IIEEJ Transactions on Image Electronics and Visual Computing

Vol. 12, No. 2 2024

The Institute of Image Electronics Engineers of Japan

Editorial Committee of IIEEJ

Contact Information

The Institute of Image Electronics Engineers of Japan (IIEEJ)
3-35-4-101, Arakawa, Arakawa-ku, Tokyo 116-0002, Japan
Tel : +81-3-5615-2893　Fax : +81-3-5615-2894
E-mail : hensyu@iieej.org
http://www.iieej.org/ (in Japanese)
http://www.iieej.org/en/ (in English)
http://www.facebook.com/IIEEJ (in Japanese)
http://www.facebook.com/IIEEJ.E (in English)

IIEEJ Transactions on
Image Electronics and Visual Computing
Vol.12 No.2 December 2024

CONTENTS

Published two times a year by the Institute of Image Electronics Engineers of Japan（IIEEJ）
3-35-4-101, Arakawa, Arakawa-ku, Tokyo 116-0002, Japan
Tel：+81-5615-2893 Fax：+81-5615-2894 E-mail：hensyu@iieej.org http://www.iieej.org/

IIEEJ **Paper**

2D-3D Registration Method for X-Ray Image Using 3D Reconstruction Based on Deep Neural Network

Pragyan SHRESTHA†, Chun XIE†, Yuichi YOSHII††, Itaru KITAHARA†(*Member*)

† University of Tsukuba, †† Tokyo Medical University

<**Summary**> This paper proposes a method for registering X-ray images with its 3D CT model by estimating 3D point clouds from X-ray images and their corresponding points on the image. Many conventional methods generate a simulated X-ray image from a 3D CT model and optimize the pose by using the similarity metrics between the simulated X-ray and the input X-ray image. On the other hand, deep learning approaches that predict pose information need a canonical coordinate system defined manually on the pre-operative CT to properly utilize the estimated pose. Therefore, we devise a fully automatic registration pipeline that is independent of coordinate system, by recovering 3D point clouds from X-ray images, estimating the corresponding points on the images, and aligning them with the given 3D CT model.

Keywords: 2D-3D registration, 3D reconstruction, camera pose estimation, pn, icp, x-ray image

1. Introduction

Radiological imaging is one of the most important technologies in modern medical systems and diagnosis. Especially in Interventional Radiology (IVR), minimally invasive surgery is performed using various imaging techniques. In orthopedic surgeries such as osteosynthesis and osteotomy, the projection of a 3D CT model is superimposed on the intraoperative X-ray image for checking the surgical progress (i.e., if the pedicle screws are placed according to the CT planning). The pre-operative CT, which is acquired before surgery, is useful for planning and simulation of fracture reduction, while intraoperative X-ray images are used for guiding purposes such as intraoperative implant positioning and bone cutting[1]. However, since X-ray images are transmissive in nature, depth information is difficult to obtain. Therefore, the surgeon must estimate the spatial location and shape of the target region mentally. Many studies have been conducted to reduce the burden on the surgeon by superimposing a preoperative CT image on the X-ray image in the appropriate posture. In general, methods for mapping a 3D CT model to a 2D X-ray image can be classified according to various criteria such as the target modality, the projection parameters to be estimated, and the similarity function[2]. This research focuses on aligning with the pre-operative CT model without manual interventions. Specifically, the following issue could be solved.

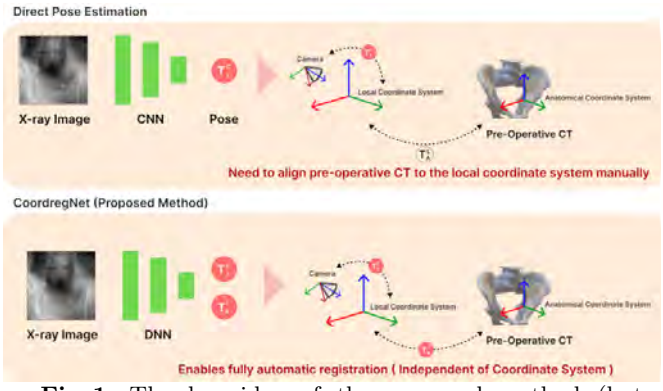


Fig. 1 The key idea of the proposed method (bottom) and issue with the recent deep learning approaches (top)

- The need to manually intervene the registration process by finding landmarks and defining the canonical coordinate system in preoperative CT.

An illustration of the problem is depicted in **Fig.1** (top). With direct pose estimation based models, 1) the pose vector is output directly conditioned on the X-ray image. 2) the pose vector only captures the relation between the camera and the canonical coordinate system (i.e., T_l^c). 3) Therefore, the pre-operative CT must be transformed to the canonical coordinates for use in clinical practice (i.e., T_a^l). This requires some form of manual intervention. With our proposed method (Fig.1 bottom), this process is alleviated by 1) decomposing the pose into 2D-3D registration component (i.e., T_l^c) and 3D-3D registration component (i.e., T_a^l) Through this, the physicians

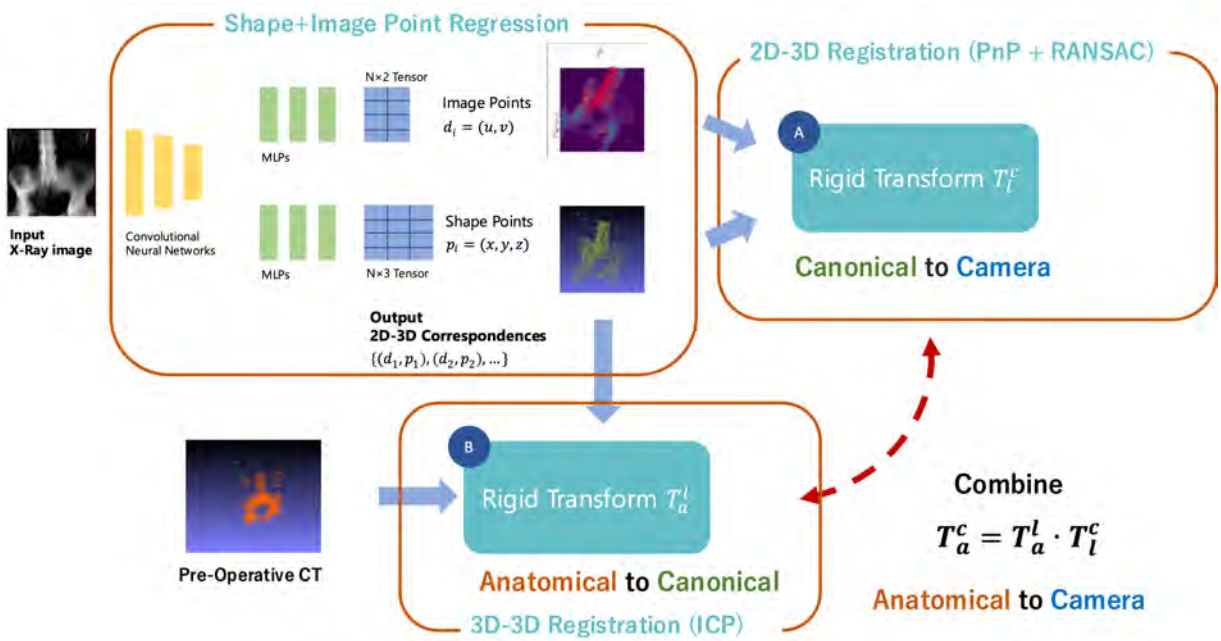


Fig. 2 The Architecture diagram and the whole pipeline of the proposed method. Component (A) represents 2D-3D registration and Component (B) represents 3D-3D registration

do not have to transform the CT model to a canonical coordinate system, which usually involves finding landmarks and calculating transforms such that certain landmarks lie on the reference points in the canonical coordinate system (e.g., anterior pelvic planes in the case of pelvis)[3].

The proposed method uses deep learning to reconstruct the 3D model from the X-ray image, estimate the corresponding points on the image, and then align the model with the preoperative CT model. An overview diagram is shown in **Fig.2**. 1) First, the 3D point cloud and corresponding points on the image are estimated from the input X-ray image. 2) In section A, the external parameters of the camera in the canonical coordinate system are obtained using Random Sample and Consensus (RANSAC)[4]and Perspective-n-Point (PnP)[5]algorithms. 3) in section B, the point cloud is extracted from the target CT image and aligned with the estimated point clouds using Iterative Closest Points (ICP)[6]. This yields a rigid body transformation that maps the target model in the anatomical coordinate system into the camera coordinate system of the X-ray device. In the experiments, a network that estimates the 3D point cloud and corresponding points on the image, from the X-ray image, is trained using the CT-ORG[7]dataset. The pose obtained using the point cloud and image correspondence points estimated from the test data was evaluated in terms of rotation and translation error. It is also compared with a deep learning model PoseNet[8], that directly regresses the camera position and orientation. We choose PoseNet for the direct pose estimation based baseline model because of its simplicity and effectiveness.

2. Related Works

An early application of deep learning to the problem of registering X-ray images with 3D CT models is the CNN-based pose estimation developed by Miao et al[9] . They applied CNN to output parameters of camera pose (6 degrees of freedom) in the final layer from X-ray images. According to Sattler et al.[10], such a pose regression network can be considered as a form of image retrieval. Moreover, It can fail if the object in the image is different from the one used in training. Bier et al.[11]used CNN to detect anatomically meaningful landmarks in X-ray images of the pelvis to map them to the 3D model. This method provides higher registration accuracy and can be performed in less time. However, it requires expert annotation in 3D and 2D to train the network.

Furthermore, Liao et al.[12]developed a network that eliminates the need for anatomical landmarks by identifying the projection points of randomly extracted points from a 3D model across all images, given multi-view X-ray images. Once the projection points are identified, triangulation can be used to register all the multi-view images with the 3D model. Given that this method is tailored for multi-view images, its registration accuracy

for a single image is comparatively low.

Jaganathan et al.[13]devised a network that updates poses sequentially. They constrain the correspondence between the projected points on the boundary of the object in the CT model and the points on the contour line in the X-ray image using point-to-plane correspondence (PPC) and conduct registration by sequentially updating the rigid transformation of the CT model. Most of the methods described above need some form of manual intervention for the registration to succeed if we are given only the pre-operative CT and no other coordinate information. The method proposed by Liao et al. can be processed in full automation however it requires multi view projection images for registration.

3. Method

3.1 Problem formulation

In general, the 2D-3D registration problem for finding rigid body transformations is formulated as follows:

$$\hat{T} = \arg\min D(I_{real}, G \circ T(V_{CT})) \tag{1}$$

where T is the transformation matrix, V_{CT} is the given 3D model, I_{real} is the X-ray image, D is the similarity metric, and G is the function that generates the simulated image from the 3D model. In the proposed method, we consider solving the problem expressed in Equation (1) by transforming it into the following Equations (2), (3) and (4).

$$\hat{T}_l^a = \arg\min D(\hat{T}_l^a \circ \bar{G}(I_{real}), V_{CT}) \tag{2}$$

$$\hat{T}_l^c = \arg\min D(\hat{T}_l^c \circ \bar{G}(I_{real}), I_{real}) \tag{3}$$

$$\hat{T} = \hat{T}_l^c \circ \hat{T}_l^a \tag{4}$$

where $\bar{G}$ is the function that outputs the 3D point cloud from the X-ray image, $\hat{T}_l^a$ is the transformation from the anatomical coordinate system used in the CT image to the canonical coordinate system defined in the point cloud coordinates, while $\hat{T}_l^c$ is the transformation from the anatomical coordinate system to the X-ray device coordinate system.

Equation (4) is the key idea in our proposed method. The transformation matrices obtained in each section (2D-3D and 3D-3D registration) are combined to produce the final transformation matrix that transforms the anatomical coordinates into camera coordinates for further overlay projections.

3.2 Network architecture

We propose a robust registration framework by decoupling 3D reconstruction and the camera pose estimation which is then integrated with Equation (4) for final alignment. Equation (2) represents the ICP part and Equation (3) represents the PnP + RANSAC part respectively in Fig.2. The proposed method uses a convolutional neural network and two MLP layers to regress the 3D point coordinates in the canonical system and their corresponding 2D coordinates on the image from the input X-ray image as shown in Fig.2. We use ResNet50[14]as the backbone for CNN. The output of the final convolutional layer is used as input to the respective MLP branches. Chamfer Distance is used for the loss function of the point cloud, and mean squared error is used for the corresponding points on the image.

3.3 Registration pipeline

During inference, PnP + RANSAC is performed based on the 2D corresponding points and the 3D point coordinates to obtain a rigid body transformation from the canonical coordinate system to the camera coordinate system. Furthermore, a 3D point cloud from the target CT is created by obtaining the gradient values, and sampling points by thresholding gradient magnitude. Since the estimated point cloud is defined in the canonical coordinate system, the point cloud of the target CT is transformed to the canonical coordinate system using ICP or manually. Finally, the point cloud of the target CT converted to the canonical coordinate system is projected using the camera pose obtained by PnP + RANSAC, overlapping with the X-ray image to achieve registration.

4. Experimental Results

4.1 Dataset

Experiments were conducted to evaluate the camera pose estimation pipeline of the proposed method. Simulated images were generated from CT data using the camera pose with added Gaussian noise for training and testing. CT-ORG[7]was used as the CT dataset for generating the simulated images. Five CT data were selected in CT-ORG that contained the pelvis. The pelvis region was extracted using a bone region segmentation mask. To generate the simulated images from this volume data, the following X-ray transformation equation was used.

$$I_{DRR} = \int_L V_{CT} dr \tag{5}$$

where L is the ray from the X-ray source to the pixel on

Table 1 Quantitative result of registering 2D X-ray images with 3D point clouds

Dataset	Method	Rot. [deg]		Trans. [mm]		X [mm]		Y [mm]		Z [mm]		Runtime [s]	
		mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
VOL-A	**Ours**	4.99	7.07	72.70	85.24	13.47	8.20	14.31	9.94	32.28	13.46	0.0078	0.0004
	PoseNet	8.91	13.63	93.42	116.42	22.86	32.50	19.35	37.05	28.24	49.11	0.0062	0.0002
VOL-B	**Ours**	14.29	18.12	226.04	245.46	11.82	7.49	10.32	5.05	100.60	113.78	0.0083	0.0006
	PoseNet	19.78	22.16	203.89	178.93	36.58	63.74	23.95	49.89	70.21	103.83	0.0063	0.0002

Algorithm 1 Generating camera poses

1: $p \sim \mathcal{N}(0, \sigma)$
2: $z \leftarrow \frac{x_{center}}{\|x_{center}\|} + p$
3: $y \leftarrow (0, 0, 1)^T + p$
4: $x \leftarrow y \times z$
5: $R \leftarrow (x, y, z)^T$

the detector plane.

Furthermore, the pose for generating the image was set so that the X-ray source was placed in a cylindrical shape and the viewpoint was oriented to the center of the volume, considering the geometry of the C-arm. Specifically, the procedures in **Algorithm 1** were used to calculate the camera pose for simulation. For the training data, the Gaussian noise vector was obtained by sampling each coordinate from a Gaussian distribution with a mean of 0.0 mm and a standard deviation of 1.0 mm.

We prepared two kinds of test data. One test set contains 1000 images generated from one of the CT volumes used in the training but were held out from the training dataset (i.e., we refer to this as VOL-A). This represents same volume different views. Another test set contains 1,000 images generated from a CT volume that was not used during training (i.e., we refer to this as VOL-B). For each test data, the camera poses were generated following Algorithm 1. However, the noise vector was sampled from a gaussian distribution with a mean of 0.0 mm and a standard deviation of 10.0 mm, which is larger than that of the training data. **Figure 3** shows the visualization of the generated camera poses in the left and the sample images of simulated X-ray images.

The training data consists of 3,000 X-ray images per volume, GT point clouds, and corresponding points on the images. In this experiment, CT data from five patients were used, making it 15,000 samples in total. The total number of samples in the test dataset was 1,000. The same training and testing dataset were used for the direct regression model of camera pose (hereafter referred to as the comparison model).

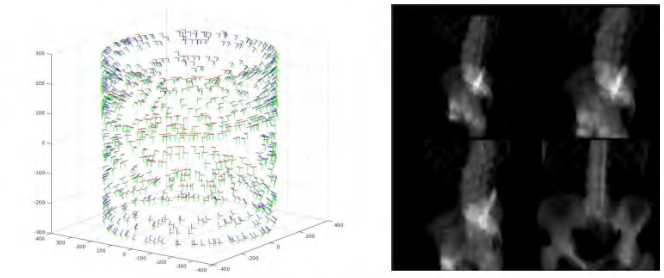


Fig. 3 Visualization of camera positions and orientations for simulation (left) and example of generate X-ray images

4.2 Implementation details

The network depicted in Fig.2 was implemented using PyTorch[15]. OpenCV[16]was used to apply PnP + RANSAC, and Open3D[17]was used for generating training point clouds from CT data. The model of the proposed method was trained for about 6 hours using RTX 3090.

4.3 2D-3D registration

We evaluate the 2D-3D registration pipeline, which is denoted by (A) in Fig.2 by considering two different scenarios. The rigid transformation obtained in this step refers to T_l^c in Equation (4). VOL-A refers to the dataset containing X-ray images that were generated from the same CT data as used in training albeit in different viewpoints. VOL-B refers to the dataset containing X-ray images generated from a completely different CT data.

Table 1 shows the registration results evaluated in terms of rotation errors and translation error of the predicted transformation with the ground truth transformation matrices for these two different datasets with the proposed method (Ours VOL-A, Ours VOL-B) and comparison method (PoseNet VOL-A, PoseNet VOL-B).

