école doctorale **sciences pour l'ingénieur et microtechniques**

Titre de la thèse : Intégration de connaissances pour le contrôle du modèle Stable Diffusion dans l'espace latent : application à la génération de données visuelles pour l'entraînement de modèles profonds en vision par ordinateur.

Laboratoire d'accueil : Connaissance et Intelligence Artificielle Distribuées (CIAD) – http://www.ciad-lab.fr

Spécialité du doctorat préparé : Informatique

Mots-clefs : IA générative, Intégration de systèmes à base de connaissances, Synthèse de scènes de conduite réaliste, Mécanismes d'attention, Contrôle dans l'espace latent.

Descriptif détaillé de la thèse :

Contexte :

Le modèle Stable Diffusion a connu un succès remarquable dans la génération d'images de haute qualité dans divers domaines grâce à son efficacité et son accessibilité en open source [1]. Cependant, il présente des limitations dans la génération de scènes de conduite en urbain, notamment des incohérences dans les formes des objets et des distorsions spatiales, des problèmes de cohérence de la scène affectant la profondeur et l'alignement [2,3]. En effet, les objets au sein de la scène peuvent manquer d'alignement, de perspective ou de relations de profondeur appropriées, ce qui conduit à des compositions non naturelles. De plus, Stable Diffusion rencontre souvent des difficultés avec les motifs répétitifs, où des textures, des arbres ou des marquages routiers apparaissent clonés de manière artificielle, réduisant ainsi le réalisme des images générées [4].
Pour résoudre ces problèmes d'incohérence lors de la génération de scènes, certains travaux [5,6,7] ont proposé d'ajouter un modèle ControlNet qui prend en entrée une condition pour guider la génération. La condition d'entrée est souvent une carte sémantique ou de profondeur représentant la scène à générer. Cependant, la génération des conditions elles-mêmes est assez complexe et présente des incohérences similaires à celles de la scène, qui se répercutent également sur celle-ci.

Objectifs :

Cette thèse vise à améliorer le processus d'apprentissage du modèle Stable Diffusion en intégrant des connaissances et des indices contextuels dans son espace latent, afin d'améliorer la synthèse de cartes de segmentation sémantique réalistes et cohérentes et/ou d'autres modalités pour collecter des ensembles de données synthétiques pour l'entraînement de modèles de vision par ordinateur. Les modalités générées seront utilisées comme unités de contrôle pour générer la scène correspondante tout en garantissant la cohérence d'un environnement urbain et des images sans incohérence. Cette approche est motivée par le fait que l'unité ControlNet permet au modèle Stable Diffusion de respecter l'image conditionnée. Par conséquent, nous croyons que si nous résolvons les problèmes au niveau de l'image conditionnée en intégrant des connaissances et des indices contextuels, nous obtiendrons des images de scène réalistes et cohérentes pour construire les ensembles de données d'entraînement.
En intégrant des connaissances structurées dans le modèle, nous garantissons une meilleure cohérence spatiale et une prise de conscience sémantique dans les sorties générées. L'approche

proposée cible les mécanismes d'attention, affinant les interactions des caractéristiques pour capturer des dépendances significatives entre différentes régions. Cette intégration guidera le processus de diffusion vers des structures sémantiquement cohérentes/plausibles, réduisant ainsi les incohérences courantes dans les modèles génératifs actuels. Les représentations latentes sensibles au contexte aideront à préserver des agencements de scènes réalistes.

Les défis de la philosophie d'apprentissage proposée sont principalement liés à la représentation du système à base de connaissances (KBS) et à son intégration dans la boucle de diffusion. En effet, le domaine de la conduite urbaine possède des indices et des connaissances riches qui sont complexes à définir complètement, mais leur richesse promet des améliorations potentielles pour l'apprentissage du modèle Stable Diffusion. Après la définition du KBS, l'étape clé consiste à le représenter de manière à permettre son intégration dans la boucle de diffusion. L'objectif de cette intégration est de fournir au modèle de diffusion une perte importante et pertinente concernant les incohérences de la scène. Nous croyons que les sorties de l'attention entre le prompt et la génération, contiennent toutes les informations nécessaires pour quantifier l'incohérence et la renvoyer comme une perte au modèle. Cette perte sera calculée sur la base de l'intégration des connaissances et des indices contextuels extraits du KBS. Une fois les cartes de segmentation générées, elles seront utilisées comme conditions dans un autre modèle de Stable Diffusion pour contrôler la génération dans l'espace d'attention, ce qui permettra de synthétiser des scènes plus réalistes et cohérentes.

