2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Scene-level Point Cloud Colorization with Semantics-and-geometry-aware Networks

Rongrong Gao[1], Tian-Zhu Xiang[2], Chenyang Lei[1], Jaesik Park[3] and Qifeng Chen[1]

***Abstract*— In robotic applications, we often obtain tons of 3D point cloud data without color information, and it is difficult to visualize point clouds in a meaningful and colorful way. Can we colorize 3D point clouds for better visualization? Existing deep learning-based colorization methods usually only take simple 3D objects as input, and their performance for complex scenes with multiple objects is limited. To this end, this paper proposes a novel semantics-and-geometry-aware colorization network, termed SGNet, for vivid scene-level point cloud colorization. Specifically, we propose a novel pipeline that explores geometric and semantic cues from point clouds containing only coordinates for color prediction. We also design two novel losses, including a colorfulness metric loss and a pairwise consistency loss, to constrain model training for genuine colorization. To the best of our knowledge, our work is the first to generate realistic colors for point clouds of large-scale indoor scenes. Extensive experiments on the widely used ScanNet benchmarks demonstrate that the proposed method achieves state-of-the-art performance on point cloud colorization.**

I. Introduction

3D sensors (*e.g.*, depth sensor, time-of-flight sensor, and LiDAR) are capable of perceiving fine 3D geometric information of the scene but unable to capture appearance details (*e.g.*, color and texture) of the surroundings, compared with image sensors. In lots of robotic applications, only 3D sensors are utilized without any color information, which makes 3D data visualization challenging. Therefore, it is desirable to visualize with vivid color because colorized 3D data is perceptually more meaningful and credible, which often conveys rich semantics clues, thus not only providing better scene understanding to human beings but also significant improvements for visual recognition [1], [2] in modern AR/VR and robotic applications. As shown in Fig. 1, compared with the original point cloud with coordinates only, with the support of color information, the colorized point cloud makes the scene easier to understand visually, greatly improving the recognizability of objects. Therefore, point cloud colorization is an emerging topic for better 3D data visualization and visual perception.

Colorizing monochromatic images or videos has been studied extensively and achieved significant progress, which is widely applied in legacy photos or video restoration [3], image compression [4], video surveillance [5] and 3D modeling [6]. However, research on point cloud colorization remains limited, mainly due to its irregular and disordered data structure, sparse geometric information, and no gray-scale hints or texture for appearance details, especially for complex indoor scenes with multiple kinds of objects.

(a) Input. (b) Our colorized. (c) Ground truth.

Fig. 1: Illustration of point cloud colorization. (a) The input 3D point cloud with coordinates only. (b) Our colorized point cloud with predicted color information. (c) The ground-truth point cloud with real color information.

Different from image data, the aforementioned unique characteristics of point clouds make it difficult to explore semantics from point clouds [7]. While as pointed out by prior work [2], [8], the colorization task inherently requires a semantic understanding of the data, *e.g.*, understanding what type of an object is colorized. As a result, colorization solutions often highlight the importance of semantics. What is worse, the high-performance training methods for images do not work well on 3D data [9]. To our knowledge, point cloud colorization is a challenging and under-explored topic in computer vision.

To obtain convincing colorization, researchers have recently regarded point cloud colorization as a conditional generative task, and some generative adversarial networks (GANs)-based colorization methods are proposed to produce bright point cloud colors, including PCNN [1], DensePoint [10], Point2color [11], and HyperColor [12]. These methods often employ a generator to produce realistic fake colors to fool the discriminator, which differentiates between the real data and the generated color. However, these methods still suffer from limitations. Firstly, they often only support the simple 3D object as input, and thus they would be cast into the shade when it comes to complex scenarios (*e.g.*, multiple objects, and cluttered background). Secondly, colored point clouds often exhibit unsatisfactory artifacts, incoherent colors, or colorizations that homogeneously color the entire scene regardless of differences between objects.

To tackle the above challenges, in this paper, we propose a point cloud colorization method, *i.e.*, semantics-and-geometry-aware colorization network (SGNet), for plausible scene-level point cloud colorization. Specifically, we propose

[1] Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong SAR, China.
[2] Inception Institute of Artificial Intelligence, Abu Dhabi, UAE.
[3] POSTECH.

a new colorization pipeline that excavates the geometric and semantic cues from coordinate-only point clouds to generate credible coloring results. The proposed method can colorize the point cloud based on the semantics or types of objects or regions. We also develop multiple constraints, including colorfulness metric and pairwise consistency, to promote the model to produce realistic and colorful color predictions for point clouds. To our knowledge, this is the first work to colorize point clouds with outstanding performance for large-scale indoor scenes. In a nutshell, the main contributions can be summarized as follows.