Figure 4 shows the reconstructed point clouds and the box plot for rotation and translation errors. For VOL-A, the proposed method had rotation error of 4.99 +/- 7.07 deg and translation error of 72.70 +/- 85.24 mm. While for VOL-B, the proposed method had rotation error of 14.29 +/- 18.12 deg and translation error of 226.04 +/-

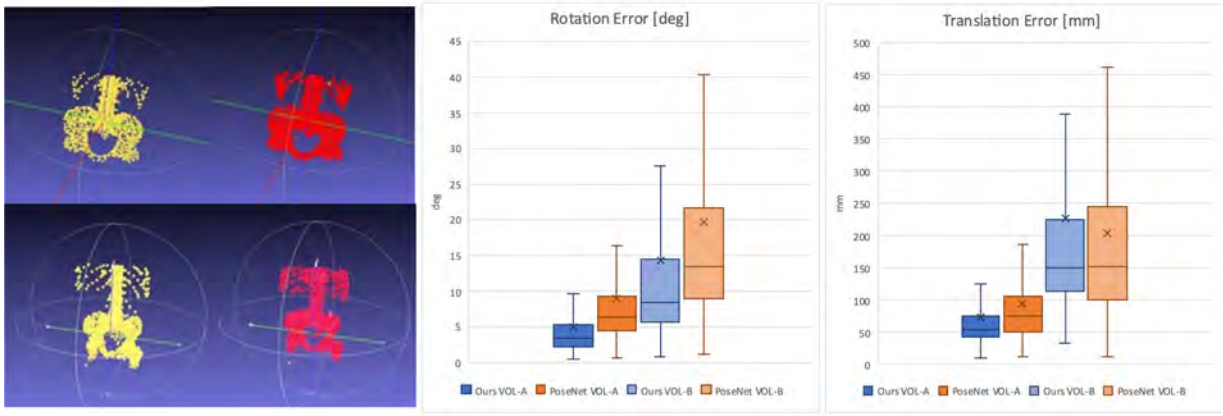


Fig. 4 Reconstructed point clouds (yellow) and ground truth (red) on the left and results for 2D-3D registration pipeline on the right

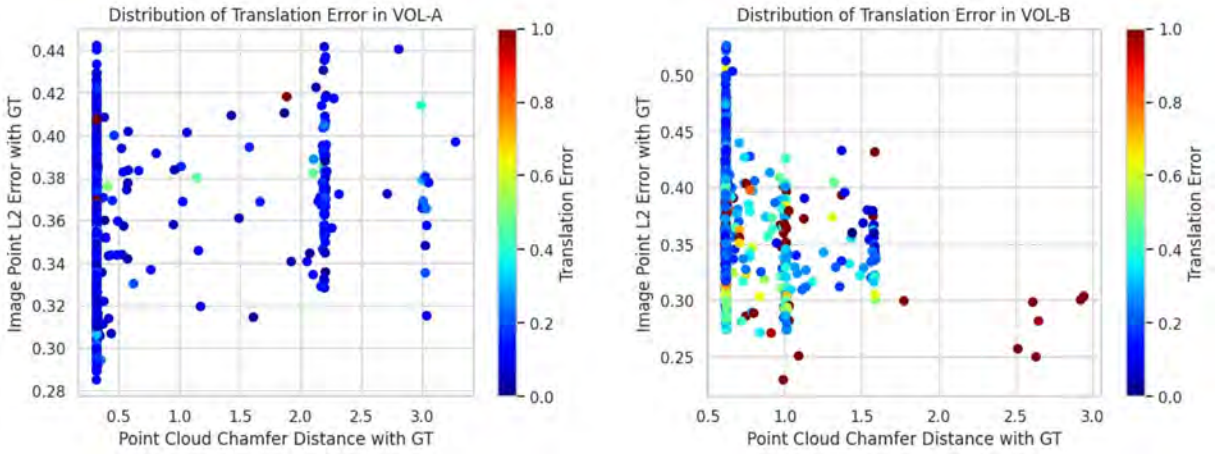


Fig. 5 Distribution of Translation Error with respect to chamfer distance of estimated point cloud with GT and L2 error of estimated image points with GT

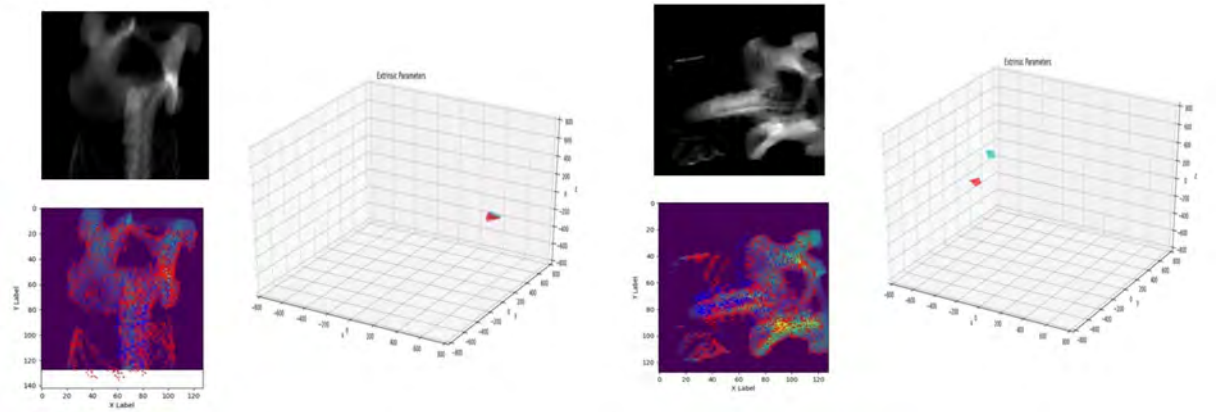


Fig. 6 Registration result with small error (Left) and large error (Right)

245.46 mm. This decrease in accuracy for data completely unseen during training is expected due to small number of CT scans used for training. The rotation error was lower in both VOL-A and VOL-B when compared with PoseNet while translation error for VOL-B was higher in the proposed method. This is due to the larger error in z-axis (depth direction) that exists in the proposed method. Since the proposed method uses perspective-n-point algorithm for pose estimation, small error in image points can lead to larger error in depth direction.

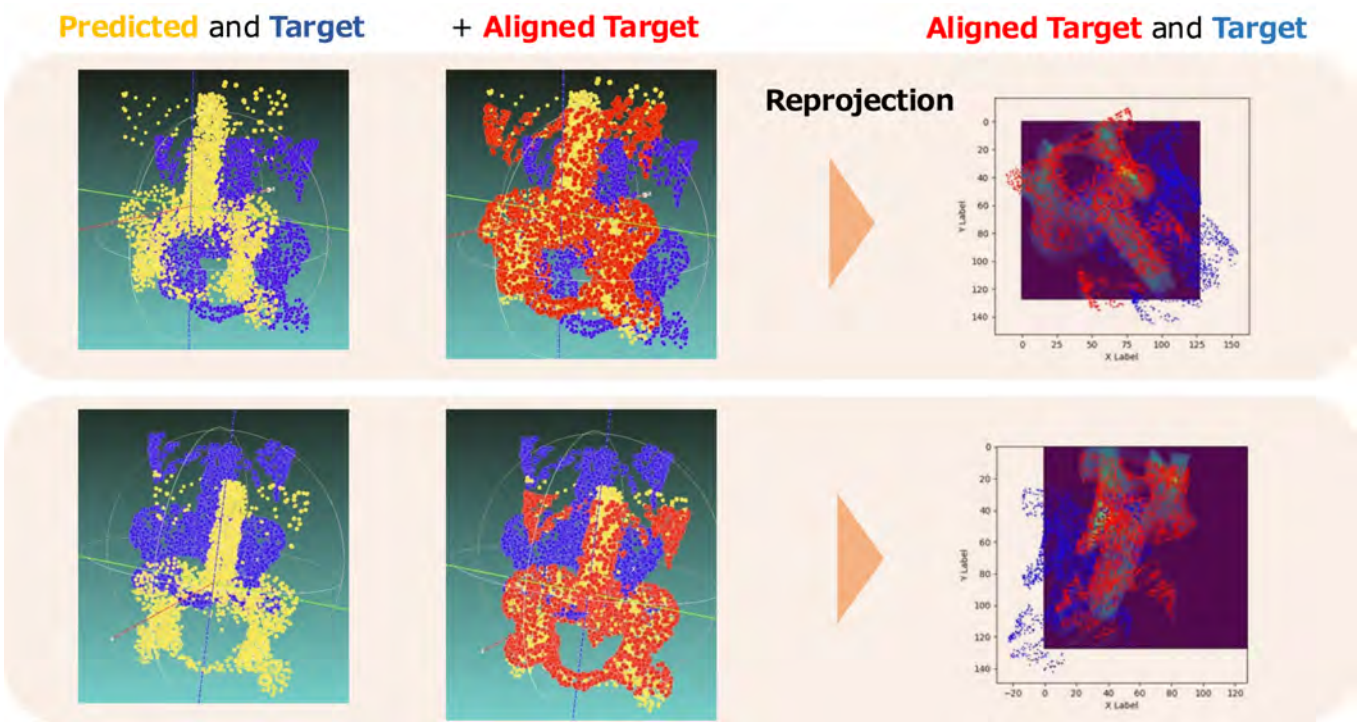


Fig. 7 Qualitative results of 3D-3D registration along with reprojection of target point clouds with and without the alignment

We have also evaluated the estimated point clouds and image point using chamfer distance and L2 error each. The impact of these errors on translation component is shown in **Fig.5**. VOL-B shows a tendency of increasing translation error with increasing point cloud error. Interestingly, VOL-A does not show such tendency. We speculate this to be caused by the effect of RANSAC on finding good correspondences on VOL-A while failing on VOL-B. An example of small registration error and large registration error each is shown in **Fig.6**. For each case, the input image is on the top left, reprojection of 3D point cloud on the bottom left, and the estimated (blue) / ground truth (right) camera pose on the right.

4.4 3D-3D registration

We evaluate the 3D-3D registration pipeline, which is denoted by (B) in Fig.2 by adding random offsets to the ground truth target point clouds. The reason for only allowing translation offsets is that generally CT scans contain forward-backward, left-right information while the origin of the anatomical coordinate system varies. The random offset vector was sampled from uniform distribution centered around zero with standard deviation of 100 mm. Furthermore, the X-ray image used in this experiment is from VOL-A because the 2D-3D registration results from VOL-B are not accurate enough for combining transformation matrices to get the reasonable final pose.

In **Fig.7**, the yellow point cloud is the estimated point cloud given by the network, and the blue point cloud is the target (i.e., point cloud obtained from pre-operative CT scan which is defined in anatomical coordinate system). In the same figure, the top row and the bottom row shows two different scenarios where the origin of the target point cloud varies. Using Point-to-point ICP algorithm, the rigid transformation for transforming target point cloud to the predicted point cloud (i.e., T_a^l in Equation (4) is computed. Using Equation (4), we can obtain the final transformation matrix that transforms the target point cloud the camera coordinate system which can then be reprojected into the image. In the right most image of each row, the reprojected points of target point cloud without alignment (i.e., using only T_l^c) is shown in blue. While the reprojected points of the target point cloud after alignment (i.e., using $\hat{T}$) is shown in red. It is evident that the latter case results in better registration because of the inherent mismatched pose in the target point cloud without alignment.

5. Discussions and Conclusion

In this research, a deep learning-based method for fully automatic registration of 2D X-ray images with 3D CT

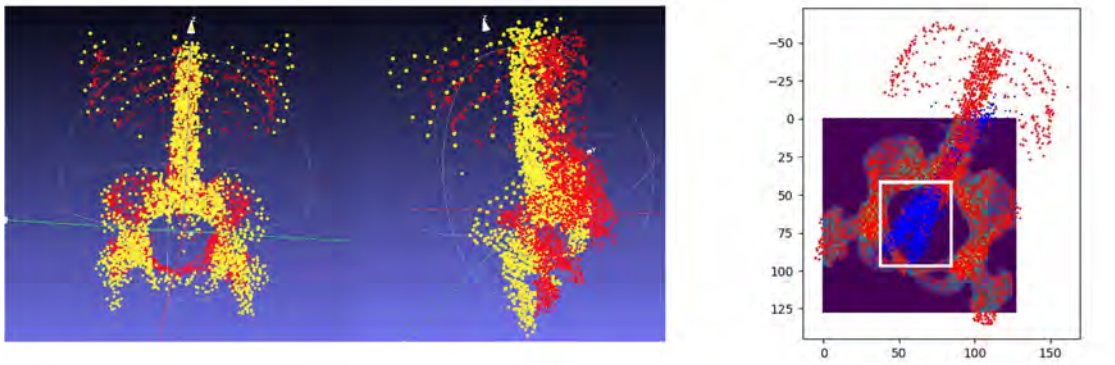


Fig. 8 Example of predicted point clouds (yellow) and ground truth point clouds (red) for VOL-B dataset on the left, predicted image points (blue) and ground truth image point (red) on the right

models was proposed. The pipeline focused on solving the issue of manually defining coordinate system on the pre-operative CT scan by decomposing the registration problem into 2D-3D registration and 3D-3D registration.

Experimental results verified the accuracy of 2D-3D registration quantitatively and it was found to have better rotation accuracy, but lower translation accuracy compared to that of direct pose estimation method. The resulting accuracy is not sufficient for clinical use however, while only qualitative results were shown for 3D-3D registration pipeline, it shows promising results for alleviating the manual intervention needed in practice. Furthermore, the runtime is under 10ms in both methods, enabling real-time usage.

Since we trained our model using only 5 CT variations, it was difficult for the model to generalize to previously unseen CT. This can be observed from the image point prediction results shown in **Fig.8**. The reconstructed point cloud originates slightly differently compared to the ground truth, yet its overall shape remains consistent. On the other hand, the predicted image points were cluttered around the center region shown in white box. This happens due to the network not generalizing to this shape and view of the CT data. Future works can address this problem by simply training the model on large number of datasets. To improve the core 2D-3D registration accuracy, one approach would be to re-place the simple multi-layer perceptrons with PointNet[18]like architecture specifically designed for point clouds.

Acknowledgments

This work was partially supported by a grant from JSPS KAKENHI grant number JP23K08618.

References

1) Y. Yoshii, T. Kusakabe, K. Akita, W. L. Tung, T. Ishii, "Reproducibility of three dimensional digital preoperative planning for the osteosynthesis of distal radius fractures," in *Journal of Orthopaedic Research*, vol. 35, no. 12, pp. 2646–2651 (Dec. 2017).
2) P. Markelj, D. Tomaževič, B. Likar, F. Pernuš, "A review of 3D/2D registration methods for image-guided interventions," in *Medical Image Analysis*, vol. 16, no. 3, pp. 642–661 (Apr. 2012).
3) C. Liu, F. Hu, Z. Li, Y. Wang, X. Zhang, "Anterior Pelvic Plane: A Potentially Useful Pelvic Anatomical Reference Plane in Assessing the Patients' Ideal Pelvic Parameters Without the Influence of Spinal Sagittal Deformity," in *Global Spine J*, vol. 12, no. 4, pp. 567–572 (May. 2022).
4) M. A. Fischler, R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," in *Communications of the ACM*, vol. 24, no. 6, pp. 381–395 (Jun. 1981).
5) X. X. Lu, "A Review of Solutions for Perspective-n-Point Problem in Camera Pose Estimation," in *Journal of Physics: Conference Series*, vol. 1087, no. 5, p. 052009 (Sep. 2018).
6) S. Rusinkiewicz, M. Levoy, "Efficient variants of the ICP algorithm," in *Proceedings Third International Conference on 3-D Digital Imaging and Modeling*, Quebec City, Que., Canada (2002).
7) B. Rister, D. Yi, K. Shivakumar, T. Nobashi, D. L. Rubin, "CT-ORG, a new dataset for multiple organ segmentation in computed tomography," in *Scientific Data*, vol. 7, no. 1, p. 381 (Nov. 2020).
8) A. Kendall, M. Grimes, R. Cipolla, "PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization," in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2938–2946 (2015).
9) S. Miao, Z. J. Wang, R. Liao, "A CNN Regression Approach for Real-Time 2D/3D Registration," in *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1352-1363 (May. 2016).
10) T. Sattler, Q. Zhou, M. Pollefeys, L. Leal-Taixé, "Understanding the Limitations of CNN-Based Absolute Camera Pose Regression," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3297–3307 (2019).
11) B. Bier et al., "X-ray-transform Invariant Anatomical Landmark Detection for Pelvic Trauma Surgery," in *International*

Conference on Medical Image Computing and Computer-Assisted Intervention (2018).

12) H. Liao, W.-A. Lin, J. Zhang, J. Zhang, J. Luo, S. K. Zhou, "Multiview 2D/3D rigid registration via a point-of-interest network for tracking and triangulation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA (2019).
13) S. Jaganathan, J. Wang, A. Borsdorf, K. Shetty, A. Maier, "Deep Iterative 2D/3D Registration," in *Medical Image Computing and Computer Assisted Intervention*, vol 12904, pp. 383–392 (2021).
14) K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778 (2015).
15) A. Paszke et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," arXiv [cs.LG], (Dec. 2019).
16) I. Culjak, D. Abram, T. Pribanic, H. Dzapo, M. Cifrek, "A brief introduction to OpenCV," in *2012 Proceedings of the 35th International Convention MIPRO*, pp. 1725–1730 (2012).
17) Q.-Y. Zhou, J. Park, V. Koltun, "Open3D: A Modern Library for 3D Data Processing," arXiv [cs.CV], (Jan. 2018).
18) C. R. Qi, H. Su, K. Mo, L. J. Guibas, "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation," arXiv [cs.CV] (Dec. 2016).

(Received November 10, 2023)
(Revised June 06, 2024)

Pragyan SHRESTHA
He received his B.E. and master's in human informatics from University of Tsukuba in 2021 and 2023 respectively. He is currently a PhD student at University of Tsukuba. His research interests include 3D reconstruction, registration and medical imaging.

Chun XIE
He received his B.Sc. in Computer Science from Hong Kong Baptist University in 2011. In 2012, he received M.E. in Computer Science and Engineering from The Chinese University of Hong Kong. Also, he received M.E. in Systems and Information Engineering and Ph.D. in Human Informatics from University of Tsukuba, Japan, in 2017 and 2020, respectively. He has been an assistant professor at University of Tsukuba since 2023. His research interests include computer vision and augmented reality.

Yuichi YOSHII
He is a professor of orthopedic surgery at Tokyo Medical University Ibaraki Medical Center. In 1996, he received his Doctor of Medicine degree from the University of Tsukuba, School of Medicine. In 2009, he completed his doctoral program at the Graduate School of Comprehensive Human Sciences in the University of Tsukuba. He is certified as a specialist by the Japanese Orthopedic Association, the Japanese Society of Hand Surgery, and the Japanese Society of Rehabilitation Medicine. His main research interests include hand surgery, biomechanics, computer vision, and computer-assisted surgery.

Itaru KITAHARA (*Member*)
He received his B.E. and M.E. degrees in Science Engineering from University of Tsukuba, Japan in 1994 and 1996, respectively. In 1996, he joined Sharp Corporation. 2000-2003, he was a research associate of University of Tsukuba. He received his Ph.D. in 2003. 2003-2005, he was a researcher at ATR. 2005-2019, he was an assistant professor and associate professor at the University of Tsukuba. Since 2019, he has been a professor at the Center for Computational Sciences, University of Tsukuba. He is also technical/academic advisor for IT companies. His research interests include computer vision, mixed reality, and intelligent image media

IIEEJ **Paper**

Jewelry Image-to-Image Translation with Consistency Regularization and Data Augmentations

Taiyo NAKAGAWA† , Tomoko OZEKI† (*Member*)

† Tokai University

<**Summary**> Image enhancement of jewelry is a difficult task because of the shape of the jewelry, its color, background elements such as shadows and glass stands, as well as the blurring of the boundary between the jewelry and the background and unique light reflections. Our preliminary results indicate that CycleGAN is effective in correcting jewelry images and that background elements in jewelry images adversely affect jewelry image correction. In this study, we propose a method to correct jewelry images with strong background elements. The results show that the target consistency of TC-ShadowGAN is effective not only in removing the background but also correcting the jewelry area in the image. In addition, data augmentation with Balanced Consistency Regularization (BCR) and Dense Consistency Regularization (DCR) are applied to increase the accuracy of the correction of the jewelry area.

Keywords: jewelry image, image-to-image translation, generative adversarial networks, target consistency, balanced consistency regularization, dense consistency regularization

1. Introduction

Jewelry is a product whose value is linked to its visual appearance. Therefore, jewelry retailers use photo retouching software to manually process images of jewelry to eliminate the difference between the actual appearance of the jewelry and the image captured by a camera for online transactions. A single jewelry image takes an expert 20 minutes to an hour. There is a need to automate image correction through machine learning.

We propose an image-to-image translation method that corrects a captured jewelry image (Domain X) to an expertly retouched image (Domain Y, Ground Truth). Applying image-to-image translation techniques to jewelry images is challenging due to the shape of the jewelry, its color, and background elements, such as the shadows and glass stands, blurring of the boundary between the jewelry and background, and the unique light reflections. Because of these unique characteristics of jewelry images, algorithms that work well on other large datasets may not work well on jewelry images. In addition, compared to landscape and animal image transformations, jewelry image transformations must be precise enough to withstand online sales. The goal is to propose a generic end-to-end image translation method, rather than a jewelry-specific model, for the challenging subject of jewelry images.

As an image-to-image translation model for jewelry images, we use TC-ShadowGAN[1], which was originally proposed for shadow removal. The performance of the GAN models depends on the quality of discriminators, which distinguish real images from translated fake ones. Therefore we apply balanced consistency regularization (BCR)[2]to improve the removal of background elements. We also apply dense consistency regularization (DCR)[3] to increase the accuracy of the correction of the jewelry area and to clarify the boundary between the jewelry and the background. We show that introducing a combination of BCR and DCR in TC-ShadowGAN improves the correction of jewelry images.

2. Related Research

There are a lot of studies on deep learning-based image-to-image translation. Especially, unpaired models, which do not need paired data, have been proposed.

In the field of low-light enhancement, the Retinex theory, which decomposes an image into reflectance and illumination, has been widely adopted. RUAS[4]and ISSR[5] are models inspired by the Retinex theory. ISSR combines image segmentation with the Retinex model to improve transformation performance. EnlightenGAN[6]uses an attention-guided U-Net as the generator, and two discriminators (global and local) to suppress overexposure

Fig. 1 Comparison of image-to-image translation models applied to jewelry image correction[10)]

and underexposure. Zero-DCE[7)]is a model to estimate pixel-wise and higher-order curves for dynamic range adjustment of a given image and does not need any paired or unpaired data during training.

