tasks [29]. For 3D object reconstruction, in particular, Transformer-based methods leverage image patch embeddings to capture intra-view correlations and then fuse the view-specific representations using pooling operations [5], [12] or Transformer encoders [4], [6], [11] that jointly model all 2D views into 3D reconstruction outputs. Despite achieving promising results, these methods cannot capture strong inter-view relations as they treat all image tokens across views equally.

To robustly model multi-view associations, several researchers have lately proposed token grouping techniques to merge similar tokens across views and enhance the accuracy of 3D object reconstruction. STTN [36] addresses this problem in the task of video inpainting by employing full-range attention (i.e., attention among all image tokens across views), leading to high computational cost. To mitigate this issue, DSTT [37] employs a short-range grouping attention strategy that enables the model to capture local information, while paying less attention to non-local relationships. However, this technique works best when images demonstrate temporal coherence and thus it is of limited usage in 3D object reconstruction. To address this problem, UMIFormer [2] employed nearest neighbor clustering to group similar tokens and estimate both intra-view and inter-view relations. On the other hand, LRGT [1] merged tokens from fixed positions in the images, successfully capturing global inter-view associations, but the grouped tokens do not necessarily represent similar characteristics of the images. Our work introduces a novel dynamic grouping strategy that allows the Transformer network to dynamically group tokens across views based on the computed attention maps. In this way, tokens sharing similar distinctive characteristics across views, being either foreground or background, are grouped together, thus enabling the decoupling of information and significantly simplifying the 3D reconstruction procedure.

III. METHODOLOGY

The proposed multi-view 3D object reconstruction method, named shortly DGMA, is based on a Transformer-based architecture modified with a novel module that performs enhanced multi-manifold self-attention and dynamic token grouping, as illustrated in Figure 2. The motivation behind DGMA lies in the need for an optimal way to merge tokens to extract information-rich inter-view features and reconstruct with high fidelity a 3D object. DGMA achieves this dynamically (i.e., during the network training) by

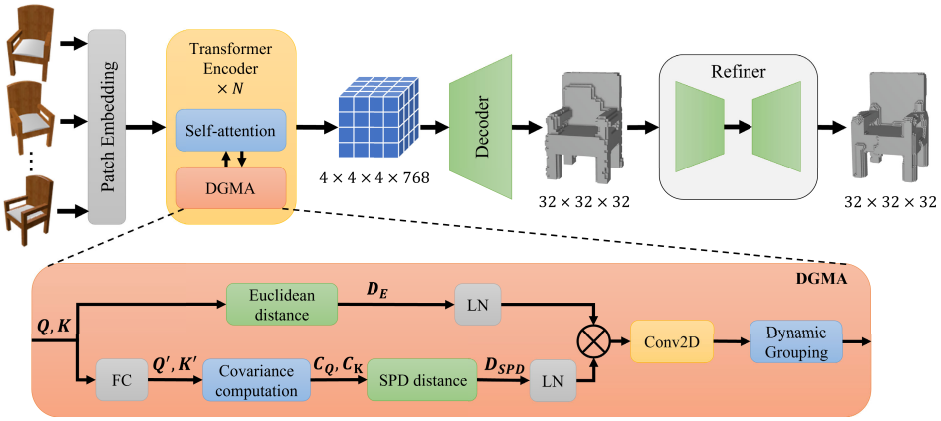

FIGURE 2. Illustration of the architecture of the proposed DGMA method.

adjusting on its own the token groupings based on attention values, enabling the efficient disentanglement of foreground and background information, as shown in Figure 3. Given a view image set $I = \{I_1, I_2, \ldots, I_V\}$ for V number of views, it is initially split into patches through a 2D convolution operation before being fed into a number N of Transformer Encoders that processes these patches and groups them into more descriptive representations.

A. MULTI-MANIFOLD ATTENTION COMPUTATION

To enhance the discrimination power of the computed attention maps inside the Transformer encoders and improve the ability of DGMA to group relevant tokens across views, the tokens are projected in two different manifolds and distances are computed in each manifold separately before being fused together. With the projection in manifolds, the aim is to leverage the different geometrical properties of the manifolds and derive a better estimation of the actual distances among tokens, leading to improved attention maps and better reconstruction results, as demonstrated in other computer vision tasks, such as image classification [21]. In this work, DGMA utilizes the Euclidean and Semi Positive Definite (SPD) manifolds for time efficiency and adds new layer normalization units to scale the computed distance maps in the range [0,1] for improved accuracy.

Given view-specific image patches or tokens $X^v = \{x_i^v\} \in \mathbb{R}^{P\times D}$, with P and D denoting their number and dimensionality, respectively, for each view image I_v, query and key vectors $Q_v, K_v \in \mathbb{R}^{P\times D}$ are initially computed through linear projection operations. Then, the distance map $D_E^v \in \mathbb{R}^{P\times P}$ of the vectors in the Euclidean manifold is computed as follows:

$$D_E^v(Q_v, K_v) = \frac{Q_v K_v^T}{\sqrt{D}} \tag{1}$$

On the other hand, for the computation of the distance map in the SPD manifold, a fully connected layer is initially employed to down-sample the query and key vectors to the covariance size S, resulting in a set of new vectors $Q'_v, K'_v \in \mathbb{R}^{P\times S}$. Afterwards, the covariance matrices $C_{Q_v}, C_{K_v} \in \mathbb{R}^{P\times P}$ of the query and key vectors are computed as:

$$C_{Q_v} = cov(Q'_v) = E[(Q'_v - E[Q'_v])(Q'_v - E[Q'_v])^T] \tag{2}$$

$$C_{K_v} = cov(K'_v) = E[(K'_v - E[K'_v])(K'_v - E[K'_v])^T] \tag{3}$$

The distance map $D_{SPD}^v \in \mathbb{R}^{P\times P}$ between the covariance matrices in the Semi-Positive Definite (SPD) manifold is then computed using the Frobenius norm $\|\cdot\|_F$ as:

$$D_{SPD}^v(C_{Q_v}, C_{K_v}) = \frac{\|C_{Q_v} - C_{K_v}\|_F}{\sqrt{S}} \tag{4}$$

