

الجمهورية الشعبية الديمقراطية الجزائرية  
People's Democratic Republic of Algeria  
وزارة التعليم العالي و البحث العلمي  
Ministry of Higher Education and Scientific Research  
المدرسة العليا للإعلام الآلي 8 ماي 5491 - سيدي بلعباس  
Higher School of Computer Science  
8 Mai 1945 - Sidi Bel Abbas



## Master's Thesis

To obtain the diploma of Master's Degree

Field of Study: **Computer Science**

Specialization: [Information system and web development (SIW)]

## Theme

---

**Text-to-3D Shape Generation Using Deep Learning**

---

Presented by  
**Asma Rahali**  
**Souha Hocini**

Defended on: [june, 2026]  
*In front of the jury composed of*

Ms. KLOUCHE Badia  
Pr. BENSLIMANE Sidi Mohamed  
Ms. ELHANNANI Souad  
Mr. KECHAR Mohamed  
Ms. ABDELHAK Soumia

President of the Jury  
Thesis Supervisor  
Examiner  
Incubator Representative  
Guest

*Academic Year: 2025/2026*

# Abstract

The creation of 3D content is central to applications such as gaming, animation, robotics simulation, and virtual and augmented reality, yet it traditionally requires considerable manual effort from skilled designers. Text-to-3D generation has recently emerged as a promising way to automate this process by leveraging large-scale vision–language models and 2D diffusion priors. This thesis is organised in two parts. The **first part** establishes the necessary background: it reviews the main 3D shape representations—explicit, implicit, and hybrid—and the principal families of deep generative models (GANs, VAEs, and diffusion models), with particular attention to Score Distillation Sampling, the CLIP vision–language model, and the prompt modalities (text, sketches, and 2D layouts) used to control generation. The **second part** surveys the state of the art through an in-depth analysis of representative methods—notably DreamStone, which uses image modality as a stepping stone to overcome the scarcity of paired text–shape data, and T-Dreamer, which injects pose-specific geometric priors to address the multi-view “Janus” problem—examining for each the datasets, methodology, and trade-offs. Together, the two parts clarify the dominant paradigms and the open challenges in fidelity, controllability, and efficiency that motivate future research.

**Keywords**— Text-to-3D generation, 3D content creation, diffusion models, Score Distillation Sampling, vision–language models, CLIP, 3D shape representation, controllable generation.

# Résumé

La création de contenu 3D est essentielle pour des applications telles que les jeux vidéo, l’animation, la simulation robotique ainsi que la réalité virtuelle et augmentée, mais elle exige traditionnellement un effort manuel considérable de concepteurs qualifiés. La génération *text-to-3D* est récemment apparue comme un moyen prometteur d’automatiser ce processus grâce aux grands modèles vision–langage et aux a priori de diffusion 2D. Ce mémoire est organisé en deux parties. La **première partie** pose les fondements nécessaires : elle passe en revue les principales représentations de formes 3D — explicites, implicites et hybrides — et les grandes familles de modèles génératifs profonds (GAN, VAE et modèles de diffusion), en accordant une attention particulière au *Score Distillation Sampling*, au modèle vision–langage CLIP, ainsi qu’aux différentes modalités de *prompt* (texte, croquis et dispositions 2D) utilisées pour contrôler la génération. La **seconde partie** dresse un état de l’art à travers une analyse approfondie de méthodes représentatives — notamment DreamStone, qui utilise la modalité image comme tremplin pour pallier la rareté des données texte–forme appariées, et T-Dreamer, qui injecte des a priori géométriques spécifiques à la pose pour traiter le problème multi-vues « de Janus » — en examinant pour chacune les données, la méthodologie et les compromis. Ensemble, ces deux parties mettent en évidence les paradigmes dominants et les défis ouverts en fidélité, contrôlabilité et efficacité qui motivent les recherches futures.

**Mots-clés** — Génération text-to-3D, création de contenu 3D, modèles de diffusion, Score Distillation Sampling, modèles vision–langage, CLIP, représentation de formes 3D, génération contrôlable.

## ملخص

يُعدّ إنشاء المحتوى ثلاثي الأبعاد أمراً محورياً في تطبيقات مثل الألعاب والرسوم المتحركة ومحاكاة الروبوتات والواقع الافتراضي والمعزز، إلا أنه يتطلب تقليدياً جهداً يدوياً كبيراً من مصممين مهرة. وقد ظهر مؤخراً توليد المحتوى ثلاثي الأبعاد من النص كوسيلة واعدة لأتمتة هذه العملية بالاعتماد على نماذج الرؤية واللغة ومُسبقات الانتشار ثنائية الأبعاد. تنقسم هذه المذكرة إلى جزأين. يرسي الجزء الأول الأسس اللازمة، إذ يستعرض أبرز تمثيلات الأشكال ثلاثية الأبعاد --- الصريحة والضمنية والهجينة --- والعائلات الرئيسية للنماذج التوليدية العميقة (الشبكات التوليدية التخاصمية، والمُرمّزات الذاتية التغيرية، ونماذج الانتشار)، مع إيلاء اهتمام خاص لتقنية تقطير الدرجات لأخذ العينات، ونموذج «كليب»، وأنماط التوجيه المستخدمة للتحكم في التوليد. أما الجزء الثاني فيقدّم عرضاً لأحدث ما توصّل إليه المجال من خلال تحليل معمّق لطرائق تمثيلية --- لا سيما «دريم ستون» و«تي-دريمر» --- مع دراسة مجموعات البيانات والمنهجية والمفاضلات لكل منهما. ويبرز الجزآن معاً النماذج المهيمنة في هذا المجال والتحديات المفتوحة في الدقة والقابلية للتحكم والكفاءة التي تحفز الأبحاث المستقبلية.

الكلمات المفتاحية: توليد المحتوى ثلاثي الأبعاد من النص، إنشاء المحتوى ثلاثي الأبعاد، نماذج الانتشار، تقطير الدرجات لأخذ العينات، نماذج الرؤية واللغة، نموذج كليب، تمثيل الأشكال ثلاثية الأبعاد، التوليد القابل للتحكم.