CycleGAN[8)]is a pioneering study in general-purpose image domain transformation that does not require any paired data by imposing cycle consistency. CUT[9)]is also a general-purpose model, which is based on the contrastive learning in the field of self-supervised learning.

Shizuno et al.[10)]showed that state-of-the art algorithms on large datasets, such as EnlightenGAN[6)], ISSR[5)], Zero-DCE[7)], RUAS[4)]and CUT[9)], do not always perform well on jewelry images due to their unique features. CycleGAN[8)] outperformed other state-of-the art algorithms as shown in **Fig. 1**. CycleGAN[8)]is a type of GAN that achieves a style transformation from domain X to domain Y. The generator G learns a style transformation from domain X to domain Y, and the generator F learns a transformation from domain Y to domain X. CycleGAN introduces cycle consistency loss,

$$L_{cyc}(G,F)=E_{x\sim X}\left[||F(G(x))-x||_1\right] + E_{y\sim Y}\left[||G(F(y))-y||_1\right] \quad (1)$$

which ensures that the bi-directional transformations between domain X and domain Y are consistent.

Nakagawa et al.[11)]applied CycleGAN[8)]to jewelry images with and without background elements and compared the correction results. The results showed that background elements in jewelry images have a negative impact on image correction by CycleGAN. In the case of images with large background elements such as glass stands, excessive brightness enhancement and noise were observed in the jewelry area when training was performed with the background elements left in place as shown in

Fig. 2 Problems in CycleGAN[11)]

Fig. 2. In images where the boundary between the background and the jewelry is blurred, a halation is generated in a part of the jewelry image.

On the other hand, CycleGAN does not progress well when training on jewelry images with the background removed, because they are quite similar to the ground truth images[11)]. This problem was partially solved by introducing the Two Time Update Rule (TTUR)[12)]. In practical applications, however, removing the background from a jewelry image as a preprocessing step is impractical from a cost perspective. In this paper, we focus on the jewelry images with background elements.

CycleGAN learns the inverse transformation from the ground truth to the captured image without additional information such as lighting conditions. As shown in **Fig. 3**, variously generated backgrounds are added to the jewelry image during the learning process. We speculate that the randomly generated background acts as a data augmentation, which indirectly improves the performance of CycleGAN's discriminator and thus the generator's ability to correct for the jewelry area in the image. At the same time, however, fake captured images obtained through training without lighting conditions may deviate from the actual captured image, which may negatively affect training of generator.

Based on these hypotheses, we expect that it is effective to increase the performance of the discriminator in the image-to-image translation model by data augmentation. Therefore, we propose a method to improve the performance of the discriminator by data augmentation and consistency regularizations in a uni-directional image-to-image translation model, which avoids the cycle consistency.

3. Proposed Method

Various methods have been proposed to improve the performance of GAN-type models: (1) balancing the learning of the generator and discriminator[11),12)](2) trying various GAN models[10)](3) hyperparameter tuning (4)

Fig. 3 The generated image $F(y)$ from the ground truth image y and the generated image $G(F(y))$ by the generator G in the learning process of CycleGAN

changing the model structure (5) data augmentation for GAN models[2](6) application of self-supervised method to improve the performance of the discriminator. Our proposed method is based on TC-ShadowGAN and introduces BCR for data augmentation and DCR to boost the quality of discriminators (**Fig. 4**).

TC-ShadowGAN[1]consists of a pair of networks, each having an encoder-decoder type generator and a discriminator. Two independent generators produce the residual images $G_1(x)$ and $G_2(x)$, which are added pixel by pixel to the original image x to produce two translated images. TC-ShadowGAN introduces a target consistency,

$$L_{\mathrm{TC}}(G_1, G_2) = E_x\left[||(x + G_1(x)) - (x + G_2(x))||_1\right] \quad (2)$$

and learns only the transformation from a shadowed image to a shadow-less image, avoiding learning the bidirectional transformations with cycle consistency loss in CycleGAN[8]. TC-ShadowGAN uses the identity loss,

$$L_{\mathrm{Identity}}(G_i) = E_y[||(y + G_i(y)) - y||_1] \quad (i = 1, 2) \quad (3)$$

which guarantees that the two generators G_1 and G_2 perform the identity transformation for real shadow-less images. The identity loss is commonly used in the image-to-image translation models and makes the convergence for training faster.

The discriminator discriminates between real and generated images. The generator G and discriminator D are simultaneously trained using adversarial loss,

$$\begin{aligned}&\min_G \max_D L_{\mathrm{adv}}(D, G)\\ &= E_{P_Y}\left[\log D(x_{\mathrm{real}})\right] + E_{P_X}\left[\log(1 - D(x_{\mathrm{fake}}))\right] \end{aligned} \quad (4)$$

where x_{real} is a real image drawn from the distribution P_Y of images in domain Y, and x_{fake} is the translated fake image of the original image drawn from the distribution P_X in domain X.

Balanced Consistency Regularization (BCR)[2]is a regularization method that makes data augmentation more effective in learning GANs. BCR performs data augmentations both on the real image x_{real} and the translated fake image x_{fake} to make the output consistent,

$$\begin{aligned}L_{BCR} = &||D(x_{\mathrm{real}}) - D(t_1(x_{\mathrm{real}}))||^2\\ &+||D(x_{\mathrm{fake}}) - D(t_2(x_{\mathrm{fake}}))||^2,\end{aligned} \quad (5)$$

where t_1 and t_2 are image transformations.

Dense Consistency Regularization (DCR)[3]is a regularization method based on the idea of self-supervised learning to improve style transformations using GANs. DCR first crops two patches x_1 and x_2 from a single real image. Next, each patch image is passed through an encoder part D_0 of the discriminator (CNN_1, CNN_2 in Fig. 4) to obtain the feature maps, $D_0(x_1), D_0(x_2)$. One feature map $D_0(x_1)$ is then passed through the DCR module, which consists of two 1x1 convolution layers and a leaky ReLU layer. The stop gradient operation is applied to the other feature map. The DCR module and the stop gradient operation are applied to prevent feature collapse to the trivial solution in representation learning. Finally, the two feature maps, $f_{\mathrm{DCR}}(D_0(x_1))$ and $D_0(x_2)$, in the overlapping area Ω of two patches, x_1, x_2, are compared by the negative cosine similarity, which enforces pointwise consistency called Dense Consistency Regularization loss[3],

$$\begin{aligned}L_{\mathrm{DCR}} = &\frac{1}{2}\mathrm{sim}\left(f_{\mathrm{DCR}}(D_0(x_1)), D_0(x_2), \Omega\right)\\ &+\frac{1}{2}\mathrm{sim}\left(f_{\mathrm{DCR}}(D_0(x_2)), D_0(x_1), \Omega\right),\end{aligned} \quad (6)$$

where $\mathrm{sim}(f_{\mathrm{DCR}}(D_0(x), D_0(x'), \Omega)$ is the negative cosine similarity between two feature maps in the overlapping area Ω. The DCR enables the discriminator to focus on important area instead of the background.

The loss for the generator is

$$L(G_1, G_2) = L_{\mathrm{adv}} + \lambda_1 L_{\mathrm{TC}} + \lambda_2 L_{\mathrm{Identity}}, \quad (7)$$

which is the same as the original TC-ShadowGAN. λ_1 and λ_2 are set to 40 and 5 for each, as in the original TC-ShadowGAN[1]. The total loss of the discriminator in our proposed model is

$$L_{\mathrm{total}} = L_{\mathrm{adv}} + \lambda_{\mathrm{BCR}} L_{\mathrm{BCR}} + \lambda_{\mathrm{DCR}} L_{\mathrm{DCR}} \quad (8)$$

where λ_{BCR} and λ_{DCR} are weights for regularization terms. $\lambda_{\mathrm{BCR}} = 10$ is experimentally found to be optimal when only BCR is introduced[2].

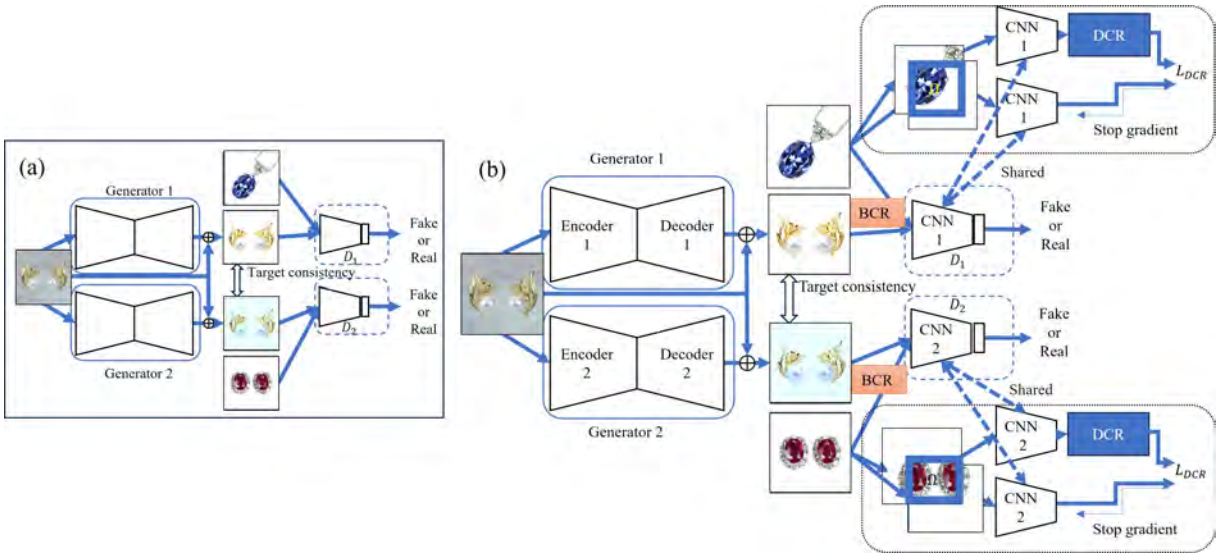


Fig. 4 Network structure: (a) Original TC-ShadowGAN (b) Proposed method

Fig. 5 Test images with strong backgrounds

4. Experiments

4.1 Dataset and method

The training dataset used in this study consists of 2,000 photographed images of jewelry and 2,000 ground truth images which were corrected by an experienced photographer using image editing software. The test dataset includes 63 images with various types of background elements, such as strong shadows, plastic plates, glass stands and paper plates in **Fig. 5**. It includes images of intricately shaped jewelry that is difficult to distinguish from the background. In all the following experiments, the learning rate is set to 0.002, the batch size to 1, and the number of epochs to 200. The input image size is 256 × 256 pixels. When introducing BCR, random crop and flip are imposed as data augmentations with $\lambda_{\mathrm{BCR}} = 10$. When introducing DCR, the patch size should be at least 0.7 times larger than the original image size and resized to 128 x 128 pixels. Color distortion, which changes brightness, saturation, hue, and contrast with a probability of 80% and grayscale with a probability of 20%, is applied to patches randomly. The fixed threshold τ that determines the overlapping area Ω in the patches is set to 0.5. When DCR and BCR are simultaneously introduced, we set the weights of losses with $\lambda_{BCR} = 5$ and $\lambda_{DCR} = 0.1$. Other parameter settings are devoted to ablation studies in Section 4.3

We use four metrics to measure the similarities between translated images and ground truth images: Peak Signal-to-Noise Ratio (PSNR) , Structural Similarity Index Measure (SSIM)[13], average of the color difference per pixel in LAB color space (LAB), and Fréchet Inception Distance (FID)[12]. In addition, in order to focus on the correction of the foreground jewelry part, background removal is performed on the corrected image to obtain an evaluation index.

4.2 Main results

Figure 6 shows the results of our proposed method, where we use TC-ShadowGAN as a base model and regularization methods of BCR and DCR are combined. We can observe that our method can remove the background more exquisitely. For example, in the ribbon pendant head, the foreground and background borders are correctly separated. In the ruby earrings, it can be seen that the two are separated only by the proposed method. In the case of the pearl earrings, CycleGAN fails to separate the background from the foreground, resulting in whitening of the pearls, but the proposed method alleviates this problem. It also improves the removal rate of black area in the corners.

Table 1 presents a quantitative evaluation using PSNR, SSIM, LAB, and FID. In the TC-ShadowGAN-based models, there are two values obtained from two generators for each metric. However, these values are nearly identical because of target consistency. Our proposed method shows consistent improvements. Note that the background is removed before obtaining the metrics

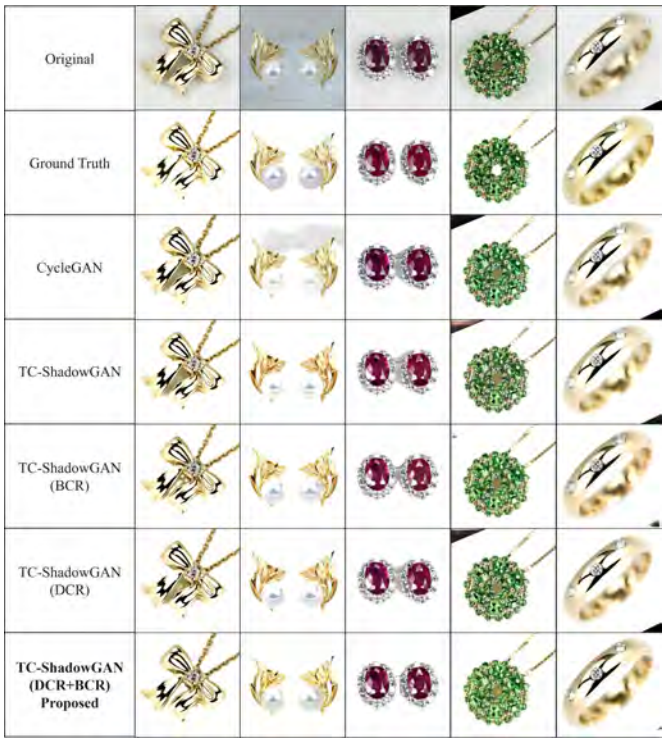


Fig. 6 Comparison of translated images

and the focus is only on the jewelry area. Compared to CycleGAN, TC-ShadowGAN with the introduction of BCR improves performance on all evaluation metrics. In addition, when DCR is introduced alone, the performance improvement is marginal compared to when DCR is not introduced, but when BCR and DCR are applied simultaneously, the performance improves.

In this experiment, we showed that the introduction of a combination of BCR and DCR into TC-ShadowGAN can remove background elements in jewelry images more powerfully, and can also correct images with fuzzy contours without halation. BCR by itself greatly improved the ability to remove background elements compared to the case where no BCR was introduced. On the other hand, the introduction of DCR improves the correction of the jewelry part and the ability to remove some background elements compared to the case where nothing is introduced. This may be due to the fact that DCR promotes attention to the foreground jewelry, and suppresses attention to the background during GAN learning.

4.3 Ablation studies

4.3.1 Comparison between TC-ShadowGAN and CycleGAN

In this Subsection, we compare the jewelry image correction by TC-ShadowGAN[1] and CycleGAN[8]. **Figure 7** shows that TC-ShadowGAN is superior to CycleGAN in correcting jewelry images with strong background elements. In particular, TC-ShadowGAN eliminates the problem pointed out in Fig. 2 that the jewelry area becomes too bright in CycleGAN. CycleGAN has to learn the inverse-transformation from the ground truth to the captured image without additional information such as background color, lightening conditions, shadow position, or color tone.

Table 1 Quantitative evaluation of methods

Algorithm	PSNR ↑	SSIM ↑	LAB ↓	FID ↓
CycleGAN	21.36	0.8974	10.81	58.63
TC-ShadowGAN	21.67	0.8967	9.716	55.44
	21.66	0.8968	9.730	55.96
TC-ShadowGAN	**22.51**	**0.9040**	**8.794**	**52.37**
BCR	22.47	**0.9035**	8.885	**52.45**
TC-ShadowGAN	21.68	0.8964	9.823	54.97
DCR	21.68	0.8963	9.811	54.15
TC-ShadowGAN	**22.68**	**0.9031**	**8.768**	**52.81**
DCR+BCR	**22.67**	**0.9029**	**8.755**	53.02

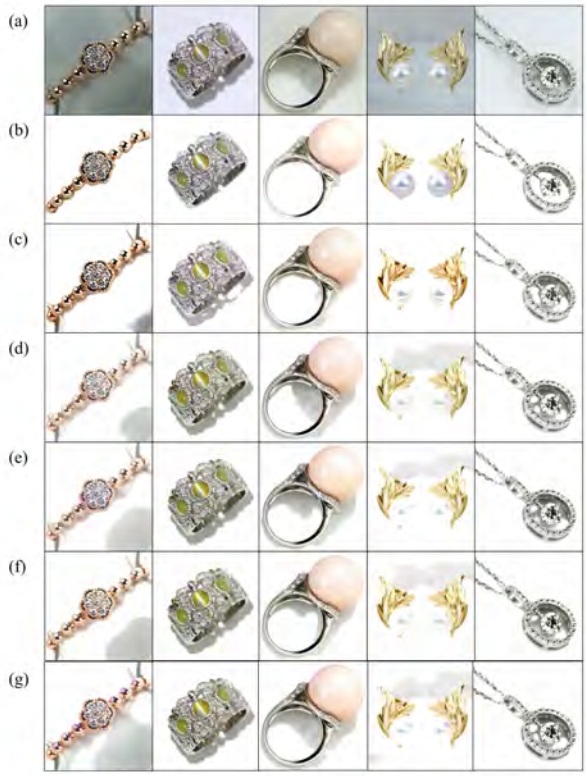


Fig. 7 Comparison between TC-ShadowGAN and CycleGAN : (a) Original image, (b) Ground truth (c) TC-ShadowGAN (d) CycleGAN (e) CycleGAN+BCR (f) CycleGAN+DCR (g) CycleGAN+BCR+DCR

We also apply BCR and DCR to CycleGAN. The performance of TC-ShadowGAN is boosted by introducing BCR and DCR. On the other hand, in the case of CycleGAN the performance is not improved or even becomes worse as shown in **Table 2** and **Fig. 7**. This result suggests that the introduction of BCR and DCR may not be suitable for GANs such as CycleGAN, where learning is

Table 2 TC-ShadowGAN and CycleGAN

Algorithm	PSNR↑	SSIM↑	LAB↓	FID↓
TC-ShadowGAN	**21.67**	0.8967	**9.716**	**55.44**
	21.66	0.8968	9.730	55.96
CycleGAN	21.36	**0.8974**	10.81	58.63
CycleGAN (BCR)	20.42	0.8837	13.11	64.15
CycleGAN (DCR)	21.18	0.8953	11.32	60.18
CycleGAN (BCR+DCR)	21.10	0.8934	11.93	63.12

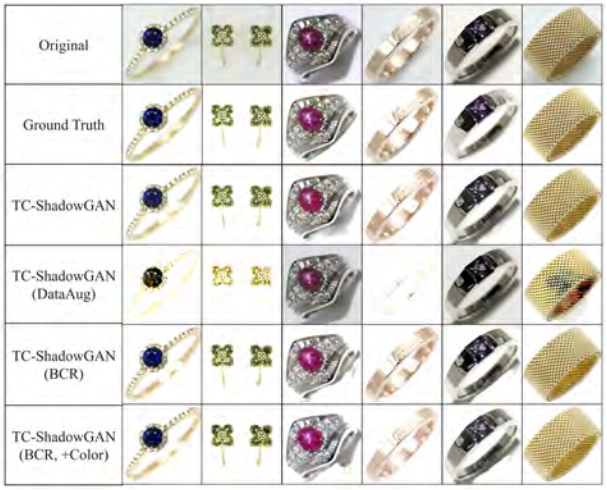


Fig. 8 Effect of BCR and type of data augmentation on corrected images

performed in a bidirectional transformation. This is presumably because the translation from ground truth to the original image, which has no correct answer, emphasizes superfluous parts such as shadows by BCR and DCR.