The attention map $A^v = \{a_{i,j}\} \in \mathbb{R}^{P\times P}$ for a single view v is finally given by the concatenation of the normalized distance maps as:

$$A^v = f(LN(D_E^v), LN(D_{SPD}^v)) \tag{5}$$

In Eq. 5, LN denotes the layer normalization operation, while $f(\cdot)$ denotes a convolutional operation that fuses the distance maps, followed by a softmax operation to compute the actual attention values.

B. TOKEN SORTING

The aim of token sorting is to identify which tokens carry important information according to the Transformer encoder. Tokens with high attention values represent distinctive characteristics of the input that facilitate the network in making correct decisions. DGMA leverages this knowledge to disentangle important foreground parts of the object from irrelevant parts of the object or the background. Given the view-specific attention map A^v, computed in Eq. 5, the attention values of all tokens in the representation X^v are taken as the diagonal elements of the attention map. Then, the tokens X^v are sorted from the smallest to the highest attention values, giving rise to the token representation $Y^v \in \mathbb{R}^{P\times D}$:

$$Y^v = \{sort(x_i^v) | a_{i,i} < a_{i+1,i+1}\}, \text{ for } i = 1, \ldots, P-1 \tag{6}$$

In the representation Y^v, the first tokens usually represent irrelevant information (e.g., image background), while the

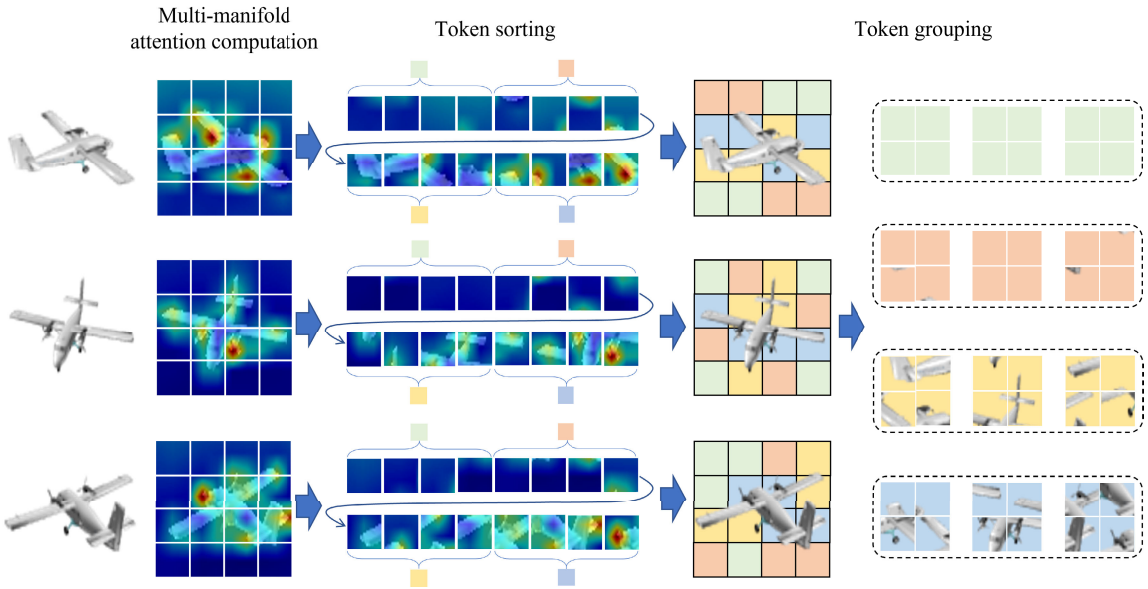

FIGURE 3. Illustration of the proposed dynamic grouping strategy. Leveraging the computed attention maps, view-specific tokens are sorted dynamically based on their attention values and placed in groups with each group, indicated by a different color, containing tokens of similar distinctive characteristics.

last tokens carry significant information (e.g., foreground object parts).

C. TOKEN GROUPING

The token grouping aims to gather tokens that represent similar distinctive characteristics across views in the same groups. In this way, the 3D reconstruction process is simplified as each token group carry information from specific parts of the input image, being foreground or background, leading to the modelling of information-rich inter-view relations and the significant improvement in the reconstruction results. Given the sorted tokens y_i^v in the representation Y^v, this procedure gathers each consecutive 4 tokens across all V views and group them together, forming a number of groupings:

$$G_j = \{y^1_{4(j-1)+1:4j}, y^2_{4(j-1)+1:4j}, \ldots, y^V_{4(j-1)+1:4j}\} \quad \text{for } j = 1, \ldots, P/4 \tag{7}$$

It can be observed that the first groups contain tokens with least significant information across views, while the latest tokens contain tokens with discriminative information across views. The groupings G_j are finally concatenated and fed to the next layer of the network. Figure 3 illustrates this procedure, with each token group shown in different color.