- We propose a novel colorization method, the semantics-and-geometry-aware colorization network (SGNet), for scene-level point cloud colorization, which takes the point cloud with only coordinates only as input for plausible color generation using a sparse fully convolutional network.
- To facilitate the color learning process, we also present two novel losses into point cloud colorization, including a colorfulness metric loss that enforces the model to produce visually vibrant colors and a pairwise consistency loss to constrain the uniformity between the ground truth and predictions, which greatly boost the performance.
- We validate the proposed method on the widely-used ScanNet indoor complex scene dataset, and extensive experiments demonstrate the effectiveness of the proposed model with superior performance to the previous state-of-the-art point cloud approaches quantitatively and qualitatively.

II. Related Work

Although Image colorization has achieved great progress, it is not easy to directly apply image colorization methods to 3D data due to big data differences. 3D data colorization often contains colorization of depth maps [13], [14], voxels [9], point clouds [1] and meshes [15]. In this paper, we focus on point cloud colorization. Compared with other 3D data, point cloud data is irregular and disordered, making colorization relatively more challenging.

Traditionally, texture mapping is used to associate point clouds with color [16], [17]. However, these methods often require predefined texture patterns or well-registered color images. They thus are not applicable for generating new colorization only from the point cloud data with no additional inputs. Given the success of generative models in image colorization, recent works apply generative models to colorize point clouds. [1] proposes a point cloud colorization network (PCCN) based on conditional generative adversarial network (cGAN) [18]. While PointNet [19] is integrated into a generator to predict the color and into a discriminator to judge whether the color is artificial or real. Similarly, Cao *et al.* [10] introduces a DensePoint dataset for point cloud colorization and presents a colorization method based on cGAN and PointNet. However, it trains each category of objects in separate networks. To support multiple style colorization, recently, Kostiuk *et al.* [12] presented the HyperColor for synthesizing auto-colored 3D models. The method adopts two improved autoencoder models to generate 3D point clouds of objects and colors for each point. However, colorizing the point cloud with consistent color and clear borderlines is difficult. In [11], Shinohara *et al.* propose a point2color model for airborne point cloud colorization, which uses the cGAN model to estimate the color of each point and the differentiable rendering to assist the colorization ability. However, this model tends to ignore small objects when colorized. We observe that point cloud colorization is in its infancy and remains an under-explored problem. This paper provides a first step toward objectively understanding why some point clouds are perceived as real-world models.

III. Method

This section will describe the proposed method for point cloud colorization in detail.

A. Semantics-and-geometry-aware Network

Our proposed model is based on the probabilistic generative model [20]. The overall architecture of the proposed method is illustrated in Fig. 2.

Intuitively, we argue that point cloud colorization potentially implies a requirement for scene understanding of point clouds. We observe, the color information can be well inferred from scene semantic information (*i.e.*, recognized objects and regions) and scene geometric information (*i.e.*, local spatial cues). Thus, we first utilize two backbone networks for the generator to learn discriminating point-wise features for describing the raw point clouds from both geometric and semantic perspectives as a pre-step. The point cloud $I \in \mathcal{R}^{N\times 3}$ (where N is the number of point clouds) only with coordinate information is directly fed into the designed generator, which exploits the 3D sparse and fully-convolutional network to explore the geometric cues and semantic cues of the point cloud to guide the color generation in the Lab color spaces. Specifically, we adopt a fully convolutional geometric feature model (FCGF), introduced in recent work [21] on 3D geometric matching, to extract such sparse geometric features $f_g \in \mathcal{R}^{N\times 16}$. The FCGF is a fully-convolutional network with sparse convolutions, which has bigger receptive fields and thus can capture broad contextual information to produce discriminating features that summarize geometric context. Besides, we adopt the MinkowskiNet [22], which is a 4D convolutional neural network with generalized sparse convolutions for high-dimensional semantic perception, to learn semantic feature representation ($f_s \in \mathcal{R}^{N\times 20}$). It can be formulated as

$$f_s = \mathcal{F}_{mk}(I), \tag{1}$$

$$f_g = \mathcal{F}_{fc}(I), \tag{2}$$

where $\mathcal{F}_{mk}$ denotes the MinkowskiNet for semantic exploration, and the $\mathcal{F}_{fc}$ indicates the FCGF model for geometric exploration. The geometric cues provide local structure information to facilitate the preservation of fine structures in colorization. The semantic cues can benefit object/region