4.3.2 Effect of data augmentation and BCR

We have investigated many types of data augmentations and found that the combination of the random crop and flip, which is used in the main results, achieves the best performance. In this section, we show the results where other types of image transformations are applied for data augmentation in the original BCR[2] and SimCLR[15]. Now we add color distortion such as brightness, saturation, contrast and hue, each of which is set to 0.5 in torchvision.transforms.ColorJitter. As a control experiment, we add color distortion as data augmentation without introducing BCR.

Figure 8 shows the effect of BCR on the corrected images. The jewelry correction performance is extremely poor when the data augmentation is applied to only real images without BCR (DataAug). When data augmentation is introduced only for real images in GAN, the discriminator learns the augmented images as part of the real images, and thus learns the distribution of the real images incorrectly. As a result, the generator learns to produce images affected by advanced and complex data augmentations such as cut out and color distortion. Therefore, random crop and flip are often used together in conventional GANs.

Table 3 Effect of data augmentation in BCR

Algorithm	PSNR↑	SSIM↑	LAB↓	FID↓
TC-ShadowGAN	**22.51**	**0.9040**	**8.794**	**52.37**
(BCR)	22.47	0.9035	8.885	52.45
TC-ShadowGAN	17.68	0.8368	17.15	137.3
(DataAug)	17.69	0.8368	17.22	136.0
TC-ShadowGAN	22.26	0.9009	9.218	58.15
(BCR, +Color)	22.25	0.9014	9.225	57.65

Consistency Regularization (CR) and its extension BCR solve this problem. CR is a regularization method that inputs real images with and without data augmentation to the discriminator and ensures consistency in the output. CR is defined as

$$L_{CR} = ||D(x_{\text{real}}) - D(t_1(x_{\text{real}}))||^2 \tag{9}$$

where x_{real} is the real image, and $t_1(\cdot)$ is the data augmentation.

Adding color distortion to the random crop and flip results in a slight drop in performance (BCR, +Color). The background is not removed sufficiently, or the jewelry is too bright. **Table 3** shows quantitative results which are consistent with the qualitative results in Fig. 8.

4.3.3 Hyperparameter for DCR

The DCR hyperparameters were determined by experimenting with various parameters using a brute-force method, focusing on the values that were considered optimal in the original paper[3]. When introducing DCR, the patch size should be at least 0.7 times larger than the original image size[3]and resized to 128 x 128 pixels. For each patch, color distortion, which changes brightness, saturation, hue, and contrast with a probability of 80% and grayscale with a probability of 20%, is applied. The fixed threshold τ that determines the overlapping area Ω in the two patches is set to two values, 0.5 and 0.7[3]. Applying DCR is effective as shown in **Fig. 9**. $\tau = 0.5$ is more suitable for removing background areas than $\tau = 0.7$ in the image of the blue earrings. However, $\tau = 0.7$ is more suitable for distinguishing foreground and background within the jewelry area as shown in the rightmost image.

We have also examined the effect of other types of data augmentation. Four types of data augmentation are added to the original data augmentation of color distortion to accelerate the learning of representations of shape: equalization, solarization, sharpness, and Sobel filter.

Quantitative evaluation of effect of DCR is shown in

Fig. 9 Dependency of parameter $\tau(= 0.5, 0.7)$ of DCR on correction results

Table 4 Hyperparameters for DCR

Algorithm	PSNR	SSIM	LAB	FID
TC-ShadowGAN	21.67	0.8967	9.716	55.44
	21.66	0.8968	9.730	55.96
TC-ShadowGAN	21.6	0.8964	9.823	54.97
(DCR, 0.5)	21.68	0.8963	9.811	54.15
TC-ShadowGAN	21.68	0.8959	9.723	55.65
(DCR, 0.7)	21.67	0.8956	9.726	55.31
TC-ShadowGAN	21.79	0.8983	9.47	55.02
(DCR, 0.5, +)	21.80	0.8984	9.455	54.40
TC-ShadowGAN	**21.93**	**0.8988**	**9.39**	**54.22**
(DCR, 0.7, +)	21.91	0.8986	9.428	55.04

Table 4. The introduction of DCR to image-to-image translation model is effective to correct jewelry images. In particular, data augmentations that make the shape of jewelry clearer, such as equalize, are effective in correcting jewelry areas in the images (DCR, $\tau = 0.7$, +). This may be the result of suppressing easily learnable elements such as color and promoting learning of less easily learnable elements such as shape.

4.3.4 Hyperparameter for simultaneous application of BCR and DCR

Zhao et al.[2]combined BCR and contrastive loss, introduced in SimCLR[15]of self-supervised learning. Their experimental results showed that the optimal parameters for combining losses were $\lambda_{\mathrm{BCR}} = 5, \lambda_{\mathrm{cntr}} = 0.1$. In this paper, we have introduced DCR instead of contrastive loss with (a) $\lambda_{\mathrm{BCR}} = 5, \lambda_{\mathrm{DCR}} = 0.1$.

We have also tried other parameters, for example, (b) $\lambda_{\mathrm{BCR}} = 10, \lambda_{\mathrm{DCR}} = 1$. Qualitative results show that $\lambda_{\mathrm{BCR}} = 5, \lambda_{\mathrm{DCR}} = 0.1$ with $\tau = 0.7$ for DCR is better than other combination of parameters in **Fig. 10**. Quantitative evaluation, which focuses on jewelry part, also

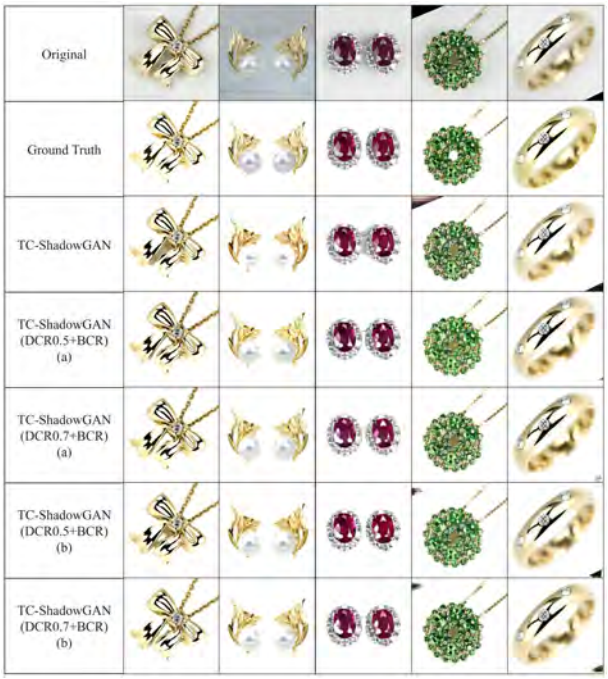


Fig. 10 Hyperparameters for combining BCR and DCR: (a) $\lambda_{\mathrm{BCR}} = 5, \lambda_{\mathrm{DCR}} = 0.1$, (b) $\lambda_{\mathrm{BCR}} = 10, \lambda_{\mathrm{DCR}} = 1$

Table 5 Hyperparameters for combining BCR and DCR: (a) $\lambda_{\mathrm{BCR}} = 5, \lambda_{\mathrm{DCR}} = 0.1$, (b) $\lambda_{\mathrm{BCR}} = 10, \lambda_{\mathrm{DCR}} = 1$

Algorithm	PSNR ↑	SSIM ↑	LAB ↓	FID ↓
(a) TC-ShadowGAN	**22.68**	**0.9031**	8.768	52.81
DCR0.5+BCR	22.67	0.9029	8.755	53.02
TC-ShadowGAN	22.59	0.9027	**8.737**	55.49
DCR0.7+BCR	22.58	0.9028	8.765	54.39
(b) TC-ShadowGAN	22.37	0.9016	9.233	52.62
DCR0.5+BCR	22.35	0.9015	9.260	51.95
TC-ShadowGAN	22.28	0.9004	9.294	52.46
DCR0.7+BCR	22.25	0.9003	9.3379	**51.84**

support that $\lambda_{\mathrm{BCR}} = 5, \lambda_{\mathrm{DCR}} = 0.1$ is better as shown in **Table 5**.

5. Conclusions

Our contributions in this paper are as follows. First, we have shown that TC-ShadowGAN is effective not only in removing backgrounds but also in correcting jewelry areas in images with strong background elements because it avoids the inverse transformation from the ground truth image to the original image by applying the target consistency. Second, introducing BCR improves the removal of background elements. The reason for this improvement is to prevent the discriminator from incorrectly learning augmented jewelry images as real jewelry images, as in conventional GANs. Third, the introduction of DCR improves the ability to correct the jewelry areas because it concentrates the GAN on the jewelry area and sup-

presses attention to the background. Lastly, together with previous studies[10], we have shown that our proposed method outperforms state-of-the-art methods in jewelry image correction. Our proposed method does not assume jewelry-specific properties and is not limited to jewelry data. Whether the method is effective for any other image groups is a subject for further study.

A possible future challenge is to introduce Adaptive Discriminator Augmentation (ADA)[16]instead of BCR. When training GANs with the original datasets, small dataset size causes overfitting of the discriminator, which in turn affects the performance of the generator. ADA adaptively controls the data augmentation of the input to the discriminator depending on the overfittying state of the model. The powerful diffusion models that have been developed in recent years should also be considered for future work.

Acknowledgement

This research was funded by Selby Corporation. We would like to take this opportunity to thank Mr. Toshiya Matsutani, Mr. Kozo Saito, Mr. Nobuyasu Suzuki, and Mr. Shinya Yamamoto for providing their jewelry image datasets and expertise in jewelry imaging. We would also like to thank Mr. Deepak Rai, Mr. Tomohiro Shizuno and Mr. Daiki Ishiguro for fruitful discussions.

References

1) C. Tan, X. Feng, B, Chen, J. Long: "TC-ShadowGAN: A Target-Consistency Generative Adversarial Network for Unpaired Shadow Removal", 2022 IEEE International Conference on Multimedia and Expo (ICME), pp.1–6 (2022).
2) Z. Zhao, Z. Zhang, T. Chen, S. Singh, H. Zhang.: "Image Augmentations for GAN Training", arXiv:2006.02595 (2020).
3) M. Ko, E. Cha, S. Suh, H. Lee, J-J. Han, J. Shin, B. Han: "Self-Supervised Dense Consistency Regularization for Image-to-Image Translation", 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.18280-18289 (2022).
4) R. Liu, L. Ma, J. Zhang, X. Fan, Z. Luo: "Retinex-Inspired Unrolling with Cooperative Prior Architecture Search for Low-Light Image Enhancement", 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.10557–10565 (2021).
5) F. Minhao, W. Wenjing, Y. Wenhan, and L. Jiaying.: "Integrating Semantic Segmentation and Retinex Model for Low-Light Image Enhancement", Proc. of the 28th ACM International Conference on Multimedia, pp.2317–2325 (2020).
6) Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, Z. Wang: "EnlightenGAN: Deep Light Enhancement Without Paired Supervision", IEEE Trans. Image Processing, Vol. 30, pp. 2340–2349 (2021).
7) C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, R. Cong: "Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement", 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1777–1786 (2020).
8) J-Y. Zhu, T. Park, P. Isola, A.A. Efros: "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks", 2017 IEEE International Conference on Computer Vision, pp.2242–2261 (2017).
9) T. Park, A.A. Efros, R. Zhang, J-Y. Zhu: "Contrastive Learning for Unpaired Image-to-Image Translation", Computer Vision — ECCV 2020, Lecture Notes in Computer Science, Vol.12354, pp.319–345 (2020).
10) T. Shizuno, T. Nakagawa, D. Ishiguro, D. Hosokawa, Y. Oshigane, T. Ozeki: "Image Enhancement of Jewelry Data Using CycleGAN", Proc. of the 36th Annual Conference of the Japanese Society for Artificial Intelligence (2022).
11) T. Nakawaga, T. Shizuno, D. Ishiguro, T. Ozeki: "Effect of Background on Image Enhancement of Jewelry Data Using CycleGAN", Proc. of FIT2022 (2022).
12) M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter: "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium", Advances in Neural Information Processing Systems 30, pp.6626–6637 (2017).
13) Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli: "Image Quality Assessment: From Error Visibility to Structural Similarity", IEEE Transactions on Image Processing, Vol. 13, No. 4, pp. 600-612 (2004).
14) C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna: "Rethinking the Inception Architecture for Computer Vision", arXiv:1512.00567 (2015).
15) T. Chen, S. Kornblith, M. Norouzi, G. Hinton: "A Simple Framework for Contrastive Learning of Visual Representations", arXiv:2002.05709 (2020).
16) T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, T. Aila: "Training Generative Adversarial Networks with Limited Data", Advances in Neural Information Processing Systems 34, pp.12104–12114 (2020).

(Received November 8, 2023)
(Revised April 6, 2024)

Taiyo NAKAGAWA
He was an undergraduate student at Tokai University when this research was conducted. His specialty was aritificial intelligence and machine learning, with a particular focus on image restoration. He currently works for a private company in Japan.

Tomoko OZEKI (*Member*)
She received the Dr. of Science degree from the Graduate School of Tokyo Institute of Technology, Tokyo, Japan, majoring in physics. She worked as a Researcher at RIKEN Brain Science Institute, Saitama, Japan, and currently works as a Professor at Tokai University. She has been engaged in research in areas of machine learning, brain science, information science, and statistical mechanics. In particular, she has devoted herself to the dynamics of neural networks and the dynamics of information processing, including associative memory, supervised learning, image restoration, and information geometry of learning machines.

IIEEJ **Paper**

SF-Net: Simultaneous Fusion Network for Semantic Segmentation and Depth Estimation

Kai WANG† (*Student Member*), Takayuki NAKAMURA†

† Graduate School of System Engineering, Wakayama University

<**Summary**> Semantic segmentation is an important technique in various applications, such as autonomous driving, medical imaging, and industrial inspection. Depth estimation, as one of the important components of scene understanding, can be used to obtain effective depth information while utilizing only RGB images. In recent years, such depth information has been used as an auxiliary feature to facilitate the semantic segmentation task. This study proposes a Simultaneous Fusion Network(SF-Net) that simultaneously learns semantic segmentation and depth estimation tasks based on a monocular camera image. The features are first extracted and strengthened by injecting contextual information using semantic labels through the feature reinforcement module and then learned simultaneously by analyzing the imaging process to establish the relationship between the size and depth of the objects in the image. A new loss function is represented by the geometric relationship. Furthermore, a feature fusion module is constructed to perform image feature fusion on the common parts of depth estimation and semantic segmentation tasks. By learning simultaneously, the accuracy of semantic segmentation can be improved by utilizing the depth information obtained from depth estimation inference. We conducted experiments using the Cityscapes dataset and the NYUDv2 dataset and verified the effectiveness of the proposed method.

Keywords: semantic segmentation, depth estimation, simultaneous learning

1. Introduction

Semantic segmentation, which aims to predict the class label of each pixel in an image, plays a vital role in various applications such as autonomous driving, medical imaging, and industrial inspection. In recent years, the methods using deep learning have shown excellent performance in various segmentation tasks such as scene understanding, medical images, and anomaly detection, but the segmentation results obtained by learning using only RGB images often have limitations. In comparison, multimodal data can provide more spatial and contextual information for accurate scene understanding, where depth maps are often used as complementary information to RGB images to improve segmentation accuracy[1)–6)]. The introduction of depth information facilitates the solution of semantic segmentation problems. It allows higher accuracy and robustness to be achieved in complex scenes. For example, there is usually a depth difference between the target object and the surrounding background or objects, and obtaining an accurate depth map helps to understand the relationship between the position of each object in front of it and the position behind it. Thus, depth information can help improve segmentation performance.

The acquisition of depth information based on visual information is more challenging than that of depth information using an active sensor such as LiDAR. Depth estimation[5),7)–11)] is also one of the important methods for scene understanding, along with semantic segmentation methods[12)–15)]. However, as analyzed by He et al.[16)], depth estimation based on visual information is ambiguous in some scenes. To improve the accuracy of monocular depth estimation, we should eliminate ambiguity as much as possible in the estimation process. As one such method for eliminating ambiguity in the depth estimation process, it is considered to be promising to utilize different information used in the semantic segmentation task. Simultaneously performing these two tasks to improve both the accuracy of depth estimation and semantic segmentation becomes an attractive direction.

Many deep learning-based fusion methods[17)–20)]aim to perform image feature fusion by skillfully designing the network structure. Though the geometric relationship between the physical size and the depth of an object in the image is considered to be useful during the fusion process, it has not been effectively utilized. He et al.[21)]effectively utilize the geometric relationship between the physical size and the depth of an object in the image. However, generally speaking, it is difficult to determine the actual physical size of each object in the image. Considering the geometric constraints between the two has certain limitations.

In this paper, we propose a simultaneous learning method for conducting depth estimation and semantic segmentation

tasks at the same time, and we also propose a novel one-stage neural network architecture for simultaneous learning. Using the perspective projection model, a zoom coefficient is proposed based on the relationship between the size of the segmented region and the depth value, and it is used to propose a new loss function that can make it stabilized to evaluate the quality of the depth estimation.

To effectively utilize the context of semantic information in an image, we design a dual attention network so that it can make more accurate predictions based on relevant feature maps. In addition, we also designed a feature fusion module to enhance the consistent features of the two tasks and thereby obtain their common feature attention maps. To summarize, the main contributions of this paper are as follows:

1. We propose a Simultaneous Fusion Network called SF-Net that simultaneously learns semantic segmentation and monocular depth estimation tasks.
2. We propose a new zoom loss function that can make it stabilized to assess the quality of depth estimation.
3. We propose an efficient feature fusion module called FFM to improve semantic segmentation performance by sharing features of the two tasks.
4. Our one-stage model achieves competitive results with other depth estimation and semantic segmentation methods on the two popular datasets. The effectiveness of our approach is demonstrated.

2. Related Works

2.1 Semantic segmentation and depth estimation

Many deep learning-based methods have been developed to solve semantic segmentation. FCN[12] proposed an end-to-end fully convolutional neural network architecture that enables pixel-level semantic segmentation of an input image of arbitrary size by replacing fully connected layers with fully convolutional layers. Inspired by this, U-Net[22] and RefineNet[13] had an encoder-decoder network architecture for a good fusion of low-level and high-level semantic information. PSP-Net[23] and DeepLabV3+[24] proposed the atrous spatial pyramid pooling(ASPP) so that it can capture global information using multi-scale information. DANet[25] and CBMA[26] had a dual-attention mechanism to enhance the representation of image features.

For the problem of learning depth from a monocular image, Make3D[27] introduced strong geometric assumptions about the scene structure and manually represented them using Markov random fields (MRF). Deep neural network-based methods have recently made great progress in monocular depth estimation tasks. Eigen et al.[7] proposed two network modules for coarse-grained global prediction and fine-grained local fine-tuning, respectively. Liu et al.[5] and Li et al.[8] proposed combining convolutional neural networks (CNNs) and conditional random fields (CRFs) to enhance the model's understanding of global contextual information. DORN[11] proposed to treat the depth estimation problem as a classification problem that preserves ordering information between categories, rather than a complex continuous value prediction. MonoDepth[10] introduced a supervised learning method that used images with depth information as labels to correspond to input images. These two methods used only unimodal information. To get higher-precision depth information, it remains a worthwhile challenge to utilize other information except that used in the depth estimation process effectively.

2.2 Multitask learning

Recently, multitask learning methods have improved the performance of various computer vision problems. Several deep learning networks using a multitask learning framework have been proposed to perform semantic segmentation and depth estimation simultaneously. Eigen et al.[28] built a network architecture containing three scales from coarse to fine to make predictions for the depth values, surface normals, and semantic labels, simultaneously.

PAD-Net[3] proposed to facilitate semantic segmentation and depth estimation using additional tasks that provide rich information for the two original tasks. CI-Net[20] introduced an attention module to enhance scene understanding and to obtain inter- and intra-class correlations. The semantic labels of the input image are used to generate an attention map to determine whether pixels belong to the same class. This allows the model to better understand the scene for subsequent predictions. It is evident that attentional mechanisms can actively contribute to multitask learning.