D. 3D VOLUME ESTIMATION

The token representation after the last Transformer encoder is finally compressed using the similar-token merger [2] and an initial 3D representation of size $4 \times 4 \times 4 \times 768$ is created. Afterwards, the initial 3D representation passes through a decoder [1] that up-scales the 3D representation to the required size of $32 \times 32 \times 32$. Finally, a refiner with a U-Net architecture [3] is employed to refine and improve the 3D representation by adding details, leading to the final 3D reconstructed object.

IV. EXPERIMENTS

A. DATASETS AND EVALUATION METRICS

To evaluate the proposed method, two well-known datasets, namely ShapeNet [13] and Pix3D [14] are employed. ShapeNet is a large dataset comprising 55 object categories and 51, 300 3D models. For fair comparison with other works, a subset of ShapeNet, consisting of 43, 783 objects rendered from 24 distinct viewpoints and from 13 categories, is utilized [9]. In contrast to the artificially generated objects of ShapeNet, Pix3D comprises aligned 3D models paired with real-world 2D images. To explore the generalization ability of the proposed method, a subset of 2894 samples, including untruncated and unoccluded view images, from the chair category of Pix3D is used for evaluation.

Similar to previous works, the 3D reconstruction performance is evaluated using 3D Intersection over Union (IoU) and F-Score@1%, with higher values indicating better performance for both metrics. IoU compares the predicted occupancy volume against the ground truth volume by quantifying the ratio of intersecting voxels between both volumes to their union, thus ensuring the calculation's independence from object size. On the other hand, F-Score@1% [15] evaluates the surface reconstruction quality by quantifying the percentage of points on object surfaces that fall within a predetermined threshold distance (1%). In line with [9], the predicted volumes are converted to point clouds and object surfaces are generated using the

marching cubes algorithm [16]. Subsequently, 8, 192 points are sampled from the object surface to compute the F-Score between predictions and ground truth.

B. IMPLEMENTATION DETAILS

For fair comparison with other works [1], [2], two models, namely DGMA and DGMA+, are considered that share the same network architecture but differ in the number of views used for training (3 views for DGMA and 8 views for DGMA+). The choice of fixed view counts is based on the observation that maintaining a consistent view count leads to superior performance compared to varying it between steps [4]. At each iteration, views are randomly sampled from the set of 24 views per object. To obtain the occupancy voxel grid, thresholds of 0.5 and 0.4 are applied for DGMA and DGMA+, respectively. The input images are of size 224×224, while the voxelized output has a size of $32 \times 32 \times 32$. To initialize the Transformer encoders, the pre-trained model of DeiT-B [20] is employed. The number of Transformer encoders N is set to 12, and the DGMA mechanism is incorporated in Transformers 5, 7, 9 and 11, while the remaining ones use the vanilla self-attention mechanism. The covariance size S is set to 12 after conducting ablation experiments. The loss function employed is the Dice loss [38], which is effective for addressing significant imbalances in the voxel grid. The training process involves the use of AdamW optimizer [19], with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and training for 110 epochs with a batch size of 16. The learning rate is initialized at $1e-4$ and is subsequently reduced by a factor of 0.1 after 60 and 90 epochs. DGMA and DGMA+ were trained on two NVIDIA RTX 3090 Ti GPUs. The code and model weights will be made publicly available on GitHub.[1]

C. EVALUATION ON SHAPENET

A comparison of DGMA and DGMA+ against state-of-the-art (SoTA) approaches using the ShapeNet dataset is conducted and the results are presented in Tables 1 and 2 in terms of IoU and F-Score@1%, respectively. It can be deduced that DGMA and DGMA+ significantly outperform the SoTA approaches for different number of views, ranging from 1 to 20. For the single-view reconstruction, DGMA achieves 0.6987 IoU, outperforming the second best LRGT with 0.6962 IoU. On the other hand, for 20 views, DGMA+ achieves SoTA performance with 0.7974 IoU, 0.7% higher than the accuracy of LRGT+. The performance improvements are even more pronounced in F-Score@1%, indicating that DGMA and DGMA+ can reconstruct with higher accuracy the object surface.

When it comes to multi-view reconstruction, a main concern is how well a model can handle a large number of inputs. To this end, DGMA and DGMA+ are compared against the SoTA works of [1], [2], [3] for 24-view inputs of the ShapeNet dataset. Tables 3 and 4 present the IoU and F-Score@1% results across the 13 ShapeNet categories. It can be seen that DGMA and DGMA+ exhibit superior performance compared to their competitors. For the display object category, specifically, DGMA+ outperforms LRGT+ by 1.13% in IoU and 1.43% in F-Score@1%, while an improvement of 1.23% in IoU and 1.97% in F-Score@1% can be noticed for the lamp object category. Similar improvements can be observed for the other object categories as well, except from telephone, in which UMIFormer+ slightly outperforms the proposed models.

Finally, Figure 4 visualizes six object reconstructions utilizing 5 and 20 input views from the test set of ShapeNet. DGMA and DGMA+ demonstrate superior performance in accurately reconstructing both the general shape and intricate details of the objects. This underscores the robust learning capability of the proposed dynamic grouping strategy and the multi-manifold attention mechanism, particularly in capturing fine-grained details. In addition, DGMA and DGMA+ manage to reduce noise levels and produce smoother surfaces with higher accuracy and fidelity than the other tested 3D reconstruction techniques.