Références bibliographiques :

[1]. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. CVPR.

[2]. Wang, Y., Zhang, L., & Li, H. (2023). *Challenges in AI-Generated Outdoor Scenes: A Study on Diffusion Models*. ICCV.

[3]. Xu, J., Chen, R., & Zhao, T. (2024). Improving Scene Coherence in Text-to-Image Generation. NeurIPS.

[4]. Chatterjee, Agneet, et al. "Getting it right: Improving spatial consistency in text-to-image models." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024.

[5]. Bevacqua, Austin, Tanmay Singha, and Duc-Son Pham. "Enhancing Semantic Segmentation with Synthetic Image Generation: A Novel Approach Using Stable Diffusion and ControlNet." 2024 International Conference on Digital Image Computing: Techniques and Applications (DICTA). IEEE, 2024.

[6]. Zhao, Shihao, et al. "Uni-controlnet: All-in-one control to text-to-image diffusion models." *Advances in Neural Information Processing Systems* 36 (2023): 11127-11150.

[7]. Zhang, Fan, et al. "Atlantis: Enabling underwater depth estimation with stable diffusion." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024.

Profil demandé :

- Master en Vision par Ordinateur, Apprentissage Automatique ou Intelligence Artificielle.
- Compétences solides en Python et PyTorch, avec de l'expérience dans les modèles génératifs profonds (en particulier les modèles de diffusion).
- Bonne compréhension des mécanismes d'attention et de la segmentation sémantique.
- Compréhension approfondie de l'IA générative, en particulier des modèles de diffusion tels que Stable Diffusion et ControlNet.
- Familiarité avec les systèmes à base de connaissances et la modélisation sensible au contexte dans l'IA.

- Une bonne maîtrise de l'anglais (oral et écrit) est exigée

Financement : MESRI établissement UTBM (3 ans)

Dossier à envoyer pour le 15 mai 2025
Début estimée du contrat : septembre/octobre 2025

Les candidatures doivent être envoyées à Prof. Hilaire <vincent.hilaire@utbm.fr>, et Dr. Kas mohamed.kas@utbm.fr par email.

Le dossier de candidature doit contenir : un CV détaillé, une copie du diplôme de Master ou tout document attestant du niveau de Master, une copie des bulletins de notes de Master, références et/ou une à deux lettres de recommandation.

Direction / codirection de la thèse :
Directeurs de thèse : Vincent Hilaire
Co-encadrement de thèse : Mohamed KAS

école doctorale **sciences pour l'ingénieur et microtechniques**

PhD title: Knowledge-driven guidance of Stable Diffusion in latent space: Application to visual data generation for training computer vision deep models

Host laboratory: Connaissance et Intelligence Artificielle Distribuées (CIAD) – http://www.ciad-lab.fr

Specialty of PhD: Computer Science

Keywords: Generative IA, Knowledge-Based System, Driving Scene Synthesis, Attention Mechanisms, control in latent space.

Job description:

Context:

Stable Diffusion model has achieved remarkable success in generating high-quality images across various domains due to its efficiency and open-source accessibility [1]. However, it faces limitations in outdoor scene generation, including inconsistencies in object shapes and spatial distortions, scene coherence issues affecting depth and alignment, and repetitive patterns that create unnatural textures [2,3]. Indeed, generated elements such as roads, buildings, and vehicles may exhibit spatial distortions or unrealistic transformations. Regarding scene coherence, the objects within the scene may lack proper alignment, perspective, or depth relationships, leading to unnatural compositions. Additionally, Stable Diffusion often struggles with repetitive patterns, where textures, trees, or road markings appear unnaturally cloned, reducing realism [4].