Some of these methods benefited from the two-stage learning strategy and achieved good results. However, they incur additional computational costs making it difficult to adapt to real-world applications. In addition, most related works focus on optimizing the shared features of the two tasks and do not consider using geometric constraints to strengthen the link between semantics and depth information. SOSD-Net[21] used a perspective projection model to find the relationship between the size of an object in an image and its depth value, called it "semantic-objectness." However, to represent this relationship, it is necessary to measure the object's actual size in the image. Generally, it is difficult to accurately measure such value.

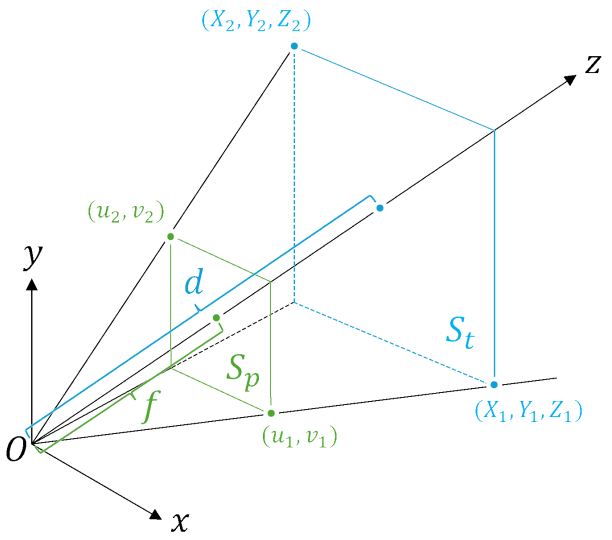


Fig. 1 Perspective projection model

3. Our Method

3.1 Relationship between the size of a segmented region and its depth value

This section describes our SF-Net for simultaneous learning of semantic segmentation and depth estimation. We first introduce a perspective projection model to represent the relationship between the size of the segmented region and the depth value, and then we describe in detail the network structure of our SF-Net and the proposed feature fusion module. Finally, our loss function for training our SF-Net is described.

The depth value for each pixel estimated from a single image is usually uncertain and changes rapidly due to some errors during depth estimation process. Such unstable changes in depth values negatively affect learning process in neural networks. Therefore, to find a metric instead of a depth value, we explore the relationship between the objects' actual size and its depth value. Since it is often difficult to accurately measure the actual size of an object in an image, we consider the connection between the two through a perspective projection model while excluding the inaccessible parameters as much as possible.

As shown in **Fig. 1**, d denotes the depth value, f denotes the focal length, S_t and S_p denote the actual area of the object and the area on the image, respectively. The ratio I of S_t and S_p can be expressed as follows:

$$I = \frac{S_t}{S_p}, \tag{1}$$

Since S_t is a fixed value, S_p changes depending on the depth value d. Therefore, when S_p becomes smaller, the ratio I becomes larger, and the corresponding d also becomes larger. Conversely, when S_p increases, both I and d decrease. This relationship between I and d can be said to be linear and can be expressed using a scale factor a as follows:

$$d = aI, a > 0, \tag{2}$$

Hereafter, we call this scale factor the "zoom factor."

Substituting the Eq. (1) into the Eq. (2) yields the following equation:

$$a = \frac{dS_p}{S_t}. \tag{3}$$

Moreover, using the projection perspective model illustrated in Fig. 1, we obtain the following equations:

$$\begin{aligned} &\Delta u = u_1 - u_2, \Delta v = v_1 - v_2, \\ &\Delta X = X_1 - X_2, \Delta Y = Y_1 - Y_2, \end{aligned} \tag{4}$$

where u_* and v_* represent the horizontal and vertical coordinates of the image in Fig. 1, and X_* and Y_* represent the horizontal and vertical coordinates of the corresponding object in Fig. 1. $\Delta*$ represents the difference between the 2 coordinate values, which are the width and height of the image and its corresponding object in the 2D plane. Then based on the proportionality, we can get the following formula:

$$\Delta u = f\frac{\Delta X}{d}, \Delta v = f\frac{\Delta Y}{d}, \tag{5}$$

Multiplying the above two gives:

$$d^2 = \frac{f^2 \Delta X \Delta Y}{\Delta u \Delta v}, \tag{6}$$

S_t and S_p is equal to $\Delta X \Delta Y$ and $\Delta u \Delta v$, respectively. Substituting these equations into Eq. (6) yields the following equation:

$$d^2 = \frac{f^2 S_t}{S_p}, \tag{7}$$

After making adjustments, the following equation is obtained:

$$\frac{S_p}{S_t} = \frac{f^2}{d^2}, \tag{8}$$

By combining Eq. (3) and Eq. (8), the zoom factor a can be obtained as Eq. (9).

$$a = \frac{df^2}{d^2} = \frac{f^2}{d}, \tag{9}$$

In Eq. (9), there is no area value, only the camera focal length and depth value. The focal length f of the camera is a fixed value, and the depth value d is a known value during training.

Depth values in the range $-1.0 < d < 1.0$ may have a detrimental effect to a zoom factor because the values of the inversely proportional equation change rapidly in this range. However, in the range $d > 1.0$, this change can not happen. In the range $d > 1.0$, even if d changes rapidly, the zoom

factor a does not change rapidly because there is an inversely proportional relationship between d and a.

Taking advantage of this characteristic, we propose to use the zoom factor a in the loss function for the depth estimation task instead of using only the depth value. Thus, drastic changes in the loss function due to sudden changes in the depth value can be avoided during the depth estimation process, and the quality of depth estimation can be evaluated stably. Finally, we describe the loss function for the depth estimation task using the function f_{Zoom} expressed by the following equation:

$$f_{Zoom}(d, \hat{d}) = |a - \hat{a}| = \left| \frac{f^2}{d} - \frac{f^2}{\hat{d}} \right| = f^2 \left| \frac{\hat{d} - d}{d\hat{d}} \right|, \quad (10)$$

where $\hat{d}$ is the predicted depth value and $\hat{a}$ is the predicted zoom factor obtained from $\hat{d}$. The detail of the total loss function is described in Section 3.3.

3.2 Network architecture

As shown in **Fig. 2**, our SF-Net consists of three parts (encoder part, feature enhancement part, and decoder part). The encoder backbone acquires the features of an image, which are enhanced by two feature enhancement modules. Furthermore, low-level features are acquired at the encoding phase and fed into the decoder layers by Encoder Feature Forwarding. Then, multi-task feature fusion is performed through the Feature Fusion Module (FFM) in the decoder phase. Finally, the semantic segmentation and depth estimation predictions are obtained in the two parallel decoding branches, respectively. The following paragraphs describe each part of our SF-Net in detail.

Encoder part We use the encoding structure of ResNet as a backbone for generating the initial feature maps. Because we need to enhance features by further utilizing category labels, we do not use the original decoding structure of ResNet. Since the convolutional layer in ResNet has a small sensory field for an extensive range of semantic information, it may lead to an insufficient understanding of the global context by the network. In addition, as the network deepens, the spatial resolution of the feature maps decreases, which may lead to difficulty in capturing the target boundaries.

Feature enhancement part The generated feature map is input to the enhancement part. It is built with an atrous spatial pyramid pooling (ASPP) and attention modules. ASPP can perform atrous convolution layer of feature maps at different sampling rates to obtain multi-scale contextual information. It uses different expansion rates to capture sensory fields at different scales to improve the network's perception of targets at different scales. With ASPP, the network can better understand the global contextual information in the image, which helps to improve the performance. The attention mechanism allows the network to assign different weights to information at different locations when processing features. This helps the network to better focus on the target region in tasks such as semantic segmentation and reduces the sensitivity to irrelevant information. As a result, this helps to determine the target boundary.

ASPP is used in Deeplabv3+, including a 1×1 convolution layer, three atrous convolution layers with different rates, and a global average pooling layer to integrate multi-scale information. Then, the output of the ASPP module is concatenated and fed into a 1×1 convolution layer(red block) to generate the final feature map. The construction of the Attention Module is inspired by DANet[25] and CBMA[26]. It includes two attention parts, which are channel attention and self-attention. As shown in **Fig. 3**, the channel attention and the self-attention module are connected in series to enhance the expressive ability of our model.

The input-output relationship for the channel attention module can be expressed as follows:

$$\begin{aligned} F_c = \sigma(&MLP(AvgPool(F_i)) \\ &+ MLP(MaxPool(F_i))) \otimes F_i, \end{aligned} \quad (11)$$

Here, $F_i \in \mathbb{R}^{C \times H \times W}$ is the input feature map, where C, H, and W is number of channels, the height, and the width of the feature map, respectively. $AvgPool$ and $MaxPool$ represent the average pooling and maximum pooling operations, respectively, and MLP represents the multi-layer perceptron, including two fully connected layers and a Relu activation function. σ denotes the sigmoid function, and $\otimes$ denotes the element-wise multiplication operation.

Then, the input-output relationship for the self-attention module can be expressed as follows:

$$F_{self} = \sigma(\frac{QK^T}{\sqrt{C_k}})V, \quad (12)$$

Q, K, and V are the query keys and values of the self-attention module, respectively. Here, we take a linear transformation of $F_c \in \mathbb{R}^{C \times M}$ using the parameter matrices $W_q \in \mathbb{R}^{C_q \times C}$, $W_k \in \mathbb{R}^{C_k \times C}$, $W_v \in \mathbb{R}^{C_v \times C}$ gives $Q = W_q F_c \in \mathbb{R}^{C_q \times M}$, $K = W_k F_c \in \mathbb{R}^{C_k \times M}$, and $V = W_v F_c \in \mathbb{R}^{C_v \times M}$. $C_{k,v,q}$ is the dimension of keys, respectively, and $M = H \times W$. The output feature F_{self} is computed by multiplying V with the attention map.

Decoder part As shown in **Fig. 4**, to perform both tasks simultaneously, we design two parallel branches for semantic segmentation and depth estimation, which are decoders for different tasks.

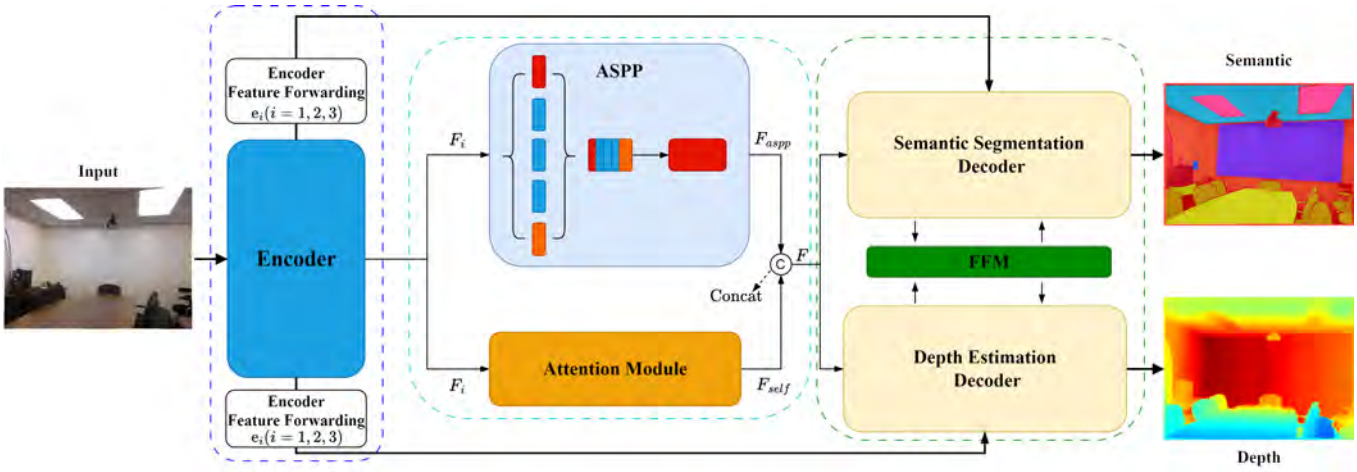


Fig. 2 The overview of our SF-Net architecture

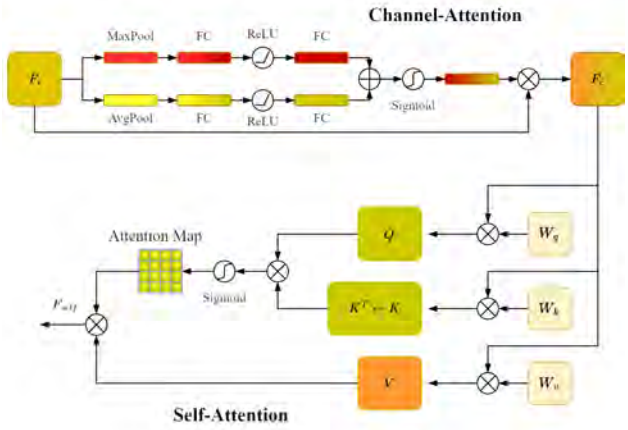


Fig. 3 Our attention module

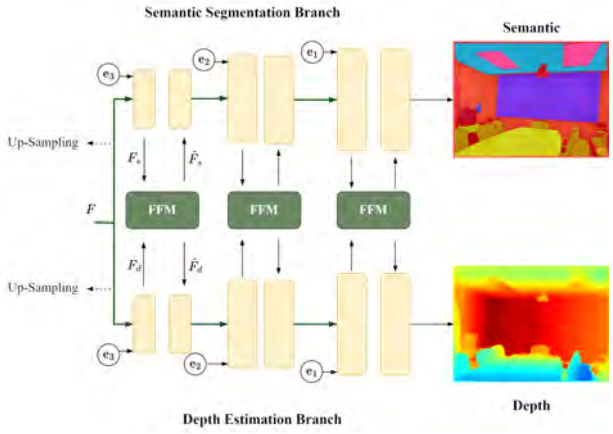


Fig. 4 Our two parallel decoding branches

Given that the feature map $F \in \mathbb{R}^{C\times H\times W}$ from the decoding phase after feature enhancement as input, the tasks of semantic segmentation and depth estimation modules in both branches are decoded at the feature layer. The general lack of low-level semantic information in high-level features makes the model prone to losing the basic geometrical structures, such as the edges of the target.

So, we add the feature $e_i(i = 1, 2, 3)$ from the encoder phase, which is used to fuse the low-level semantic information. Specifically, as the initial input to the decoding phase, the initial feature maps of the two tasks are added with the feature $e_i(i = 1, 2, 3) \in \mathbb{R}^{C\times H\times W}$ of the encoding phase to obtain the initial feature fusion of low-level semantic features, and can be expressed as $F_s = F_d = F + e_i$. The feature maps of the two tasks are then upsampled three times to restore the image to its original size. Finally, the semantic segmentation and depth estimation results are obtained, respectively.

To extract common and complementary features among different tasks, we propose a feature fusion module that enhances the feature representation capability in the model. The structure of the FFM is shown in **Fig. 5**. The FFM fuses the features extracted from different tasks to generate a shared weight map and then enhances both tasks to output the information containing the enhanced details. It focuses on common and complementary features of tasks and facilitates information transfer between tasks in a multi-task learning environment. The feature fusion process of FFM can be summarized as follows. Two feature maps are first fused and processed once with maximum pooling and upsampling.

$$F_{fuse} = Conv(concat(F_s, F_d)), \tag{13}$$

$$F_{up} = Upsampling(Conv(MaxPool(F_{fuse}))), \tag{14}$$

where $MaxPool$ and $UpSampling$ represent the maximum pooling and up-sampling operations. $Conv$ represents the convolution operation of 3×3. Then the fused feature weight maps are obtained:

$$F_{weight} = \sigma(Conv(F_{up})), \tag{15}$$

where σ represents the sigmoid function. Finally, the output feature maps of the 2 branches are then generated:

$$\hat{F}_s = F_s + F_s \otimes F_{weight}, \tag{16}$$

$$\hat{F_d} = F_d + F_d \otimes F_{weight}, \tag{17}$$

where $\otimes$ denotes element-wise multiplication.

In FFM, a shared weight map is calculated using the sigmoid function to fuse the features required for the two tasks. Finally, each feature map and shared weight map are integrated and output. The two output feature maps are upsampled separately and fed into the next level layer in FFM.

3.3 Loss function

The loss function of the proposed network is the sum of the loss functions of two branch tasks: a semantic segmentation task and a depth estimation task. The loss function for the semantic segmentation task uses a cross-entropy loss function. It can be described as follows:

$$\mathcal{L}_s = -\frac{1}{N}\sum_{n=1}^{N}\sum_{m=1}^{M} y_m^n \log \hat{y}_m^n, \tag{18}$$

Here, N is the number of pixels, M is the number of classes, y_m^n is the truth label of the mth class of the nth pixel, $\hat{y}_m^n$ is the predicted label of the mth class of the nth pixel.

In the depth estimation task, we must first consider the pixel-level error to minimize the difference between the prediction result and the true value. The most straightforward way to compute this error is to use the L2 norm. The loss function for depth estimation can be described as follows:

$$\mathcal{L}_d = \frac{1}{N}\sum_{n=1}^{N}(\hat{d}_n - d_n)^2, \tag{19}$$

Here,d_n is the truth depth value of the nth pixel, and $\hat{d}_n$ is the predicted depth value of the nth pixel.

However, the L2 norm is sensitive to outliers, and using the L2 norm may lead to unstable depth estimation due to the drastic changes in the depth values of individual pixels. Therefore, we propose a new loss function based on Section 3.1, which is more robust to outliers and can smooth out drastic changes in depth values due to outliers. It can be described as follows:

$$\mathcal{L}_{Zoom} = \frac{1}{N}\sum_{n=1}^{N}(f^2\left|\frac{\hat{d}_n - d_n}{d_n\hat{d}_n}\right|), \tag{20}$$

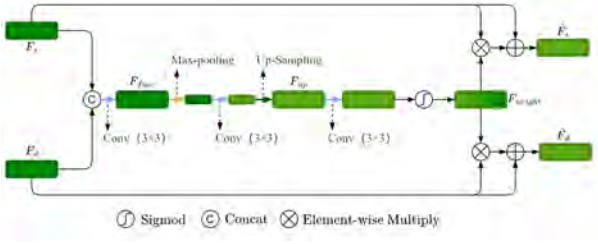


Fig. 5 Our feature fusion module(FFM)

In the actual computation, since the depth value of the background class is 0, it may lead to a situation where $d_n\hat{d}_n = 0$ during the calculation. This makes the function value explode and results in the inability to correctly calculate the scaling loss function for the corresponding pixel of the background class. Therefore, as shown in Eq. (21), a trick is used in the calculation process, which is not obtained by direct derivation but intentionally separates $\hat{d}_n - d_n$ and $d_n\hat{d}_n$ in a single calculation. In this way, even if the depth value of the background class is 0 in the calculation, it will only make a few items $d_n\hat{d}_n = 0$ and will not affect the overall value. The whole calculation process will also not be affected.

$$\mathcal{L}_{Zoom} = f^2\frac{\frac{1}{N}\sum_{n=1}^{N}\left|\hat{d}_n - d_n\right|}{\frac{1}{N}\sum_{n=1}^{N}\left|d_n\hat{d}_n\right|}, \tag{21}$$

where f is a fixed value of the camera's focal length.

Finally, the total loss function becomes Eq. (22) by combining the three different loss functions. Here, $\lambda_{1,2,3}$ is the weighting coefficient of each loss function, respectively.

$$\mathcal{L} = \lambda_1\mathcal{L}_s + \lambda_2\mathcal{L}_d + \lambda_3\mathcal{L}_{Zoom}. \tag{22}$$

4. Experiment

To verify the effectiveness of SF-Net, we conducted experiments using the publicly available datasets CityScapes[29] and NYUDv2[30].

4.1 Dataset and data augmentation

The CityScapes dataset is a large-scale dataset for urban scene understanding. It contains 5,000 high-quality pixel-level annotated images, of which 2,975 are used for training, 500 for validation, and 1,525 for testing. The training and test data in the dataset includes only the ground truth data for semantic segmentation. The dataset contains 19 classes. Following the data augmentation method of SOSD-Net[21], because of the computational overhead constraints, we resized the CityScapes image size to 256×512, and the training data are augmented on the fly during the training phase. The images are scaled with a randomly selected ratio among $\{0.5, 0.75, 1, 1.25, 1.5, 1.75\}$. In addition, images are also transformed using color transformations on HSV color space and flip with a chance of 0.5.