The aim of experimenting with Pix3D is to evaluate the generalization capacity of the proposed method in single-view 3D reconstruction under real-world scenarios with complex backgrounds. Consistent with prior studies [9], [10], a training dataset was formed by utilizing the chair objects from ShapeNet, with view images rendered by RenderforCNN [17] and random backgrounds from the SUN database [18]. Each object in the dataset is represented by a set of 60 images captured from various viewpoints. Table 5 compares the performance of the proposed method against other SoTA approaches, presented in [1], [2], [3]. All methods were trained for single-view 3D reconstruction using the generated training dataset and according to the implementation details provided in their respective papers. In LRGT†, both LGA and IFS were removed and substituted with the original self-attention layer. In a similar fashion, DGMA† is the result of substituting all DGMA modules with vanilla self-attention. Since DGMA is designed to group multi-view inputs, the performance of the proposed method is hindered in Pix3D that contains only a single view for each object. From the results of Table 5, it can be deduced that DGMA outperforms all other methods by at least 0.8% in IoU and 1.5% in F1-score@1%. The removal of the DGMA modules leads to an additional performance improvement, making DGMA† achieve SoTA performance, surpassing LRGT† by 0.6% in IoU and 0.3% in F1-score@1%. Finally, Figure 5 performs a qualitative analysis of the 3D reconstruction results in Pix3D. It can be seen that the proposed DGMA and DGMA† can accurately estimate the 3D shapes of the objects by successfully reconstructing the legs of the chairs, while other SoTA approaches face challenges in this area. The results in Pix3D demonstrate that the proposed method can excel even in scenarios where objects and background share similar

[1] https://github.com/VCLdimitrop/DGMA

TABLE 1. Performance evaluation of multi-view 3D reconstruction methods on ShapeNet using IoU ↑. Best results are highlighted in bold.

Methods	Number of Views								
	1	2	3	4	5	8	12	16	20
3D-R2N2 [8]	0.5600	0.6030	0.6170	0.6250	0.6340	0.6350	0.6360	0.6360	0.6360
AttSets [7]	0.6420	0.6620	0.6700	0.6750	0.6770	0.6850	0.6680	0.6920	0.6930
Pix2Vox/A [10]	0.6610	0.6860	0.6930	0.6970	0.6990	0.7020	0.7040	0.7050	0.7060
Pix2Vox++/A [9]	0.6700	0.6950	0.7040	0.7080	0.7110	0.7150	0.7170	0.7180	0.7190
Legoformer [4]	0.5190	0.6440	0.6790	0.6940	0.7030	0.7130	0.7170	0.7190	0.7210
EVoIT [6]	-	-	-	0.6090	-	0.6980	0.7200	0.7290	0.7350
3D-RETR [5]	0.6740	0.7070	0.7160	0.7200	0.7230	0.7270	0.7290	0.7300	0.7310
3D-C2FT [11]	0.6290	0.6780	0.6950	0.7020	0.7080	0.7160	0.7200	0.7230	0.7250
GARNet [3]	0.6730	0.7050	0.7160	0.7220	0.7260	0.7310	0.7340	0.7360	0.7370
GARNet+ [3]	0.6550	0.6960	0.7120	0.7190	0.7250	0.7330	0.7370	0.7400	0.7420
UMIFormer [2]	0.6802	0.7384	0.7518	0.7573	0.7612	0.7661	0.7682	0.7696	0.7702
UMIFormer+ [2]	0.5672	0.7115	0.7447	0.7588	0.7681	0.7790	0.7843	0.7873	0.7886
LRGT [1]	0.6962	0.7462	0.7590	0.7653	0.7692	0.7744	0.7766	0.7781	0.7786
LRGT+ [1]	0.5847	0.7145	0.7476	0.7625	0.7719	0.7833	0.7888	0.7912	0.7922
DGMA	**0.6987**	**0.7487**	**0.7632**	**0.7692**	0.7734	0.7781	0.7800	0.7810	0.7812
DGMA+	0.5961	0.7173	0.7518	0.7666	**0.7767**	**0.7876**	**0.7934**	**0.7962**	**0.7974**

TABLE 2. Performance evaluation of multi-view 3D reconstruction methods on ShapeNet using F1-score@1% ↑. Best results are highlighted in bold.

Methods	Number of Views								
	1	2	3	4	5	8	12	16	20
3D-R2N2 [8]	0.351	0.368	0.372	0.378	0.382	0.383	0.382	0.382	0.383
AttSets [7]	0.395	0.418	0.426	0.430	0.432	0.444	0.445	0.447	0.448
Pix2Vox/F [10]	0.364	0.393	0.404	0.409	0.412	0.417	0.420	0.423	0.423
Pix2Vox++/F [9]	0.394	0.422	0.432	0.437	0.440	0.446	0.449	0.450	0.451
Legoformer [4]	0.282	0.392	0.428	0.444	0.453	0.464	0.470	0.472	0.472
EVoIT [6]	-	-	-	0.358	-	0.448	0.475	0.486	0.492
3D-RETR [5]	-	-	-	-	-	-	-	-	-
3D-C2FT [11]	0.371	0.424	0.443	0.452	0.458	0.468	0.476	0.477	0.479
GARNet [3]	0.418	0.455	0.468	0.475	0.479	0.486	0.489	0.491	0.492
GARNet+ [3]	0.399	0.446	0.465	0.475	0.481	0.491	0.498	0.501	0.504
UMIFormer [2]	0.4281	0.4919	0.5067	0.5127	0.5168	0.5213	0.5232	0.5245	0.5251
UMIFormer+ [2]	0.3177	0.4568	0.4947	0.5104	0.5216	0.5348	0.5415	0.5451	0.5466
LRGT [1]	0.4461	0.5005	0.5148	0.5214	0.5257	0.5311	0.5337	0.5347	0.5353
LRGT+ [1]	0.3378	0.4618	0.4989	0.5161	0.5271	0.5403	0.5467	0.5497	0.5510
DGMA	**0.4518**	**0.5041**	**0.5204**	**0.5265**	0.5307	0.5353	0.5375	0.5380	0.5383
DGMA+	0.3475	0.4668	0.5057	0.5227	**0.5345**	**0.5470**	**0.5539**	**0.5570**	**0.5585**