To deal with these issues in terms of inconsistency during scene generation, some works [5,6,7] proposed to add a ControlNet model that takes an input to condition the generation. The input condition often is a semantic map or depth representing the scene to be generated. However, the generation of the conditions themselves is quite challenging and presents similar inconsistencies to the scene ones, which are reflected also on the scene.

Objectives

This PhD project aims to enhance the learning process of Stable Diffusion by integrating knowledge and contextual cues into its latent space, improving the synthesis of realistic and coherent semantic segmentation maps and/or other modalities in order to collect synthetic datasets for training deep vision models. The generated modalities will be used as control units to generate the corresponding scene while ensuring the coherence of an urban environment and inconsistency-free images. This approach is motivated by the fact that ControlNet unit helped Stable Diffusion to respect the condition image. Therefore, we believe that if we solve the issues on the condition image level by integrating knowledge and contextual cues, we will get realistic and coherent scene images to build the training datasets.

By embedding structured knowledge into the model, we ensure better spatial consistency and semantic awareness in the generated outputs. The approach targets attention mechanisms, refining feature interactions to capture meaningful dependencies between different regions and relationships between classes. This integration will guide the diffusion process toward semantically coherent/plausible structures, reducing inconsistencies common in current generative models. Context-aware latent representations will help in preserving realistic scene layouts.

The proposed learning philosophy challenges are mainly linked to the knowledge-based system (KBS) representation and its integration into the latent diffusion (denoising) loop. Indeed, the urban driving field has rich cues and knowledge which is complex to be completely defined, but its richness promises potential improvements for the learning of Stable Diffusion. After the definition of the KBS, the key step is to represent it in a way that will lead to its integration in the diffusion loop. The integration aim is to provide the diffusion model with prominent loss about the inconsistencies and scene coherence. We believe that the cross-attention outputs, linked to their tokens (words), contain all the necessary information to quantify the inconsistency and feed it back as a loss to the model. This loss will be calculated based on the integration of the knowledge and contextual cues extracted from the KBS. Once the maps are generated, they will be used as conditions in another Stable Diffusion model to control the generation at attention space, which will lead to synthetize more realistic and consistent scene.

References:

[1]. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. CVPR.

[2]. Wang, Y., Zhang, L., & Li, H. (2023). *Challenges in AI-Generated Outdoor Scenes: A Study on Diffusion Models*. ICCV.

[3]. Xu, J., Chen, R., & Zhao, T. (2024). Improving Scene Coherence in Text-to-Image Generation. NeurIPS.

[4]. Chatterjee, Agneet, et al. "Getting it right: Improving spatial consistency in text-to-image models." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024.

[5]. Bevacqua, Austin, Tanmay Singha, and Duc-Son Pham. "Enhancing Semantic Segmentation with Synthetic Image Generation: A Novel Approach Using Stable Diffusion and ControlNet." 2024 International Conference on Digital Image Computing: Techniques and Applications (DICTA). IEEE, 2024.

[6]. Zhao, Shihao, et al. "Uni-controlnet: All-in-one control to text-to-image diffusion models." *Advances in Neural Information Processing Systems* 36 (2023): 11127-11150.

[7]. Zhang, Fan, et al. "Atlantis: Enabling underwater depth estimation with stable diffusion." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024.

Candidate Profile:

- Master's degree in Computer Vision, Machine Learning, or Artificial Intelligence.
- Strong skills in Python and PyTorch, with experience in deep generative models (especially diffusion models).
- Good understanding of attention mechanisms, cross-attention, and semantic segmentation.
- Solid understanding of generative AI, especially diffusion models like Stable Diffusion and ControlNet.

- Familiarity with knowledge-based systems and context-aware modeling in AI.
- Advanced level in English writing and speaking is required.

Financing Institution: French Ministry / UTBM (3 years)

Application deadline: May, 15th 2025
Expected Start of contract: September/October 2025

Applications must be sent to Prof. Hilaire vincent.hilaire@utbm.fr, and Dr. mohamed.kas@utbm.fr by email.

The application must include: a detailed CV, a copy of the Master degree or any document attesting the Master level, a copy of the Master transcripts, references and/or one to two recommendation letters.

Supervisor(s):
Supervisors: Prof. Vincent Hilaire
Co-supervisor: Dr. Mohamed Kas