The NYUDv2 dataset is a dataset for indoor scene understanding. It contains 1,449 RGB-D images, of which 795 are used for training, and 654 for testing. The dataset contains 40 classes. Furthermore, the dataset includes the ground truth data for semantic segmentation and depth estimation. So, the proposed method is evaluated in terms of two aspects, which are semantic segmentation and depth estimation performance. Because of the computational overhead constraints, the input im-

ages are resized to 240×320. The training data are augmented on the fly during the training phase. The images are scaled with a randomly selected ratio among $\{1, 1.25, 1.5, 1.75\}$. In addition, the images are also transformed using color transformations on HSV color space and flip with a chance of 0.5.

4.2 Experiment details

Our method is implemented on the Ubuntu 22.04LTS platform using PyTorch. The computer environment used for the experiment is composed of an Intel(R) Core(TM) i9-9900K CPU@3.60GHz and an NVIDIA GeForce RTX2080ti graphics card. During the initialization stage, the weight layers in the first part of the architecture are initialized using the corresponding pre-trained model (ResNet-50) on the ImageNet classification task[31]. We use the Adam[32]optimizer and set the initial learning rate to $5e-4$. We set $\lambda_1 = 1, \lambda_2 = \lambda_3 = 2.5e-3$ to balance the weights between loss functions.

To evaluate the performance of semantic segmentation, we use mean intersection over union (mIoU), average accuracy (mPA), and pixel accuracy (Acc) as evaluation criteria. To evaluate the performance of depth estimation, we use the following metrics for evaluation:

- Average relative error (rel): It is the average of the absolute value of the relative error of the depth value of each pixel.

$$rel = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{d_i^* - d_i}{d_i}\right| \tag{23}$$

- Root mean square error (rms): It is the square root of the average of the square of the relative error of the depth value of each pixel.

$$rms = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(d_i^* - d_i\right)^2} \tag{24}$$

- Absolute log_{10} error (log_{10}): It is the absolute value of the relative error in the logarithm depth value of each pixel.

$$log_{10} = \frac{1}{N}\sum_{i=1}^{N}\left|\log_{10}(d_i^*) - \log_{10}(d_i)\right| \tag{25}$$

- Accuracy with threshold t: percentage (%) of d_i such that $\max(\frac{d_i}{d_i^*}, \frac{d_i^*}{d_i}) = \delta_t < 1.25^t, t = 1, 2, 3$

,where d_i is the predicted depth value of the ith pixel, d_i^* is the true depth value of the ith pixel, and N is the number of pixels.

4.3 Ablation experiment

We conducted ablation experiments with different conditions to validate the proposed model's effectiveness and the proposed method's modules. These experiments include semantic segmentation results and the fusion task of semantic segmentation and depth estimation.

Table 1 Semantic segmentation results of our proposed model based on CityScapes dataset

	Att	ASPP	$\mathcal{L}_d$	$\mathcal{L}_{Zoom}$	mIoU(%)	mPA	Acc
w/o FFM	✓				56.7	67.6	91.9
		✓			68.0	77.3	94.5
	✓	✓			69.1	77.6	94.8
w/ FFM	✓	✓	✓		70.9	79.6	95.2
	✓	✓	✓	✓	**71.3**	**79.7**	**95.3**

Table 2 Semantic segmentation performance comparison of multitasking algorithms based on CityScapes dataset

Methods	mIoU(%)
Kendall[33]	64.2
GradNorm[34]	64.8
Ozan[35]	66.6
ESOSD-Net[21]	68.2
SF-Net	**71.3**

The conditions for an ablation experiment are with/without the attention module(Att) alone, the ASPP module alone, and the combination of the attention module and ASPP. Furthermore, the conditions for an ablation experiment are with/without $\mathcal{L}_s + \mathcal{L}_d$ and $\mathcal{L}_s + \mathcal{L}_d + \mathcal{L}_{Zoom}$.

4.3.1 CityScapes experiment results

As shown in **Table 1**, derived from the validation set of the CityScapes dataset, the test set is not used because it does not include the ground truth data. Using two feature enhancement modules improves the accuracy of semantic segmentation by 12.4% over the attention module alone and 1.1% over the ASPP module alone in the cityscape dataset. With the addition of the depth estimation task, the accuracy of semantic segmentation is further improved. The accuracy was 70.9% when $\mathcal{L}_d$ was used alone and 71.3% when $\mathcal{L}_{Zoom}$ was also used, gaining a 2.2% improvement. This shows that the Zoom loss function proposed in this study can effectively improve the accuracy of semantic segmentation.

In addition, to verify the performance of the proposed model, as shown in **Table 2**, we compared ESOSD-Net[21], Kendall et al.'s method[33], and GradNorm[34]and Ozan et al.'s method[35]. The results of which are derived from the validation set of the CityScapes dataset, the test set is not used because the comparison methods use the validation set for performance evaluation. Most of the comparison methods in multi-tasks based on the Cityscapes dataset mainly apply the most important mIoU metrics for evaluation and do not provide the corresponding mPA and Acc metrics. So, to ensure objectivity, we similarly use only the mIoU metrics for comparison here. It can be seen that SF-Net improves the semantic segmentation results by about $3.1\% \sim 6.9\%$ compared to conventional methods. **Figure 6** shows the semantic segmentation results. In Fig. 6, the first row shows input images, the second shows the true value, and the third shows the prediction results.

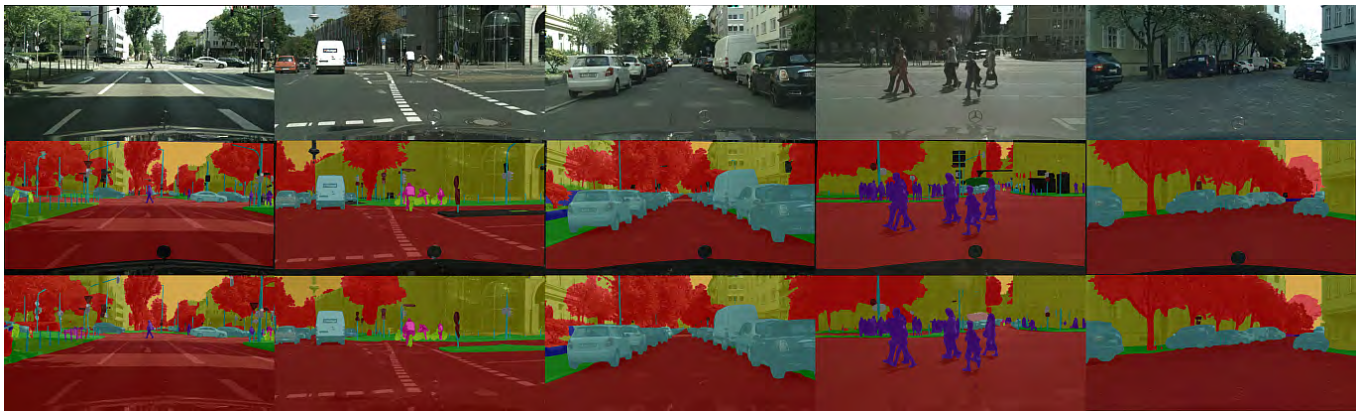


Fig. 6 Examples of segmentation results based on CityScapes dataset

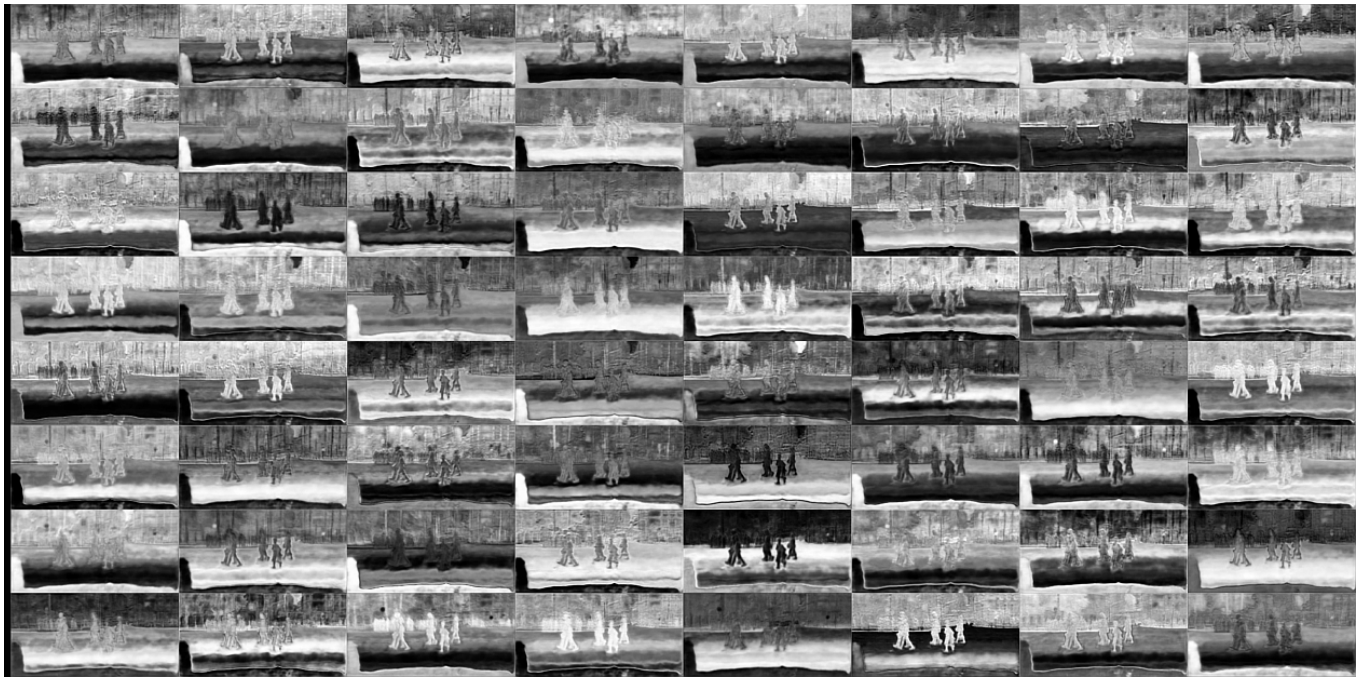


Fig. 7 Examples of weight map of FFM

Table 3 Semantic segmentation results of our proposed model based on NYUDv2 dataset

	Att	ASPP	$\mathcal{L}_d$	$\mathcal{L}_{Zoom}$	mIoU(%)	mPA	Acc
w/o FFM	✓				39.7	51.3	72.0
		✓			46.1	59.3	74.9
	✓	✓			46.8	59.3	75.1
w/ FFM	✓	✓	✓		47.6	59.3	77.4
	✓	✓	✓	✓	**48.1**	**60.8**	**77.4**

Similarly, to reveal how FFM focuses on shared features, we visualize the weight map of F_{weight}. This weight map is from the last FFM in the decoding phase with 64 feature channels, as shown in **Fig. 7**. It can be seen that the weight mapping focuses on the common features of the two tasks, such as the edges, and shapes of the objects. These are consistent with the common features of the two tasks.

4.3.2 NYUDv2 experimental results

As shown in **Table 3**, these results are derived from the test set of NYUDv2 dataset. The semantic segmentation accuracies are improved by 7.1% (Att) and 0.7% (ASPP) when using two feature enhancement modules (ASPP & Att) and when using a single feature enhancement module. When using two feature enhancement modules for the semantic segmentation and depth estimation task, the semantic segmentation accuracy was 48.1%, and the proposed loss function term $\mathcal{L}_{Zoom}$ improved by 0.5% compared to the conventional loss function term $\mathcal{L}_d$.

To verify the performance of the proposed model on the NYUDv2 dataset, we also compared our method with other related methods which included one-stage and two-stage($^{\triangle}$ represents that this is a two-stage method.) strategies, as shown in **Table 4**. It can be seen that the semantic segmentation model of the proposed model obtains competitive results compared to related methods. We see that the proposed method has a relatively low mIoU compared to the prediction results

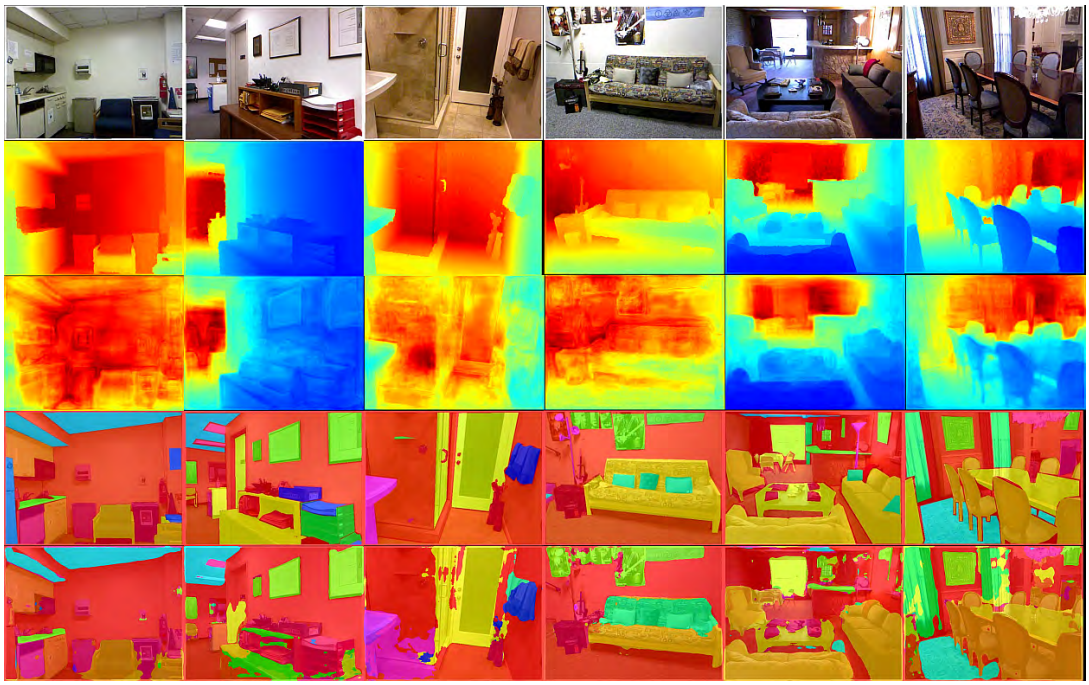


Fig. 8 Examples of prediction results based on NYUDv2 dataset

Table 4 Semantic segmentation performance comparison of multitasking algorithms based on NYUDv2 dataset

Methods	mIoU(%)	mPA	Acc
FCN[12]	29.2	42.2	60.0
Deng et al.[14]	31.5	-	63.8
Eigen et al.[28]	34.1	45.1	65.6
Arsalan et al.[△4]	39.2	52.3	68.6
Context[15]	40.6	53.6	70.0
CI-Net[20]	42.6	-	72.7
ESOSD-Net[21]	45.0	**64.7**	73.3
RefineNet[13]	46.5	58.9	73.6
PAD-Net[△3]	50.2	62.3	75.2
SF-Net	**48.1**	60.8	**77.4**

of PAD-Net, and also have relatively low mPA values compared to the results of ESOSD-Net. On the other hand, our method obtains relatively high Acc value. We consider that the reason for this phenomenon is that the number of samples of certain classes in the dataset is extremely small, and the proposal model is not sufficiently learned for a few classes, leading to unsatisfactory prediction results. Therefore, the proposed model is yet to be optimized for the class balancing problem.

Table 5 shows the depth estimation results of the ablation experiments based on the NYUDv2 dataset. These results are derived from the test set of NYUDv2 dataset. When both feature enhancement modules are used, rel is reduced from 0.153 to 0.144, rms is reduced from 0.499 to 0.462, and log_{10} is reduced from 0.068 to 0.060 as compared to no feature enhancement module. After using the proposed Zoom loss function, rel was reduced to 0.125, rms to 0.385, and log_{10} to 0.052. These values are the best values in the ablation experiments.

Table 5 Depth estimation results of our proposed model based on NYUDv2 dataset

Att	ASPP	$\mathcal{L}_d$	$\mathcal{L}_{Zoom}$	rel	rms	log_{10}
		✓		0.153	0.499	0.068
✓		✓		0.165	0.539	0.071
	✓	✓		0.146	0.470	0.062
✓	✓	✓		0.144	0.462	0.060
✓	✓	✓	✓	**0.125**	**0.385**	**0.052**

As shown in **Table 6**, we have compared the depth values of our proposed method with other algorithms in the NYUDv2 dataset as well. DORN still gives the best results on the rel and log_{10} metrics, since it is optimized for a separate depth estimation task and was trained using the full 120K (including imprecisely labeled data) images of NYUDv2. Although the experimental conditions are different, the results of our method are close to those of DORN. It can be seen that our method achieves the best results in the vast majority of indicators compared to other methods.

Figure 8 shows the prediction results based on the NYUDv2 dataset, which include the results of depth estimation and semantic segmentation obtained by the proposed model. The first row shows the input images, the second row shows the true value of the depth images, the third row shows the predicted result of the depth images, the fourth row shows the true values of the semantic segmentation, and the fifth row shows the predicted results of the semantic segmentation.

5. Conclusion

In this study, we developed a new fusion network structure (SF-Net) for simultaneously learning semantic segmentation

Table 6 Depth estimation performance comparison of multi-tasking algorithms based on NYUDv2 dataset

Methods	Error(lower is better)			Accuracy(higher is better)		
	rel	rms	log_{10}	δ_1	δ_2	δ_3
Make3D[27]	0.349	1.214	-	0.447	0.745	0.897
Liu et al.[36]	0.335	1.06	0.127	-	-	-
Li et al.[8]	0.232	0.821	0.094	-	-	-
Liu et al.[5]	0.230	0.824	0.095	0.614	0.883	0.975
HCRF$^{\Delta}$[37]	0.220	0.745	0.094	0.605	0.890	0.970
Eigen et al.[7]	0.215	0.907	-	0.611	0.887	0.971
Roy et al.[9]	0.187	0.744	0.078	-	-	-
Eigen et al.[28]	0.158	0.641	-	0.769	0.950	0.988
Jafari et al$^{\Delta}$.[38]	0.157	0.673	0.068	0.762	0.948	0.988
He et al.[16]	0.151	0.572	0.064	0.789	0.948	0.98
SOSD-Net[21]	0.145	0.514	0.062	0.805	0.962	0.992
Laina et al.[39]	0.129	0.583	0.056	0.811	0.953	0.988
CI-Net[20]	0.129	0.504	-	0.812	0.957	0.990
PAD-Net$^{\Delta}$[3]	0.120	0.582	0.055	0.817	0.954	0.987
DORN[11]	**0.115**	0.509	**0.051**	0.828	0.965	0.992
SF-Net	0.125	**0.385**	0.052	**0.856**	**0.979**	**0.996**

and monocular depth estimation. We derived a zoom coefficient that represents the relationship between the size of the divided region and the depth value and proposed a new loss function using it. We constructed a new attention module to utilize the semantic information in image features effectively. Furthermore, we proposed a feature fusion module that improves the performance of each task by calculating shared feature weights using feature maps extracted from different tasks. We verified the effectiveness of our proposed method through experiments and comparisons using the Cityscapes dataset and NYUDv2 dataset. Utilizing the depth images in the dataset to supervise the depth estimation task can effectively improve the accuracy of semantic segmentation. Moreover, the simultaneous learning between semantic segmentation and depth estimation tasks proved to be mutually beneficial and assisted each other. It was worth emphasizing that the loss function formulated based on the proposed zoom factor shows improvements in semantic segmentation tasks. Our future work will improve the forwarding of low-level features. We are considering introducing a gating mechanism to solve the noise problem in low-level features and optimize the model performance.