TABLE 3. Performance comparison of 24-view 3D reconstruction on the ShapeNet test dataset using Intersection over Union (IoU). The highest score in each category is highlighted in bold.

Category	Methods							
	GARNet [3]	GARNet+ [3]	UMIFormer [2]	UMIFormer+ [2]	LRGT [1]	LRGT+ [1]	DGMA	DGMA+
airplane	0.724	0.739	0.769	0.789	0.778	0.793	0.780	**0.798**
bench	0.698	0.707	0.738	0.761	0.753	0.768	0.754	**0.773**
cabinet	0.841	0.840	0.861	0.877	0.869	0.881	0.872	**0.882**
car	0.888	0.894	0.895	0.903	0.900	0.904	0.900	**0.907**
chair	0.674	0.683	0.713	0.735	0.722	0.744	0.724	**0.747**
display	0.668	0.665	0.742	0.768	0.750	0.767	0.756	**0.778**
lamp	0.516	0.513	0.570	0.610	0.580	0.611	0.586	**0.623**
speaker	0.773	0.772	0.820	0.840	0.825	0.839	0.825	**0.841**
rifle	0.697	0.709	0.760	0.784	0.763	0.783	0.778	**0.792**
sofa	0.807	0.810	0.825	0.840	0.834	**0.846**	0.836	**0.846**
table	0.693	0.692	0.726	0.744	0.737	0.748	0.739	**0.758**
telephone	0.871	0.879	0.887	**0.904**	0.895	0.898	0.890	0.901
watercraft	0.693	0.696	0.723	0.745	0.734	0.747	0.734	**0.748**
Overall	0.737	0.742	0.771	0.790	0.779	0.793	0.781	**0.798**

textures, objects are positioned at varying distances within complex backgrounds that cast shadows on the objects.

D. ABLATION STUDY

To evaluate the contribution of each component, i.e., Dynamic Grouping (DG) strategy and enhanced Multi-Manifold Attention (MA), in the performance of the proposed method, we test DGMA for 3 input views on ShapeNet. Table 6 presents the experimental results using the metric of IoU. In all experiments, the same decoder and refiner networks are utilized to ensure consistency. More specifically, the decoder network is a Transformer

TABLE 4. Performance comparison of 24-view 3D reconstruction on the ShapeNet test dataset using F1-score@1%. The highest score in each category is highlighted in bold.

Category	Methods							
	GARNet [3]	GARNet+ [3]	UMIFormer [2]	UMIFormer+ [2]	LRGT [1]	LRGT+ [1]	DGMA	DGMA+
airplane	0.606	0.628	0.667	0.691	0.678	0.696	0.680	**0.704**
bench	0.536	0.551	0.498	0.600	0.591	0.607	0.595	**0.614**
cabinet	0.473	0.473	0.498	0.515	0.498	0.520	0.507	**0.527**
car	0.608	0.623	0.622	0.641	0.633	0.643	0.635	**0.649**
chair	0.369	0.384	0.399	0.419	0.408	0.428	0.410	**0.435**
display	0.386	0.396	0.454	0.485	0.466	0.485	0.474	**0.499**
lamp	0.366	0.369	0.410	0.451	0.422	0.456	0.420	**0.465**
speaker	0.338	0.346	0.392	0.418	0.399	0.418	0.398	**0.422**
rifle	0.634	0.647	0.707	0.736	0.711	0.734	0.729	**0.746**
sofa	0.489	0.500	0.505	0.528	0.521	0.536	0.525	**0.539**
table	0.449	0.452	0.467	0.481	0.476	0.485	0.481	**0.496**
telephone	0.698	0.716	0.709	**0.736**	0.719	0.726	0.715	0.734
watercraft	0.494	0.504	0.534	0.567	0.55	0.571	0.550	**0.575**
Overall	0.493	0.505	0.525	0.548	0.536	0.552	0.539	**0.560**

FIGURE 4. Qualitative results on ShapeNet test set with 5 and 20 input views.

TABLE 5. Evaluation on Pix3D using IoU / F1-score@1% ↑.

GARNet [3]	UMIFormer [2]	LRGT [1]	LRGT† [1]	DGMA	DGMA†
0.167 / 0.092	0.312 / 0.153	0.305 / 0.145	0.318 / 0.157	0.313 / 0.15	**0.324 / 0.16**

TABLE 6. Evaluation of various components of DGMA on ShapeNet using IoU.