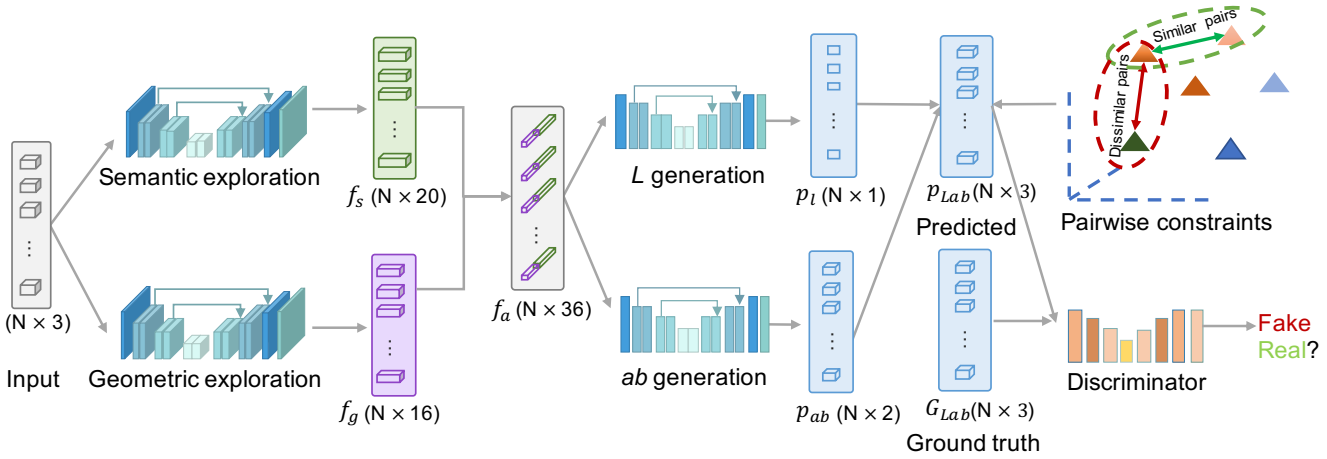


Fig. 2: The architecture of the proposed semantics-and-geometry-aware network for point cloud colorization.

recognition, thereby boosting the differential colorization for different types of objects/regions.

After that, the learned geometric features and the learned semantic features of each point are integrated by concatenation to obtain an enhanced discriminating feature vector ($f_a \in \mathcal{R}^{N\times 36}$) for the following color semantic exploration in Lab color space. Unlike the RGB color model, LAB is designed to approximate human vision and thus can well represent color information in the physical world [23]. Considering that lightness (L) and chromaticity (a and b) are independent in Lab color space, we exploit two branches to learn lightness representation and chromaticity information, respectively. Specifically, the aggregated feature f_a is fed into two UNet-like structures with skip connections and residual blocks, *i.e.*, L generation and *ab* generation, respectively to mine valuable color cues and produce the predictions of Lab. One is used to predict L channel value $p_l \in \mathcal{R}^{N\times 1}$ for each point and the other is for a and b channels $p_{ab} \in \mathcal{R}^{N\times 2}$, which can be defined as

$$f_a = \mathcal{F}_{cat}(f_s, f_g), \tag{3}$$
$$p_l = \mathcal{F}_l(f_a), \tag{4}$$
$$p_{ab} = \mathcal{F}_{ab}(f_a), \tag{5}$$

where $\mathcal{F}_{cat}$ is concatenation operation, $\mathcal{F}_l$ is L generation and $\mathcal{F}_{ab}$ is *ab* generation. Then, we merge these two predictions to get the final colorized point cloud P. And more qualitative experiments with different branching strategies is shown in supplementary material.

B. Loss Functions

To boost colorization learning, we incorporate two novel losses for point cloud colorization: the colorfulness metric loss and the pairwise consistency loss. The former forces the model to learn authentic colors with a broad color distribution for raw point clouds, and the latter constrains model predictions to be as consistent as possible with ground truth. Together with commonly used smoothed L_1 loss and adversarial loss, the proposed model is trained to produce vibrant and realistic colors for point clouds.

1) Colorfulness metric loss (ℓ_{cm}): To evaluate the colorfulness of our colorized point cloud, we can compare the color distribution of our colorized point cloud and the ground-truth point cloud. Therefore, the colorization quality can be evaluated by the difference of color histograms, *i.e.*, colorfulness metric [24], which is defined as

$$\ell_{cm} = \left(6dist_L + dist_a + dist_b\right)/8, \tag{6}$$

where $dist_L$, $dist_a$ and $dist_b$ are point-wise KL-divergence of the histograms of the L, a, and b channels of the point cloud with predicted color and the ground truth. The proposed loss (ℓ_{cm}) is to minimize the difference in color distribution between the prediction and the ground truth. We compute the color histogram on the luminance channel, which has been shown to better correlate with human perception of color, and the weight of $dist_L$ is larger than the other two channel terms because human eyes can capture subtler color changes of L channels.