References

1) A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, W. Burgard: "Multimodal deep learning for robust rgb-d object recognition", 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp.681–687 (2015).
2) X. Xu, Y. Li, G. Wu, J. Luo: "Multi-modal deep feature learning for rgb-d object detection", *Pattern Recognition*, Vol.72, pp.300–313 (2017).
3) D. Xu, W. Ouyang, X. Wang, N. Sebe: "Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp.675–684 (2018).
4) A. Mousavian, H. Pirsiavash, J. Košecká: "Joint semantic segmentation and depth estimation with deep convolutional networks", 2016 Fourth International Conference on 3D Vision (3DV), pp.611–619 (2016).
5) F. Liu, C. Shen, G. Lin: "Deep convolutional neural fields for depth estimation from a single image", Proc. of the IEEE conference on computer vision and pattern recognition, pp.5162–5170 (2015).
6) L. Zhu, Z. Kang, M. Zhou, X. Yang, Z. Wang, Z. Cao, C. Ye: "Cmanet: Cross-modality attention network for indoor-scene semantic segmentation", *Sensors*, Vol.22, No.21, p.8520 (2022).
7) D. Eigen, C. Puhrsch, R. Fergus: "Depth map prediction from a single image using a multi-scale deep network", *Advances in neural information processing systems*, Vol.27 (2014).
8) B. Li, C. Shen, Y. Dai, A. Van Den Hengel, M. He: "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs", Proc. of the IEEE conference on computer vision and pattern recognition, pp.1119–1127 (2015).
9) A. Roy, S. Todorovic: "Monocular depth estimation using neural regression forest", Proc. of the IEEE conference on computer vision and pattern recognition, pp.5506–5514 (2016).
10) C. Godard, O. Mac Aodha, G. J. Brostow: "Unsupervised monocular depth estimation with left-right consistency", Proc. of the IEEE conference on computer vision and pattern recognition, pp.270–279 (2017).
11) H. Fu, M. Gong, C. Wang, K. Batmanghelich, D. Tao: "Deep ordinal regression network for monocular depth estimation", Proc. of the IEEE conference on computer vision and pattern recognition, pp.2002–2011 (2018).
12) J. Long, E. Shelhamer, T. Darrell: "Fully convolutional networks for semantic segmentation", Proc. of the IEEE conference on computer vision and pattern recognition, pp.3431–3440 (2015).
13) G. Lin, A. Milan, C. Shen, I. Reid: "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation", Proc. of the IEEE conference on computer vision and pattern recognition, pp.1925–1934 (2017).
14) Z. Deng, S. Todorovic, L. Jan Latecki: "Semantic segmentation of rgbd images with mutex constraints", Proc. of the IEEE international conference on computer vision, pp.1733–1741 (2015).
15) G. Lin, C. Shen, A. Van Den Hengel, I. Reid: "Efficient piecewise training of deep structured models for semantic segmentation", Proc. of the IEEE conference on computer vision and pattern recognition, pp.3194–3203 (2016).
16) L. He, G. Wang, Z. Hu: "Learning depth from single images with deep neural network embedding focal length", *IEEE Trans. on Image Processing*, Vol.27, No.9, pp.4676–4689 (2018).
17) P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, A. L. Yuille: "Towards unified depth and semantic prediction from a single image", Proc. of the IEEE conference on computer vision and pattern recognition, pp.2800–2809 (2015).
18) I. Misra, A. Shrivastava, A. Gupta, M. Hebert: "Cross-stitch networks for multi-task learning", Proc. of the IEEE conference on computer vision and pattern recognition, pp.3994–4003 (2016).
19) Z. Zhang, Z. Cui, C. Xu, Z. Jie, X. Li, J. Yang: "Joint task-recursive learning for semantic segmentation and depth estimation", Proc. of the European Conference on Computer Vision (ECCV), pp.235–251 (2018).
20) T. Gao, W. Wei, Z. Cai, Z. Fan, S. Q. Xie, X. Wang, Q. Yu: "Ci-net: A joint depth estimation and semantic segmentation network using contextual information", *Applied Intelligence*, Vol.52, No.15, pp.18 167–18 186 (2022).
21) L. He, J. Lu, G. Wang, S. Song, J. Zhou: "Sosd-net: Joint semantic object segmentation and depth estimation from monocular images", *Neurocomputing*, Vol.440, pp.251–263 (2021).
22) O. Ronneberger, P. Fischer, T. Brox: "U-net: Convolutional networks for biomedical image segmentation", Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pp.234–241 (2015).
23) H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia: "Pyramid scene parsing network",

Proc. of the IEEE conference on computer vision and pattern recognition, pp.2881–2890 (2017).

24) L. C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam: "Encoder-decoder with atrous separable convolution for semantic image segmentation", Proc. of the European conference on computer vision (ECCV), pp.801–818 (2018).
25) J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu: "Dual attention network for scene segmentation", Proc. of the IEEE/CVF conference on computer vision and pattern recognition, pp.3146–3154 (2019).
26) S. Woo, J. Park, J. Y. Lee, I. S. Kweon: "Cbam: Convolutional block attention module", Proc. of the European conference on computer vision (ECCV), pp.3–19 (2018).
27) A. Saxena, M. Sun, A. Y. Ng: "Make3d: Learning 3d scene structure from a single still image", *IEEE Trans. on pattern analysis and machine intelligence*, Vol.31, No.5, pp.824–840 (2008).
28) D. Eigen, R. Fergus: "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture", Proc. of the IEEE international conference on computer vision, pp.2650–2658 (2015).
29) M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele: "The cityscapes dataset for semantic urban scene understanding", Proc. of the IEEE conference on computer vision and pattern recognition, pp.3213–3223 (2016).
30) N. Silberman, D. Hoiem, P. Kohli, R. Fergus: "Indoor segmentation and support inference from rgbd images.", *ECCV (5)*, Vol.7576, pp.746–760 (2012).
31) J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, L. Fei-Fei: "Imagenet: A large-scale hierarchical image database", 2009 IEEE conference on computer vision and pattern recognition, pp.248–255 (2009).
32) D. P. Kingma, J. Ba: "Adam: A method for stochastic optimization", *arXiv preprint arXiv:1412.6980* (2014).
33) A. Kendall, Y. Gal, R. Cipolla: "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics", Proc. of the IEEE conference on computer vision and pattern recognition, pp.7482–7491 (2018).
34) Z. Chen, V. Badrinarayanan, C. Y. Lee, A. Rabinovich: "Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks", International conference on machine learning, pp.794–803 (2018).
35) O. Sener, V. Koltun: "Multi-task learning as multi-objective optimization", *Advances in neural information processing systems*, Vol.31 (2018).
36) M. Liu, M. Salzmann, X. He: "Discrete-continuous depth estimation from a single image", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp.716–723 (2014).
37) P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, A. L. Yuille: "Towards unified depth and semantic prediction from a single image", Proc. of the IEEE conference on computer vision and pattern recognition, pp.2800–2809 (2015).
38) O. H. Jafari, O. Groth, A. Kirillov, M. Y. Yang, C. Rother: "Analyzing modular cnn architectures for joint depth prediction and semantic segmentation", 2017 IEEE International Conference on Robotics and Automation (ICRA), pp.4620–4627 (2017).
39) I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, N. Navab: "Deeper depth prediction with fully convolutional residual networks", 2016 Fourth international conference on 3D vision (3DV), pp.239–248 (2016).

Kai WANG (*Student Member*)
He received a master's degree from the Graduate School of Systems Engineering, Wakayama University, Japan in 2023. The same year, he entered the Graduate School of Systems Engineering to work toward a Ph.D. And has been engaged in research on semantic segmentation, which has continued to this day. His research interests include deep learning, image segmentation, and computer vision. He is a member of IIEEJ and RSJ.

Takayuki NAKAMURA
He received his M.S. degrees in Graduate School of Engineering, Osaka University, Japan in 1993. In 1996, he received the Dr. Eng. Degree from Graduate School of Engineering, Osaka University, Japan. In 1996, he became a Research Fellow of the Japan Society for the Promotion of Science. In 1997, he joined the Graduate School of Information Science, Nara Institute of Science and Technology, as an assistant professor. In 2002, he joined the Department of Information and Communication Systems, Faculty of Systems Engineering, Wakayama University, as an associate professor. In 2013, he became a professor in the same department at the same university. His research interests include robot vision, robot learning systems based on visual information, self-position estimation methods for mobile robots and 3D point cloud data processing. He is a member of JSME, RSJ, and SICE.

IIEEJ **Paper**

Bit Depth Enhancement Considering Semantic Contextual Information via Spatial Feature Transform

Taishi IRIYAMA†(*Member*), Yuki WATANABE†, Takashi KOMURO†

† Department of Information and Computer Sciences, Saitama University

<**Summary**> In this paper, we propose a novel bit depth enhancement (BDE) model that considers semantic contextual information by incorporating spatial feature transform (SFT) layers into the BDE model. In the proposed method, we adopt the pixel-wise class probability maps of the input image obtained by semantic segmentation as prior information for SFT. The SFT layer transforms from the input feature maps into the modulated feature maps by considering the contextual information modulated using affine parameters generated from the prior information. The proposed method considers the pixel-wise details by a proposed network that preserves spatial dimensions and the contextual information by incorporating the SFT layers conditioned with semantic information. Moreover, the proposed method adopts a perceptual loss function to recover the visually natural luminance changes by considering the contextual information. The experimental results show that the proposed BDE method achieves better performance compared with existing DNN-based BDE methods for restoring 8-bit depth from 3,4, and 6-bit depths. In addition, we investigate how to provide the contextual information to the BDE model, and show that providing it through the SFT layer is effective compared with other providing methods.

Keywords: bit depth enhancement, de-quantization, de-banding, convolutional neural network, conditional normalization

1. Introduction

In recent years, the advancement of display technology has led to the development of high-end consumer display devices capable of representing more extensive color information, such as those with 10 and 12-bit depth. These technologies have improved visual experiences by representing more colors and smoother gradations, resulting in realistic and visually appealing images. To fully harness the potential of these advanced display devices, it is crucial that image and video contents are recorded with a higher bit depth than the display capability. However, high-bit depth contents have the inherent problem of increasing the amount of data required for storage and transmission.

To address this issue, bit depth enhancement (BDE) techniques have gained significant attention in the field of image processing. BDE aims to expand low-bit depth images to high-bit depth images by estimating the missing least significant bits (LSBs), enabling the enhancement of existing low-bit depth content to match the capabilities of modern high-bit depth displays. Nevertheless, simple BDE techniques often fall short in producing satisfactory results and introduce artifacts that decrease the visual quality of the enhanced images. Two common artifacts caused by quantization errors are false contours and missing details. False contours appear as visible steps or bands in areas that should have smooth gradations of color or brightness. Missing details refer to the loss of subtle luminance changes within the quantization step.

To address these artifacts and improve the performance of BDE, several BDE methods have utilized learning-based approaches in recent years[1)–8)]. In learning based-BDE tasks, a variety of deep neural network (DNN) architectures have been explored, each presenting distinct advantages and inherent limitations. Encoder-decoder-based architectures compress spatial dimensions to capture global contextual information, potentially compromising pixel-wise detail information. Residual network (ResNet)-based architectures that employ skip connections alongside convolution layers with same padding to maintain spatial dimensions, can effectively preserve pixel-wise detail information. However, the trade-off between capturing a large receptive field and preserving pixel-wise details remains. Although increasing the depth of neural networks enhances their ability to capture global

contextual information, it also results in a higher computational cost.

Meanwhile, conditional normalization methods have been shown to be effective in integrating contextual information into DNNs. These techniques, such as adaptive instance normalization (AdaIN)[9], feature-wise linear modulation (FiLM)[10], and spatial feature transform (SFT)[11], adjust the feature maps of a network based on specific conditional information. With conditional normalization, the network preserves the spatial details of the features while infusing them with semantic contextual information as the conditions.

In this paper, we propose a novel BDE model that preserves pixel-wise details and global contextual information by leveraging a ResNet-based architecture in combination with SFT layers. The proposed method adopts pixel-wise class probability maps of the input image, obtained through semantic segmentation, as prior information for the SFT layers. The SFT layer transforms the input feature maps into modulated feature maps by incorporating contextual information using affine parameters generated from the prior information. Moreover, the proposed method adopts the VGG-based perceptual loss function to consider the semantic contextual information and recover the visually natural luminance changes. Experiment results show that the proposed BDE method achieves better performance compared with existing DNN-based BDE methods for restoring 8-bit depth from 3,4, and 6-bit depths. Furthermore, we investigate various approaches to provide contextual information to the BDE model and show that incorporating it through the SFT layers is the most effective method.

2. Related Work

2.1 Bit depth enhancement

BDE methods have been widely explored in the literature to increase the number of bits used to represent each pixel, thereby expanding the range of possible color values and improving the overall visual quality of the image.

One of the simplest BDE method is zero padding (ZP). In ZP, zeros are added after the LSB of each pixel in a low-bit depth image to obtain an image with the desired higher bit depth. Another traditional BDE method is multiplication by ideal gain (MIG). MIG involves multiplying each pixel value in the low-bit depth image by a constant factor to scale it up to the desired higher bit depth. Bit-replication (BR)[12]is yet another traditional BDE method. In BR, the most significant bits (MSBs) of each pixel in the low-bit depth image are replicated to achieve the desired higher bit depth. While these traditional BDE methods are computationally efficient and easy to implement, it does not provide any additional information to the image and may result in visible artifacts, especially in areas with smooth gradients or subtle color variations.

To address the limitations of these methods and effectively remove false contour artifacts, advanced context-aware algorithms have been proposed such as BDE by contour region reconstruction (CRR)[13]and content adaptive (CA) BDE[14], and intensity potential for adaptive de-quantization (IPAD)[15]. CRR calculates high-bit depth pixel values based on the distances to the nearest contour edges. Although CRR can largely eliminate false contours, it blurs out details in regions with local extrema. CA adaptively enhances the bit depth of an image based on its local content. It improves CRR by utilizing neighboring pixel values to interpolate the missing bits, addressing the blurry details in regions with local extrema. However, the enhanced images still suffer from over smoothing and unnatural false contours. IPAD takes a different approach by utilizing an intensity potential field to model the complicated relationships among pixels and adaptively de-quantizes the image based on the potential field. While IPAD achieves higher accuracy compared to the above mentioned methods, it still suffers from false contour and missing detail artifacts in scenarios with a large number of missing LSBs.

2.2 Learning-based bit depth enhancement

In recent years, DNNs have emerged as a powerful tool for various image processing tasks, including BDE. Learning-based BDE methods leverage the hierarchical feature extraction capabilities of DNNs to capture complex contextual information and estimate enhanced images with reduced artifacts and improved visual quality.

Byun et al. proposed BitNet[3], a convolutional neural network (CNN)-based BDE model with an encoder-decoder-based architecture. The encoder part of BitNet consists of multiple convolutional layers that gradually downsample the input image and extract high-level features. The decoder part then upsamples the feature maps and reconstructs the enhanced image. By using a multi-layered encoder, BitNet achieves a large receptive field with a relatively small number of layers, enabling it to capture global contextual information and effectively suppress false contours over wide and smooth regions in the

image. However, the spatial compression of feature maps in the encoder may result in the loss of pixel-wise details, especially in areas with high-frequency components. To address the issue of recovering fine details, Zhao et al. introduced a deep BDE network (BDEN)[5], a BDE model based on a deep ResNet architecture[16]. BDEN utilizes convolution layers with same padding to extract detailed feature maps while preserving their spatial dimensions. By maintaining the resolution of the feature maps throughout the network, BDEN is able to accurately recover missing LSBs in high-frequency regions of the image. However, the deep-layered architecture required to obtain a large receptive field comes at the cost of increased computational complexity and memory consumption. To mitigate the computational cost while still achieving a large receptive field, Zhao et al. also proposed a lighter but efficient BDE network (LBDEN)[6], which incorporates dilated convolutions into the BDEN architecture. By replacing some of the standard convolution layers in BDEN with dilated convolution layers, LBDEN can expand its receptive field without increasing the number of parameters and computational cost, and capture a global contextual information while maintaining pixel-wise details. However, the use of dilated convolutions may result in a reduction of local connectivity and spatial coherence, as the gaps between kernel elements can cause the network to miss important local patterns.

Due to this trade-off between capturing local and global information while maintaining computational cost, simultaneously addressing false contour and missing detail artifacts remains a significant challenge in the field of BDE.

2.3 Conditional normalization

Conditional normalization methods adjust feature maps through affine transformations based on parameters derived from specific conditional information[9)–11)]. These methods have been explored for incorporating contextual information into DNN architectures.

Huang et al. proposed an adaptive instance normalization (AdaIN) for style transfer, which aligns the mean and standard deviation of content features with those of style features[9]. Perez et al. proposed feature-wise linear modulation (FiLM), which learns affine parameters for each feature map from conditional information and normalizes the intermediate features accordingly[10]. By applying these learned affine transformations to normalize the intermediate features, FiLM enables the network to modulate its feature representations based on the given context. Wang et al. presented spatial feature transform (SFT), which applies different normalizations for each feature map and spatial dimension using a learnable mapping function[11]. SFT offers more spatially flexible feature modulation compared to FiLM.

3. Proposed Method

In this paper, we propose a novel BDE model that leverages a ResNet-based CNN architecture with SFT layers to preserve pixel-wise details and global contextual information. The motivation of the proposed method is to achieve both suppression of false contours and restoration of details by providing the class probability maps obtained by semantic segmentation model as contextual information. These contextual information can be used to modulate the intermediate features of the network to achieve a contextualized prediction of the LSBs.

3.1 Spatial feature transform

A SFT obtains the modulation parameters $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ for each element of the intermediate feature maps $\boldsymbol{F}$ based on the given conditions Ψ, and modulates the feature maps based on the obtained parameters[11]. The SFT is formulated by

$$\mathrm{SFT}(\boldsymbol{F}|\Psi) = \boldsymbol{\gamma} \odot \boldsymbol{F} + \boldsymbol{\beta} \tag{1}$$

where, $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ indicate the affine parameters for scaling and shifting, $\boldsymbol{F}$ is feature maps and $\odot$ is Hadamard product.

The structure of the SFT layer for incorporation into DNNs is shown in **Fig.1**. In the SFT layer, affine parameters for scaling and shifting are separately generated from the given conditional information using two convolutional layers and Leaky ReLU[17]. Then, the SFT layer outputs the modulated feature maps by the element-wise multiplication and the element-wise summation of the generated affine parameters to the input feature maps.

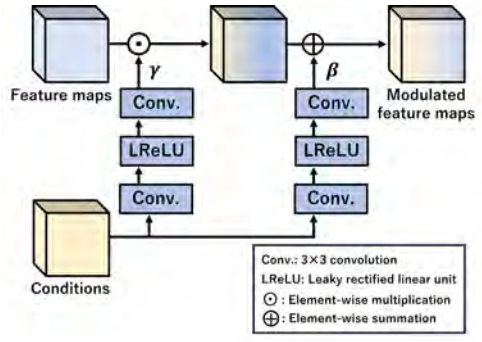


Fig. 1 Illustration of SFT layer

3.2 Network architecture

We design a novel network architecture that incorporates the SFT layer into the ResNet-based BDE network. The detail of network structure is shown in **Fig.2**. The network consists of three main parts: initial embedding, feature extraction, and reconstruction. The initial embedding stage comprises a single convolutional layer that embeds the input image into feature maps with 64 channels. This step helps to capture the low-level features and prepares the input for further processing. The feature extraction stage is a stack of 16 residual blocks (RBs), where each RB consists of two convolutional layers and a skip connection that bypasses the input. This stage is responsible for extracting high-level features and capturing the contextual information necessary for the enhancement process. Additionally, we incorporate SFT layers before each convolutional layer in each RB and after the stack of RBs. The structure of the SFT layer is the same as shown in Fig. 1. The input to each SFT layer is the conditional information extracted from the segmentation map obtained by the pre-trained segmentation network applied to the input image. The conditional information is processed by four convolutional layers to extract the relevant contextual information. The reconstruction stage consists of four convolutional layers and is responsible for mapping the residuals between the input low-bit depth image and the target high-bit depth image from the extracted features. This stage aims to generate the final enhanced image by incorporating the learned contextual information and preserving the pixel-wise details. All convolutional layers in the network have a 3×3 kernel size. The activation function used in the condition network is Leaky ReLU, while ReLU is used in the other parts of the network. By leveraging the SFT layers and incorporating semantic contextual information, our proposed network architecture is designed to effectively enhance low-bit depth images while capturing the semantic contextual modulation and preserving the fine details of the input image.