DG	MA	3 views	5 views	8 views	12 views	16 views	20 views
✗	✗	0.7588	0.7680	0.7718	0.7731	0.7741	0.7742
✗	✓	0.7605	0.7703	0.7751	0.7776	0.7789	0.7795
✓	✗	0.7609	0.7713	0.7765	0.7789	0.7802	0.7807
✓	✓	**0.7632**	**0.7734**	**0.7781**	**0.7800**	**0.7810**	**0.7812**

decoder [1], while the refiner is based on the U-Net architecture, as proposed in [3]. This decision enables a clearer analysis of the role of each DGMA component in the 3D reconstruction performance of the proposed network architecture, depicted in Figure 2. Initially, a baseline, without the addition of the proposed DG and MA components,

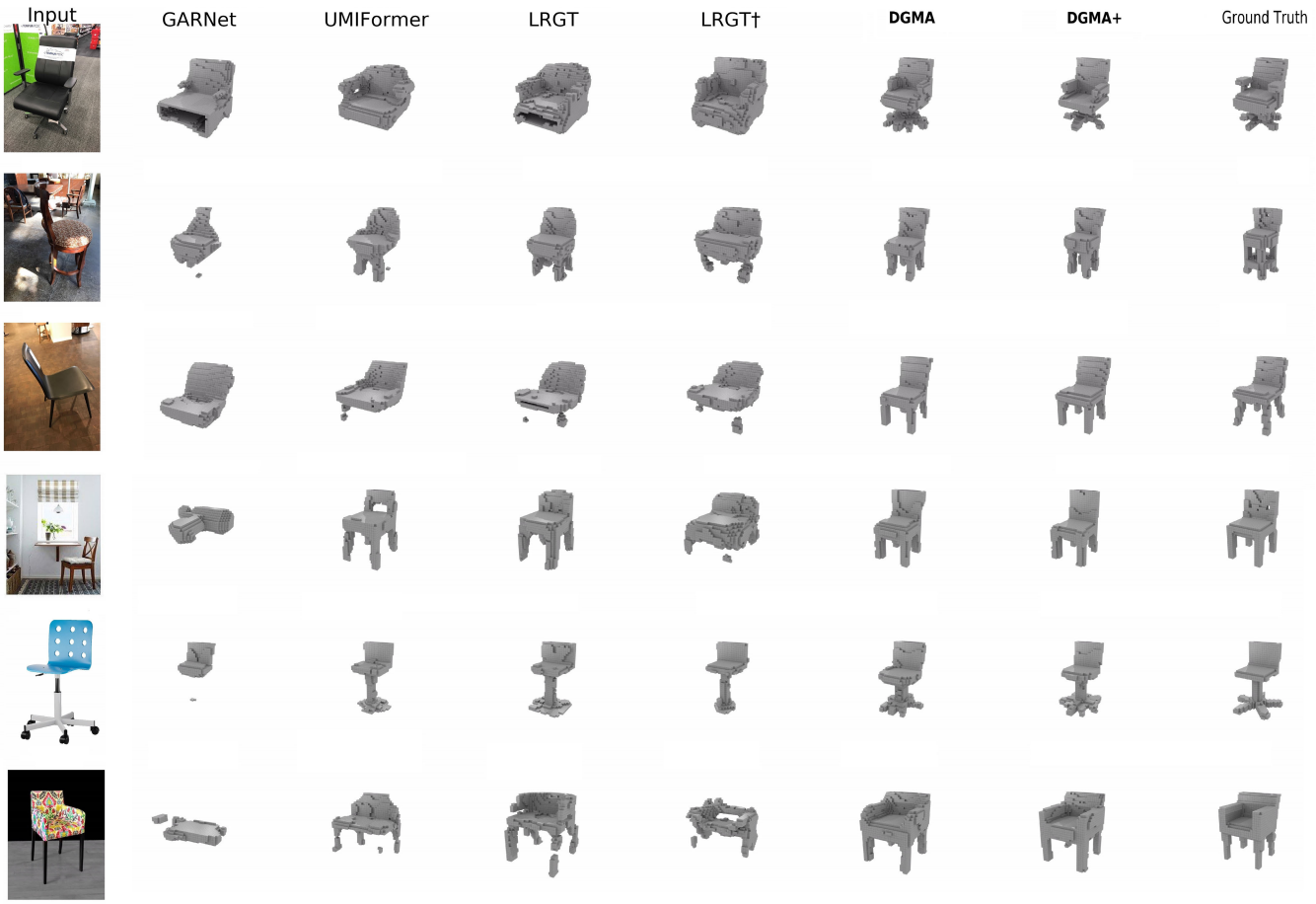

FIGURE 5. Single-view 3D reconstruction results on Pix3D dataset.

TABLE 7. Experimentation with the covariance size S on ShapeNet using IoU.

Covariance size S	3 views	5 views	8 views	12 views	16 views	20 views
8	0.7588	0.7695	0.7747	0.7778	0.7796	0.7801
12	**0.7632**	**0.7734**	**0.7781**	**0.7800**	**0.7810**	**0.7812**
16	0.7626	0.7728	0.7776	0.7793	0.7803	0.7806

is evaluated, leading to a performance of 0.7588 IoU for 3 views and 0.7742 IoU for 20 views. The introduction of the enhanced MA mechanism, which leverages both Euclidean and SPD manifolds to compute highly accurate attention maps, leads to an improvement in 3D reconstruction quality, achieving an IoU of 0.7605 for 3 views and 0.7795 for 20 views. On the other hand, the addition of the novel DG strategy enhances the 3D reconstruction accuracy (i.e., IoU of 0.7609 for 3 views and 0.7807 for 20 views) by intelligently grouping image tokens based on their attention values, facilitating better feature disentanglement and simplification of the reconstruction process. Finally, the use of both DG and MA components further boosts the 3D reconstruction performance of the proposed method, leading to best results for varying number of views, ranging from 3 to 20.