2) Pairwise consistency loss (ℓ_{con}): Meanwhile, we hope to keep the inherent smoothness characteristics of predicted colors remaining the same. Given the input point clouds, we design a pairwise consistency loss to measure the similarity of their outputs to achieve random point pairwise consistency, as shown in Fig. 3.

Specifically, we assume that the pair of points that share similar colors should maintain similar colors for the predicted point clouds and vice versa. Thus, this consistency loss can be represented as

$$\ell_{con} = \sum_{i,j\in\mathbb{S},i\neq j} e^{-\frac{\|g_i - g_j\|}{\lambda}} \|p_i - p_j\|, \tag{7}$$

where $\|\cdot\|$ denotes the Euclidean distance, and $\mathbb{S}$ is the randomly selected subset from the input point cloud. i and j denote a random pair of points. p_i and g_i are the predicted color and the ground-truth color at point i. λ is the hyper-parameter which is trivially set 0.01 in the paper.

3) Smoothed L_1 loss (ℓ_s): For the point clouds collected with lidar or RGB-D camera, there are usually lots of noise and outliers points, and to make the network more robust to

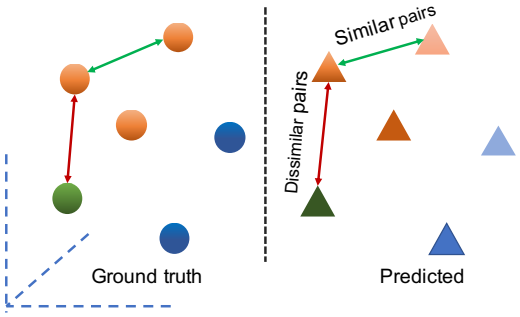


Fig. 3: Illustration of the pairwise consistency for point clouds. The clusters represent the ground-truth point clouds with real colors (left) and the point cloud with generated colors (right). The similar pair (cyan line) is the point pair sharing similar color and the dissimilar pair (red line) is one if not. Note that the pairwise consistency keeps the relationships between selected similar/dissimilar pairs the same for ground truth and generated point clouds.

outliers and avoid the gradient explosion during the training process, we introduce the smoothed L_1 loss as one of the loss function term. Suppose g_i is the ground-truth color for the i-th point and p_i is the predicted color of the i-th point, the smoothed L_1 loss is defined as

$$\ell_s = \sum_i \begin{cases} 0.5\|g_i - p_i\|^2/\delta & \text{if } \|g_i - p_i\| < \delta, \\ \|g_i - p_i\|/\delta - 0.5 & \text{otherwise,} \end{cases} \tag{8}$$

where δ is the threshold and i is the point index. The threshold helps to make the network training robust to outliers.

4) Total loss (ℓ_L): One observation of the colorization problem is that the ground truth of the indoor scene is only one of the possible colorization strategies. What we hope to synthesize is the most user-friendly colorization result. To enhance the ability to differentiate generated data from real data and improve the realism of the generated colorized point cloud, we can adopt an adversarial loss for the generator. Given a (well-trained) discriminator, the objective is to train the generator so that the discriminator believes the generated result is realistic. Thus the adversarial loss [25] is adopted as the generator loss ℓ_g:

$$\ell_g = \mathbb{E}_x[log(1 - D(x, G(x)))], \tag{9}$$

where x is the input $N \times 3$ tensor representing a point cloud in our case. $G(\cdot)$ and $D(\cdot)$ denote the generator and the discriminator respectively. Thus, the total loss for SGNet can be formulated as

$$\mathcal{L} = \ell_g + \alpha\ell_s + \beta\ell_{cm} + \sigma\ell_{con}, \tag{10}$$

where α, β, and σ are hyper-parameters used to control the strength of each loss term. We empirically set α = 5, β = 0.1, and σ = 0.001 in all experiments.

IV. Experiments

This section introduces the experimental setup, including datasets, evaluation metrics, comparison methods, and implementation details. Then, we provide the comparison experiments with other baselines and ablation studies.

A. Experimental Setup

1) Datasets: We use the widely-used ScanNet dataset [26], which consists of 1,513 scans covering 20 object categories from 707 unique indoor environments, such as beds, cabinets, tables, chairs, lamps, and windows. Each scan is an indoor RGB-D scene, represented as a point cloud. Each point contains x, y, z coordinates and r, g, b colors. The splits of training, validation, and testing sets are the same as the original setting. We train our model on the whole dataset from scratch.