Table 7 presents IoU scores for the covariance size S taking the values of 8, 12, and 16. Notably, a size of 8 results in a distinct performance decline, indicating a loss of essential information during the dimensionality reduction. On the other hand, a size of 16 leads to a slight performance deterioration due to information redundancy. As a result, the size of 12 was chosen for striking the best balance between information retention, reconstruction accuracy and computational efficiency.

V. CONCLUSION

This work proposes a Transformer-based 3D object reconstruction method that relies on a novel dynamic token grouping strategy that enables the network to merge tokens across different views on-the-fly based on their attention values. In addition, an enhanced multi-manifold attention mechanism is utilized, allowing the computation of a more descriptive attention map to better support the proposed token grouping strategy. The experimental results verify the ability of the proposed method to achieve SoTA results in both multi-view and single-view 3D reconstruction tasks.

REFERENCES

[1] L. Yang, Z. Zhu, X. L. J. Nong, and Y. Liang, "Long-range grouping transformer for multi-view 3D reconstruction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 18211–18221.

[2] Z. Zhu, L. Yang, N. Li, C. Jiang, and Y. Liang, "UMIFormer: Mining the correlations between similar tokens for multi-view 3D reconstruction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 18180–18189.

[3] Z. Zhu, L. Yang, X. Lin, L. Yang, and Y. Liang, "GARNet: Global-aware multi-view 3D reconstruction network and the cost-performance tradeoff," *Pattern Recognit.*, vol. 142, Oct. 2023, Art. no. 109674.

[4] F. Yagubbayli, Y. Wang, A. Tonioni, and F. Tombari, "LegoFormer: Transformers for block-by-block multi-view 3D reconstruction," 2021, *arXiv:2106.12102*.

[5] Z. Shi, Z. Meng, Y. Xing, Y. Ma, and R. Wattenhofer, "3D-RETR: End-to-end single and multi-view 3D reconstruction with transformers," in *Proc. BMVC*, 2021, pp. 1–14.

[6] D. Wang, X. Cui, X. Chen, Z. Zou, T. Shi, S. Salcudean, Z. J. Wang, and R. Ward, "Multi-view 3D reconstruction with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5702–5711.

[7] B. Yang, S. Wang, A. Markham, and N. Trigoni, "Robust attentional aggregation of deep feature sets for multi-view 3D reconstruction," *Int. J. Comput. Vis.*, vol. 128, no. 1, pp. 53–73, Jan. 2020.

[8] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese, "3D-R2N2: A unified approach for single and multi-view 3D object reconstruction," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 628–644.

[9] H. Xie, H. Yao, S. Zhang, S. Zhou, and W. Sun, "Pix2Vox++: Multi-scale context-aware 3D object reconstruction from single and multiple images," *Int. J. Comput. Vis.*, vol. 128, no. 12, pp. 2919–2935, Dec. 2020.

[10] H. Xie, H. Yao, X. Sun, S. Zhou, and S. Zhang, "Pix2Vox: Context-aware 3D reconstruction from single and multi-view images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2690–2698.

[11] L. C. O. Tiong, D. Sigmund, and A. B. J. Teoh, "3D-C2FT: Coarse-to-fine transformer for multi-view 3D reconstruction," in *Proc. Asian Conf. Comput. Vis.*, 2022, pp. 1438–1454.

[12] X. Zhu, X. Yao, J. Zhang, M. Zhu, L. You, X. Yang, J. Zhang, H. Zhao, and D. Zeng, "TMSDNet: Transformer with multi-scale dense network for single and multi-view 3D reconstruction," *Comput. Animation Virtual Worlds*, vol. 35, no. 1, Jan. 2024, Art. no. e2201.

[13] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, "ShapeNet: An information-rich 3D model repository," 2015, *arXiv:1512.03012*.

[14] X. Sun, J. Wu, X. Zhang, Z. Zhang, C. Zhang, T. Xue, J. B. Tenenbaum, and W. T. Freeman, "Pix3D: Dataset and methods for single-image 3D shape modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2974–2983.

[15] M. Tatarchenko, S. R. Richter, R. Ranftl, Z. Li, V. Koltun, and T. Brox, "What do single-view 3D reconstruction networks learn?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3400–3409.

[16] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3D surface construction algorithm," *ACM SIGGRAPH Comput. Graph.*, vol. 21, no. 4, pp. 163–169, Aug. 1987.

[17] H. Su, C. R. Qi, Y. Li, and L. J. Guibas, "Render for CNN: Viewpoint estimation in images using CNNs trained with rendered 3D model views," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2686–2694.

[18] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "SUN database: Large-scale scene recognition from abbey to zoo," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3485–3492.

[19] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

[20] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[21] D. Konstantinidis, I. Papastratis, K. Dimitropoulos, and P. Daras, "Multi-manifold attention for vision transformers," *IEEE Access*, vol. 11, pp. 123433–123444, 2023.

[22] X.-F. Han, H. Laga, and M. Bennamoun, "Image-based 3D object reconstruction: State-of-the-art and trends in the deep learning era," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1578–1604, May 2021.

[23] J. Cheng, C. Leng, J. Wu, H. Cui, and H. Lu, "Fast and accurate image matching with cascade hashing for 3D reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1–8.

[24] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3D shape recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 945–953.

[25] O. Özyeşil, V. Voroninski, R. Basri, and A. Singer, "A survey of structure from motion," *Acta Numerica*, vol. 26, pp. 305–364, May 2017.