2) Evaluation metrics.: We utilize three quantitative metrics to evaluate our method, including mean square error (MSE), fréchet point cloud distance (FPD), and colorfulness metric (CM). We also provide user study results. Among these metrics, MSE is used to measure the mean difference between the predicted color and the ground truth color of the point clouds. CM is a histogram-based metric that measures how realistic the resulting color distribution is close to the ground truth. Besides, an obvious observation is that realistic colorization results should maintain similar colors within the same semantic objects. Thus, inspired by Shu *et al.* [27], a simple extension of the fréchet inception distance for point clouds (*i.e.* FPD) is adopted to calculate the 2-Wasserstein distance between predicted and ground-truth Gaussian measures in feature spaces, defined as

$$\begin{aligned} \mathrm{FPD}(\mathbf{f(p)},\mathbf{f(g)}) = &\left\|\mathbf{m_{f(p)}} - \mathbf{m_{f(g)}}\right\|_2^2 + \\ &\mathrm{Tr}\left(\Sigma_{\mathbf{f(p)}} + \Sigma_{\mathbf{f(g)}} - 2\left(\Sigma_{\mathbf{f(p)}}\Sigma_{\mathbf{f(g)}}\right)^{\frac{1}{2}}\right), \end{aligned} \tag{11}$$

where $\mathbf{m_{f(p)}}$, $\Sigma_{\mathbf{f(p)}}$ and $\mathbf{m_{f(g)}}$, $\Sigma_{\mathbf{f(g)}}$ are the mean vector and covariance matrix of the point cloud calculated from prediction and ground truth, respectively. The smaller the FPD, the better the performance.

3) Baselines and Implementation details: To demonstrate the effectiveness of the proposed method, we choose regression-based colorization and DensPointNet [10] as baselines for comparison. The regression method is a straightforward solution that directly regresses the color vector from the input point cloud. DensePointNet also is a generative adversarial network based on a modified PointNet network. It does not fit our case because it is specially designed for small-scale object-level point clouds. So we first down-sample each scene point cloud into 16,392 points and then train the model on the whole dataset from scratch with the same settings based on the code released by the authors. Then for evaluation, we down-sampled all the results of comparison methods into 16,392 points with the random sampling method.

We use the Adam optimizer and set the learning rate to 5e-5 and 1e-4 for the generator and discriminator. The model is trained for 80 epochs with a batch size of 12 on NVIDIA 2080Ti GPUs. During the training stage, we update both generator and discriminator at each step.

TABLE I: Quantitative comparison with other methods for point cloud colorization using three evaluation metrics. "↓" or "↑" indicates that the method with the smaller or larger metrics is better than the others. The best results are highlighted in bold.

Method	MSE ↓	FPD ↓	CM ↑
Regression	0.084 ± 0.015	3.333 ± 1.676	0.064 ± 0.013
DensePointNet	0.081 ± 0.009	4.106 ± 1.273	0.080 ± 0.010
SGNet (ours)	$\mathbf{0.041 \pm 0.019}$	$\mathbf{2.116 \pm 1.438}$	$\mathbf{0.221 \pm 0.008}$

B. Experimental Results

1) Qualitative evaluation: Fig. 4 shows the qualitative comparisons of our proposed method with other comparison methods on a typical sample from the dataset. These results intuitively show the superior performance of the proposed method. Compared with other methods, the proposed method provides more vibrant and realistic colorization results, which may benefit from the proposed loss constraints. Moreover, different objects in the scene are differentiated and colored with clear object structure (*i.e.*, table, and monitor in the sample), which may benefit from the integration of semantic and geometric cues. Due to space limitations, more visual comparisons can be found in the *supp*.

2) Quantitative evaluation: Tab. I summarize the quantitative results of our method against other baselines. It is clear that our method significantly outperforms all other models under the three evaluation metrics. Compared to DensePointNet, the performance gains of MSE, FPD, and CM are 0.040 (↑49%), 2.01(↑49%), and 0.141 (↑176%), respectively, on the testing dataset. Our proposed method achieves state-of-the-art performance, providing a new strong baseline for the point cloud colorization community, which is expected to advance the field. Specifically, compared with the two baseline methods, the proposed model gets smaller MSE, FPD values, which indicates the fact that it generates more reasonable colors for different objects/regions. And the colorfulness metric of our algorithm is larger than the compared methods, which is consistent with the qualitative results in the former section.

3) The perceptual user study: To evaluate whether the result is user-friendly to the audience, we follow the user study protocol. The perceptual user study is the key experiment to evaluate the performance of different methods when the ground truth of not user-friendly performance. In our experiments, we randomly select 20 samples from 100 testing samples, then let raters who are professional researchers related to the field assess the quality of these colorizations with a three alternative-forced choice test to pick the best-of-three colorization. We display the re-colorized point clouds sequentially in random order based on three comparison methods. Then we report the mean fooling rate over 13 colorization and 20 different raters for each seed. Tab. II summarizes the results of our perceptual experiment. Our method is consistently rated preferable by most users. Obviously, with these three kinds of evaluation methods, we can see that our method outperforms all the other state-of-the-art methods.

TABLE II: Preference score in the user study for the colorization results of different methods. The best results are highlighted in bold.

Method	Regression	DensePointNet	Ours
Percentage	3.3%	30.4%	**66.3%**

TABLE III: Ablation study with different input feature settings, including only geometric feature (f_g), only semantic feature (f_s), and integration of both features ($f_g + f_s$). The best results are highlighted in bold.

Input features	MSE ↓	FPD ↓	CM ↑
f_g	0.079 ± 0.028	2.681 ± 1.227	0.069 ± 0.010
f_s	0.048 ± 0.023	$\mathbf{2.106 \pm 1.454}$	0.071 ± 0.010
$f_g + f_s$	$\mathbf{0.041 \pm 0.019}$	2.116 ± 1.438	$\mathbf{0.221 \pm 0.008}$

C. Ablation Study

In this section, we perform comprehensive ablation analyses to validate the effectiveness of each key component, including input feature adoption (*e.g.*, semantic feature, and geometric feature) and the selection of loss terms. More analysis (*e.g.*, parameter sensitivity analysis, and prediction branch selection) can be found in the supplementary material.

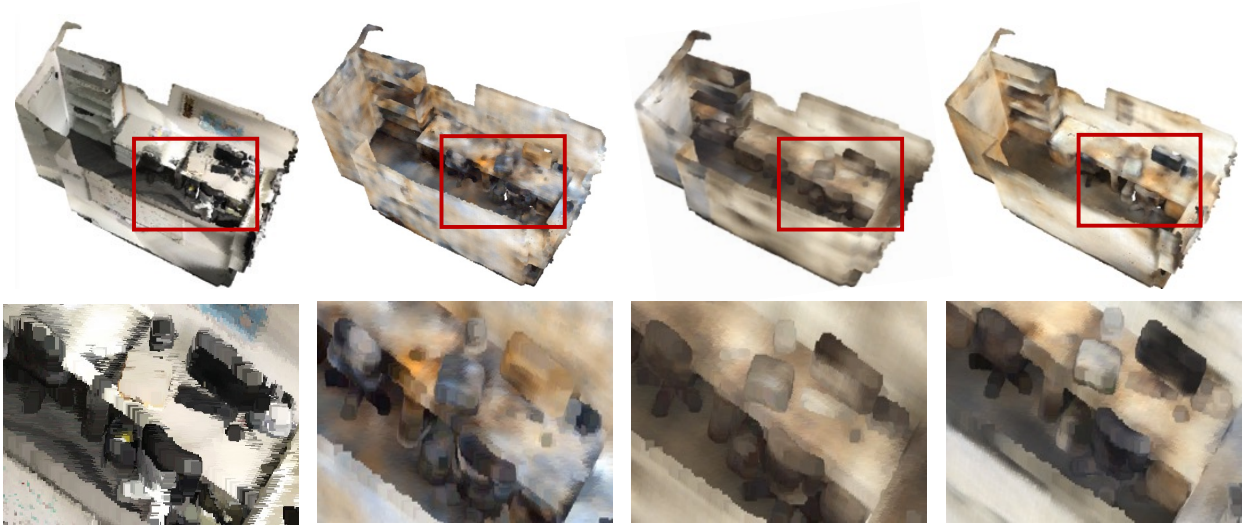
Effectiveness of input features (f_g&f_s). We test the effectiveness of different input features for point cloud colorization, as shown in Tab. III. Compared with the geometric feature input, the semantic feature exhibits relatively slightly better colorization performance, which may stem from mining semantic clues of the scene.

Combining the two can encode more valuable cues and achieve significant performance improvements, with gains of about 0.038, 0.565, and 0.152 compared to f_g in terms of MSE, FPD, and CM, 0.007 and 0.150 compared to f_s on MSE and CM respectively. Fig. 5 provides the visual results. It can be seen that the geometric version well preserves the object structure of the scene (*e.g.*, the chair on the right), and the semantic version considers the semantic differences of objects/regions for colorization (*e.g.*, the monitor and walls). The proposed method combines the benefits of both and provides differentiated coloring with a clear structure for different objects/regions.