[26] L. Gao, Y. Zhao, J. Han, and H. Liu, "Research on multi-view 3D reconstruction technology based on SFM," *Sensors*, vol. 22, no. 12, p. 4366, Jun. 2022.

[27] A. Kar, C. Häne, and J. Malik, "Learning a multi-view stereo machine," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–12.

[28] M. Wang, L. Wang, and Y. Fang, "3DensiNet: A robust neural network architecture towards 3D volumetric object prediction from 2D image," in *Proc. 25th ACM Int. Conf. Multimedia*, vol. 15, Oct. 2017, pp. 961–969.

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, and S. Gelly, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–21.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

[31] L. Jiang, S. Shi, X. Qi, and J. Jia, "GAL: Geometric adversarial loss for single-view 3D-object reconstruction," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 802–816.

[32] B. Yang, S. Rosa, A. Markham, N. Trigoni, and H. Wen, "Dense 3D object reconstruction from a single depth view," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 12, pp. 2820–2834, Dec. 2019.

[33] P. Mandikal, K. L. Navaneet, M. Agarwal, and B. R. Venkatesh, "3D-LMNet: Latent embedding matching for accurate and diverse 3D point cloud reconstruction from a single image," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2018, pp. 1–12.

[34] A. A. Soltani, H. Huang, J. Wu, T. D. Kulkarni, and J. B. Tenenbaum, "Synthesizing 3D shapes via modeling multi-view depth maps and silhouettes with deep generative networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2511–2519.

[35] X. Hu, F. Zhu, L. Liu, J. Xie, J. Tang, N. Wang, F. Shen, and L. Shao, "Structure-aware 3D shape synthesis from single-view images," in *Proc. BMVC*, 2018, p. 230.

[36] Y. Zeng, J. Fu, and H. Chao, "Learning joint spatial-temporal transformations for video inpainting," in *Proc. Eur. Conf. Comput. Vis.*, Glasgow, U.K. Cham, Switzerland: Springer, 2020, pp. 528–543.

[37] R. Liu, H. Deng, Y. Huang, X. Shi, L. Lu, W. Sun, X. Wang, J. Dai, and H. Li, "Decoupled spatial-temporal transformer for video inpainting," 2021, *arXiv:2104.06637*.

[38] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (DV)*, Oct. 2016, pp. 565–571.

GEORGIOS KALITSIOS received the Diploma degree in electrical and computer engineering and the M.Sc. degree in artificial intelligence from the Aristotle University of Thessaloniki, Greece, in 2020 and 2022, respectively. Since October 2020, he has been a Research Associate with the Information Technologies Institute (ITI), Centre for Research and Technology Hellas (CERTH). His research interests include computer vision, machine/deep learning, image processing, and artificial intelligence.

DIMITRIOS KONSTANTINIDIS received the B.S. degree in electrical and electronic engineering from AUTH, in 2009, the Advanced M.S. degree in artificial intelligence from KU Leuven, in 2012, and the Ph.D. degree from the Imperial College of London with the topic of monitoring urban changes from satellite images, in 2017. He is currently with the Information Technologies Institute (ITI), Centre for Research and Technology Hellas (CERTH), as a Postdoctoral Researcher. His research interests include computer vision, image processing, machine learning, and artificial intelligence.

PETROS DARAS (Senior Member, IEEE) received the B.S. degree in electrical and computer engineering, the M.S. degree in medical informatics, and the Ph.D. degree in electrical and computer engineering from the Aristotle University of Thessaloniki, Greece, in 1999, 2002, and 2005, respectively. He is currently the Research Director (Grade A') of the Visual Computing Laboratory, Information Technologies Institute, CERTH. His main research interests include 3D object recognition, indexing, classification and retrieval, real-time 3D reconstruction of dynamic scenes, compression and coding of 3D meshes, bioinformatics, and medical image processing. His involvement with those research areas has led to the co-authoring of 70 papers in refereed journals (of which 33 in IEEE journals), 194 in international conferences, 35 book chapters, and four books. He has been involved in more than 59 projects, funded by the EC and the Greek Ministry of Research and Technology, having attracted as Researcher (July 2006–May 2020): 21.487.553,18 EUR. He is a member of the Technical Chamber of Greece. He served as the Chair for the IEEE Interest Group on Image, Video and Mesh Coding (2012–2014) and has been a Key Member of the IEEE Interest Group on 3D Rendering, Processing and Communications of IEEE Multimedia, since 2010. He regularly acts as a Reviewer/Evaluator of European Commission. He served as a Coordinator for the Future Media and 3D Internet—Task Force, the ''User Centric Media'' Cluster, and a member of the ''3D Media'' Cluster, the Future Internet Architectures (FIArch) Group and the Future Internet Assembly of the Networked Media Unit of the EC.

KOSMAS DIMITROPOULOS received the Diploma degree in electrical and computer engineering and the Ph.D. degree in applied informatics. He is currently a Principal Researcher (Grade B') with the Information Technologies Institute (ITI), Centre for Research and Technology Hellas (CERTH), and the Academic Director of the AIMOVE Post-Master Programme for CERTH (AIMove—''Artificial Intelligence and Movement for Robotics and Interactive Systems'') with MINES ParisTech/PSL University. His main research interests include multi-dimensional data modeling and analysis, artificial intelligence, and human–computer interaction. He has been involved in several European and national research projects and has served as a regular reviewer for a number of international journals and conferences.