Effectiveness of loss terms. Tab. IV shows the results of different settings of loss terms. Compared to baseline $\ell_g + \ell_s$, adding ℓ_{cm} forces the model to produce more colorful colors with a wider color distribution and thus obtains the best performance on CM. The ℓ_{con} is used to constrain the relative color relationships between points to be consistent with the

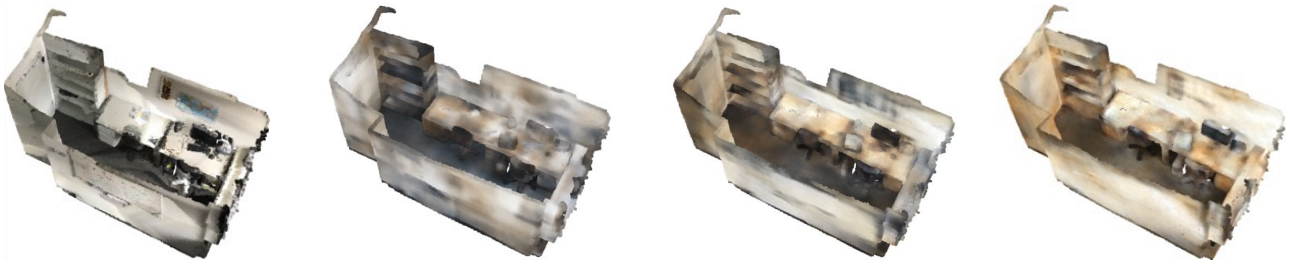
TABLE IV: Ablation study with different loss terms. The best results are highlighted in bold.

Method	MSE ↓	FPD ↓	CM ↑
$\ell_g + \ell_s$	0.046 ± 0.056	2.840 ± 1.051	0.190 ± 0.009
$\ell_g + \ell_s + \ell_{cm}$	0.069 ± 0.053	2.301 ± 1.334	$\mathbf{0.300 \pm 0.006}$
$\ell_g + \ell_s + \ell_{con}$	0.042 ± 0.020	2.285 ± 1.649	0.079 ± 0.010
$\ell_g + \ell_s + \ell_{cm} + \ell_{con}$	$\mathbf{0.041 \pm 0.019}$	$\mathbf{2.116 \pm 1.438}$	0.221 ± 0.008



(a) Ground truth (b) Regression (c) DensePointNet (d) Ours

Fig. 4: Visual comparisons of the proposed method with other baselines on the ScanNet dataset. Our method achieves superior colorization performance to other baselines. Specifically, compared with (b) regression-based method, the proposed model generates more consistent and plausible colors for different objects/regions. Compared with (c) DensePointNet, our model produces differentiated colors for different objects/regions in the scene, *e.g.*, table, and monitor (top). More visual results can be found in the *supp.*



(a) Ground truth (b) Geometric features (c) Semantic features (d) Geometric + Semantic

Fig. 5: Colorization results based on different input features. From the results, we can see that (b) geometric feature-based model preserves the clear object structures, *e.g.*, the chair on the right, and the (c) semantic feature-based model distinguishes the objects/regions and generates different colorization, *e.g.*, monitor and walls. The proposed model combines the benefits of both and produces appealing colorization results. Please zoom in for more details.

(a) Ground truth (b) $l_g + l_s$ (c)) $l_g + l_s + l_{cm}$ (d) $l_g + l_s + l_{con}$ (e) $l_g + l_s + l_{cm} + l_{con}$

Fig. 6: Colorization results based on different settings of loss terms. From left to right columns are results with different loss functions. It is clear that with the help of the two proposed loss terms, our loss function constrains the model to generate realistic colorization with more consistent and colorful colors. Please zoom in for more details.

ground truth, thus achieving good performance on MSE and FPD. Combining these losses, the proposed model achieves the best-balanced performance, with the best result on MSE and FPD and the second-best on CM. As shown in Fig. 6, the method of (c) introduces the ℓ_{cm} term, which improves the color distribution of the scene. In contrast, the method of

(d) combines the ℓ_{con} term, which colorizes the scene more consistently and smoothly, such as desks in the office scenes. As the consistency loss strengthens the inner smoothness of the same object, which contradicts the definition of the colorfulness metric, adding the ℓ_{con} term degrades the colorfulness of the final result. It can be concluded that a realistic colorization result should be a balance between the object semantics consistency and colorfulness.

V. Conclusion

This paper proposes a novel semantics-and-geometry-aware network (SGNet) for realistic scene-level point cloud colorization. Specifically, our model fully explores the geometric and semantic cues in a probabilistic generative adversarial framework for an automatic color generation. Furthermore, we design a colorfulness metric loss and a pairwise consistency loss, which enforce the model to generate consistent colors with a wide color distribution. Extensive experiments demonstrate superior performance over other state-of-the-art competitors. Our method produces vivid colorization with a clear structure for different objects and regions. To our knowledge, we are the first to generate realistic colors for point clouds of large-scale indoor scenes. We hope that our study will strongly boost growth in this community.

VI. Acknowledgement

Jaesik Park was supported by IITP grant funded by the Korea government(MSIT) (No.2019-0-01906, Artifcial Intelligence Graduate School Program(POSTECH) and No.2022-0-00290: Visual Intelligence for Space-Time Understanding and Generation based on Multi-layered Visual Common Sense).

References

[1] J. Liu, S. Dai, and X. Li, "Pccn: Point cloud colorization network," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 3716–3720.

[2] J. Zhao, J. Han, L. Shao, and C. G. Snoek, "Pixelated semantic colorization," *International Journal of Computer Vision*, pp. 1–17, 2019.

[3] C. Lei and Q. Chen, "Fully automatic video colorization with self-regularization and diversity," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3753–3761.

[4] Y. Xiao, J. Wu, J. Zhang, P. Zhou, Y. Zheng, C.-S. Leung, and L. Kavan, "Interactive deep colorization and its application for image compression," *IEEE Transactions on Visualization & Computer Graphics*, no. 1, pp. 1–1, 2020.

[5] G. Wu, Y. Zheng, Z. Guo, Z. Cai, X. Shi, X. Ding, Y. Huang, Y. Guo, and R. Shibasaki, "Learn to recover visible color for video surveillance in a day," in *European Conference on Computer Vision*. Springer, 2020, pp. 495–511.

[6] M. Waechter, N. Moehrle, and M. Goesele, "Let there be color! large-scale texturing of 3d reconstructions," in *European conference on computer vision*. Springer, 2014, pp. 836–850.

[7] Y. Guo, H. Wang, Q. Hu, H. Liu, L. Liu, and M. Bennamoun, "Deep learning for 3D point clouds: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 12, pp. 4338–4364, 2020.

[8] M. Kumar, D. Weissenborn, and N. Kalchbrenner, "Colorization transformer," *arXiv preprint arXiv:2102.04432*, 2021.

[9] Z. Yang, L. Liu, and Q. Huang, "Learning generative neural networks for 3d colorization," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018, pp. 2580–2587.

[10] X. Cao and K. Nagao, "Point cloud colorization based on densely annotated 3d shape dataset," in *International Conference on Multimedia Modeling*. Springer, 2019, pp. 436–446.

[11] T. Shinohara, H. Xiu, and M. Matsuoka, "Point2color: 3d point cloud colorization using a conditional generative network and differentiable rendering for airborne lidar," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1062–1071.

[12] I. Kostiuk, P. Stachura, S. K. Tadeja, T. Trzciński, and P. Spurek, "Hypercolor: A hypernetwork approach for synthesizing auto-colored 3d models for game scenes population," *arXiv preprint arXiv:2108.01411*, 2021.

[13] F. M. Carlucci, P. Russo, and B. Caputo, "DE^2co: Deep depth colorization," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2386–2393, 2018.

[14] C.-S. Lai, Z. You, C.-C. Huang, Y.-H. Tsai, and W.-C. Chiu, "Colorization of depth map via disentanglement," in *European Conference on Computer Vision*. Springer, 2020, pp. 450–466.

[15] G. Leifman and A. Tal, "Mesh colorization," *Computer Graphics Forum*, vol. 31, no. 2, pp. 421–430, 2012.

[16] J. F. Blinn and M. E. Newell, "Texture and reflection in computer generated images," *Communications of the ACM*, vol. 19, no. 10, pp. 542–547, 1976.

[17] B. Lévy, "Constrained texture mapping for polygonal meshes," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 2001, pp. 417–424.

[18] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[19] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[20] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[21] C. Choy, J. Park, and V. Koltun, "Fully convolutional geometric features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8958–8966.

[22] C. Choy, J. Gwak, and S. Savarese, "4d spatio-temporal convnets: Minkowski convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3075–3084.

[23] R. C. Gonzalez, *Digital image processing*. Pearson education india, 2009.

[24] I. Viola, S. Subramanyam, and P. Cesar, "A color-based objective quality metric for point cloud contents," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2020, pp. 1–6.

[25] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, pp. 1–9, 2014.

[26] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5828–5839.

[27] D. W. Shu, S. W. Park, and J. Kwon, "3d point cloud generative adversarial network based on tree structured graph convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3859–3868.