3.3 Loss function

In the context of BDE tasks, it has been observed that relying solely on the mean squared error (MSE) loss function can lead to the retention of artifacts, such as false contours, in the enhanced images. To address this issue, various alternative loss functions have been explored and considered.

Liu et al.[18]demonstrated that replacing the MSE loss with the perceptual loss, which is essentially the MSE loss computed on the features extracted from the pre-trained VGG-19 network[19], can significantly suppress false contours in the reconstructed images. The perceptual loss takes into account the high-level features learned by the VGG-19 network, which capture more semantic and perceptual information compared to pixel-wise differences.

Inspired by this finding, we adopt a perceptual loss in the proposed method to effectively suppress false contour artifacts while preserving the semantic and perceptual properties of the image. In BDE tasks, while low-level features are typically emphasized due to the importance of local details, the proposed method also considers semantic contextual information obtained obtained from segmentation model. To balance these aspects, we have empirically selected the feature maps from the 8th layer of the VGG-19 network as the basis for computing the perceptual loss.

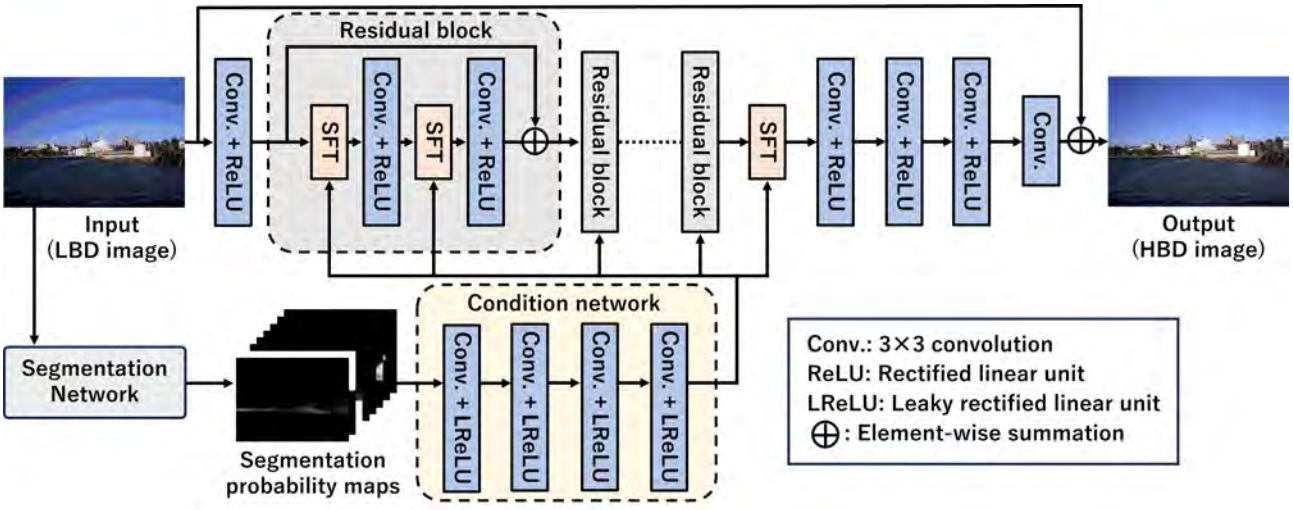


Fig. 2 Illustration of network architecture

The perceptual loss is formulated as follows:

$$L = \frac{1}{WHC} \sum_{i=1}^{HWC} (VGG_i(I) - VGG_i(\hat{I}))^2 \quad (2)$$

where i represents the index of the feature maps in the 8th layer of the VGG-19 network, and W, H, and C denote the width, height, and channel dimensions of the feature map, respectively. I represents the ground truth high-bit depth image, and $\hat{I}$ represents the enhanced image generated by our proposed model.

4. Experiments

In the experiments, we use OutdoorSeg dataset[11]used for 8-bit images, and the Sintel dataset[20]and the MIT-Adobe FiveK dataset[21]for 16-bit images. The OutdoorSeg dataset is natural image of outdoor scene consisting of 9,900 images with 8-bit depth, including 8,447 images collected from the ADE dataset[22], 899 from the Flickr website and 554 from the COCO dataset[23]. The Sintel dataset is a short animation film consisting of 21,312 frames with 16-bit (436 × 1024 pixels). The MIT-Adobe FiveK dataset is a natural image with different tones adjusted by five photography experts, each consisting of 5,000 photographs with 16-bit. For 8-bit training dataset, we randomly selected 1,414 images from the OutdoorSeg dataset, and randomly cropped them to 64×64 patches, and quantized into the 3, 4 and 6-bit depth and dequantized into 8 bit depth by the ZP. For 16-bit training dataset, we randomly selected 2,000 images: 1,000 from the Sintel dataset and 1,000 from images a0001-2000 of the MIT-Adobe FiveK dataset adjusted by expert C, and randomly cropped them into 64×64 patches, and quantized into the 3 and 4-bit depth and dequantized into 16-bit depth by the ZP. Based on the experimental settings used in previous study[7], the MIT-Adobe FiveK dataset is downsampled to half the size in each spatial dimension before cropping. The mini-batch size is set to 4. Our model is optimized by Adam optimizer[24] with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. The learning rate is 2×10^{-4} initially and is halved every 200 epochs. For testing, we use 300 images from the OutdoorSceneTest300 dataset as 8-bit depth images, and 100 images from a4901–5000 of the MIT-Adobe FiveK dataset as 16-bit depth images.

4.1 Segmentation network

Following SFTGAN[11], we use an 8-class segmentation model proposed by Liu et al.[25]pre-trained on the COCO dataset and fine-tuned with the ADE dataset as the segmentation network in proposed method.

To confirm the performance of this pre-trained segmentation model for low-bit depth images, we calculated the mean Intersection over Union (mIoU) between the predicted segmentation maps for low-bit depth images and the ground truth labels. **Table 1** shows the mIoU values for 3, 4, 6, and 8-bit depth images, and **Fig.3** illustrates the segmentation maps for 3 and 8-bit depth images. As shown in these results, the reduction in mIoU for low-bit depth images is limited, indicating that this segmentation model is robust to decreases in bit depth. While there is a slight decrease of 0.0325 in mIoU for 3-bit depth images, the segmentation maps show close results to the 8-bit depth images, confirming that this segmentation model maintains relatively consistent performance even at lower bit depths. Therefore, in this experiments, we use this segmentation model regardless of the quantization level.

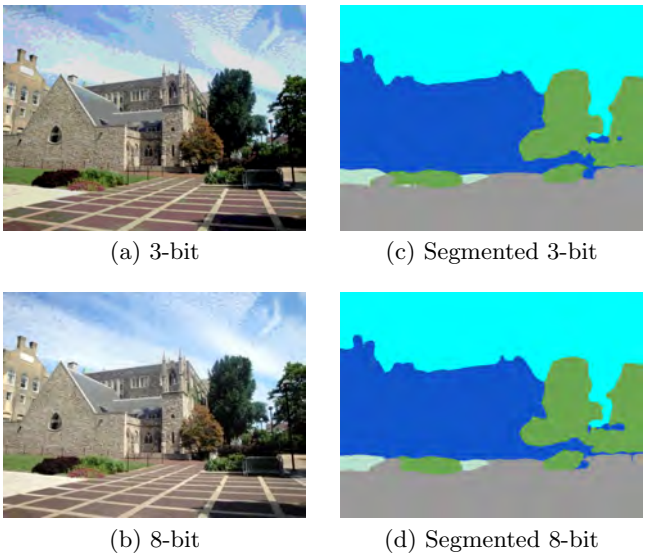
4.2 Comparison with conventional methods

The proposed method is compared with conventional CNN-based BDE methods, BitNet[3]and BWBDR[7]. For an objective evaluation of BDE performance, two assessment indices are adopted: composite peak signal-to-noise ratio (CPSNR) and structural similarity (SSIM). The CPSNR is calculated as follows:

$$\mathrm{CPSNR} = 10 \log_{10} \left(\frac{L^2}{\mathrm{MSE}} \right) \quad (3)$$

Table 1 mIoU values for different bit depth

Bit depth	3-bit	4-bit	6-bit	8-bit
mIoU	0.5339	0.5632	0.5696	0.5664



(a) 3-bit (c) Segmented 3-bit

(b) 8-bit (d) Segmented 8-bit

Fig. 3 Segmentation results for different bit depth from training data

where L is the dynamic range of the pixel values, determined by the bit-depth b of the reference image, and given by $L = 2^b - 1$. The MSE between reference image y and enhanced image x is calculated as:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 \tag{4}$$

where x_i and y_i represent the pixel values of x and y, respectively, and N is the total number of pixels. The SSIM is calculated as follows:

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{5}$$

where μ_x and μ_y are the mean values of x and y, σ_x^2 and σ_y^2 are the variances, and σ_{xy} is the covariance between x and y. These values are computed within a sliding window of size 11×11. The constants C_1 and C_2 are defined as $C_1 = (0.01L)^2$ and $C_2 = (0.03L)^2$.

Table 2 shows the CPSNR and SSIM results of restoring 8-bit depth from 3, 4 and 6-bit decpths for the comparison methods and the proposed method. The best value is indicated in red and the second best value is indicated in blue. As shown in this table, our method provides better accuracy compared with other BDE methods in all assessment indices. Particularly in comparison with BWBDR, which shows the second best accuracy in all assessment indices, our method demonstrates an average improvement of approximately 0.97 [dB] in CPSNR and 0.021 in SSIM for the restoration of 3, 4 and 6-bit depths. To further validate these improvements, we conducted paired t-tests comparing our method with BWBDR, using a significance level of 0.05. The results showed statistically significant differences in almost cases except for SSIM in 3-8 bit restoration.

Table 3 shows the CPSNR and SSIM results of restoring 16-bit depth from 3 and 4-bit depths for the comparison methods and the proposed method. As shown in this table, for 3-16 bit restoration, our method provides competitive result with BWBDR. For 4-16 bit restoration, while our method underperforms compared to BWBDR, it provides conpetitive results with BitNet.

As a subjective evaluation, **Fig.4** shows a part of the BDE results of restoring 8-bit depth from 3-bit depths by the comparison methods and our method. For "OST: 52" in Fig. 4, it can be seen that false contours contained in the input remain on BitNet, and color distortion and blurring are observed on BWBDR. In our method, false contours in the input are smoothly suppressed, and the colors are reproduced to close the ground truth compared with other methods. For "OST: 77" in Fig. 4, the detailed texture of the water surface becomes more homogenized and less distinct on BitNet and BWBDR, whereas our method recovers much of the textural detail. For "OST: 265" in Fig. 4, the BitNet results in a blurring of the texture of the grassland, and BWBDR induces a color that is slightly greener than the ground truth. On the other hand, our method successfully recovers the grassland texture with a level of detail comparable to that of the ground truth.

4.3 Ablation study

To evaluate the effectiveness of segmentation information in BDE tasks, we conduct ablation studies on segmentation information and the SFT layers. In these experiments, we prepare four models: the proposed model, a model without segmentation information (w/o seg), a

Table 3 Objective evaluation of conventional BDE methods for restoring 16-bit depth

Method		3-16 bit	4-16 bit
ZP (Input)	CPSNR	23.20 ± 0.36	29.36 ± 0.28
	SSIM	0.7133 ± 0.0911	0.8745 ± 0.0488
BitNet	CPSNR	33.02 ± 1.22	38.24 ± 1.24
	SSIM	0.8918 ± 0.0436	0.9531 ± 0.0198
BWBDR	CPSNR	33.16 ± 1.16	38.75 ± 1.21
	SSIM	0.8950 ± 0.0502	0.9574 ± 0.0181
Ours	CPSNR	33.18 ± 1.24	38.39 ± 1.62
	SSIM	0.8950 ± 0.0516	0.9529 ± 0.0235

Table 2 Objective evaluation of conventional BDE methods for restoring 8-bit depth

Method		3-8 bit	4-8 bit	6-8 bit
ZP (Input)	CPSNR	22.69 ± 0.47	28.75 ± 0.39	40.97 ± 0.50
	SSIM	0.8008 ± 0.0656	0.9172 ± 0.0360	0.9930 ± 0.0045
BitNet	CPSNR	32.84 ± 1.19	38.22 ± 1.14	47.21 ± 0.95
	SSIM	0.9302 ± 0.0219	0.9728 ± 0.0117	0.9958 ± 0.0020
BWBDR	CPSNR	33.41 ± 1.23	39.25 ± 1.10	48.37 ± 0.59
	SSIM	0.9423 ± 0.0202	0.9776 ± 0.0116	0.9968 ± 0.0021
Ours	CPSNR	34.13* ± 1.69	40.32* ± 2.36	49.49* ± 1.81
	SSIM	0.9443 ± 0.0263	0.9804* ± 0.0171	0.9974* ± 0.0026

* Indicates statistically significant difference ($p < 0.05$) compared to the BWBDR

(a) OST: 52 (b) Input (ZP) (c) BitNet

(d) BWBDR (e) Ours (f) Ground truth

(g) OST: 77 (h) Input (ZP) (i) BitNet

(j) BWBDR (k) Ours (l) Ground truth

(m) OST: 265 (n) Input (ZP) (o) BitNet

(p) BWBDR (q) Ours (r) Ground truth

Fig. 4 Subjective evaluation for BDE results of recovering 8-bit from 3-bit on OST dataset

model without the SFT layers (w/o SFT), and a model without both segmentation information and the SFT layers (w/o both). For the model without segmentation information, we design the SFT layers to receive zero vectors as conditional information instead of semantic class probability maps, effectively eliminating the consideration of contextual information. For the model without the SFT layers, we remove all SFT layers from the proposed model and concatenate the semantic class probability maps along the channel dimension of the input low-bit depth image. For the model without both segmentation information and the SFT layers, we use the standard ResNet-based CNN architecture without any module or information. This model serves as the baseline to evaluate the individual and combined contributions of segmentation information and the SFT layers to the BDE task.

Table 4 shows the CPSNR and SSIM results of restoring 8-bit depth from 3, 4 and 6-bit depths for the ablation models and the proposed model. As shown in this table, the proposed model achieves superior performance in almost all metrics and comparable performance in some others. Comparing the "w/o seg" to the baseline model, we observe that it performs comparably to or better than the baseline in most metrics. This can be attributed to the increased number of parameters in the model due to the SFT layers. On the other hand, comparing the "w/o SFT" to the baseline model, we observe that it has inferior performance compared to the baseline in the restoration from 4-8 bit and 6-8 bit. This can be explained by the relative decrease in the amount of input image information at the input layer, caused by concatenating the class probability maps along the channel direction. These results demonstrate the effectiveness of the proposed BDE method in efficiently infusing semantic contextual information through the SFT layers. To further validate these observations, we conducted paired t-tests with a significance level of 0.05 comparing our proposed model with "w/o seg" and "w/o SFT" respectively. The results showed that a statistically significant difference is only observed in the CPSNR for 4-8 bit restoration when comparing our proposed model to the "w/o seg" model. While the numerical improvements in other cases are promising, they were not found to be statistically significant.

5. Discussion

In comparison with the conventional methods, the proposed method has shown improved performance in restoring 8-bit depth. The improvements in both objective and subjective evaluations indicate that semantic information enables more context-aware BDE, particularly in challenging areas such as restoration of missing details and the suppression of false contours. However, the proposed method did not achieve the same improvements in restoring 16-bit depth. This limitation can be attributed to the greater diversity of the MIT-Adobe FiveK dataset used for 16-bit restoration, which contains numerous classes that the segmentation model used in the proposed method is unable to classify. These unseen classes result in less accurate semantic information, which in turn affects the restoration accuracy. To further verify the BDE performance on high bit-depth images and the generalization capability across diverse images, it is beneficial to use semantic segmentation models capable of handling more diverse classes and adopt training datasets containing more diverse semantic information.

We analyzed the relationship using the Pearson correlation coefficient between segmentation accuracy, measured by mIoU, and image assessment indices, CPSNR and SSIM, for restoring 8-bit depth from 3-bit, 4-bit, and 6-bit depths. A weak positive correlation was observed between mIoU and CPSNR, which gradually strengthened as bit depth increased. Specifically, the correlation coefficients between mIoU and CPSNR are 0.1951 for 3-bit, 0.2556 for 4-bit, and 0.2780 for 6-bit. These results suggest that semantic segmentation as prior information contributes positively to the BDE accuracy. In contrast, the correlation coefficients between mIoU and SSIM ex-

Table 4 Objective evaluation of ablation study

		ZP (Input)	w/o both	w/o seg	w/o SFT	Ours
3-8 bit	CPSNR	22.69 ± 0.47	33.90 ± 1.55	33.97 ± 1.57	34.07 ± 1.61	**34.13** ± 1.69
	SSIM	0.8008 ± 0.0656	0.9418 ± 0.0259	0.9429 ± 0.0249	**0.9445** ± 0.0252	0.9443 ± 0.0259
4-8 bit	CPSNR	28.75 ± 0.39	40.16 ± 2.39	39.89 ± 2.35	40.12 ± 2.35	**40.32*** ± 2.36
	SSIM	0.9172 ± 0.0360	0.9803 ± 0.0161	0.9787 ± 0.0173	0.9800 ± 0.0163	**0.9804** ± 0.0168
6-8 bit	CPSNR	40.97 ± 0.50	49.41 ± 1.78	49.31 ± 1.88	49.41 ± 1.80	**49.49** ± 1.81
	SSIM	0.9930 ± 0.0045	**0.9974** ± 0.0022	0.9973 ± 0.0021	**0.9974** ± 0.0023	**0.9974** ± 0.0025

* Indicates statistically significant difference ($p < 0.05$) compared to the "w/o seg"

hibited a different pattern, with coefficients of 0.1197, 0.0728, and -0.0032 for 3-bit, 4-bit, and 6-bit, respectively. These results suggest the possibility that the proposed model synthesizes texture structures that are not strictly faithful to the original, based on semantic contextual information obtained from the segmentation model.

The design of the loss function requires careful consideration. While the BDE task primarily focuses on low-level features due to its emphasis on local details, proposed method utilizes high-level features to consider semantic information from segmentation. To achieve this, we adopted a perceptual loss, utilizing the feature maps from the 8th layer of the VGG-19 network. However, we recognize that there is room for reconsideration in the selection of layer. It is expected that deeper layers capture higher-level features and more semantic information. On the other hand, relying on such deeper layers may reduce the ability to accurately reproduce fine textures and details.

6. Conclusion

In this paper, we proposed a novel BDE model considering the semantic contextual information by incorporating the SFT layers into the BDE model. The SFT layer considers the semantic contextual information based on the pixel-wise class probability maps of input image obtained by semantic segmentation model. In the experiments, we showed that the proposed BDE method achieves objectively superior performance and subjectively suppresses artifacts compared with existing CNN-based BDE methods for restoring 8-bit depth from 3,4, and 6-bit depths. In addition, we investigated how to provide the contextual information to the BDE model and showed that providing it through the SFT layer is effective compared with other providing methods.

In the current study, we used a segmentation model pre-trained on 8-bit images. However, by fine-tuning the segmentation learning specifically for low-bit images, it may be possible to obtain more appropriate contextual information. Furthermore, using semantic segmentation models capable of handling more diverse classes could prove beneficial. This is expected to enhance generalization performance across diverse datasets and improve results in high bit-depth restoration.

References

1) C. Peng, L. Cai, Z. Fu, X. Li: "CNN-Based Bit-Depth Enhancement by the Suppression of False Contour and Color Distortion", Proc. of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, pp. 1145–1151 (2019).
2) Y. Su, W. Sun, J. Liu, G. Zhai, P. Jing: "Photo-Realistic Image Bit-Depth Enhancement via Residual Transposed Convolutional Neural Network", Neurocomputing, Vol. 347, pp. 200–211 (2019).
3) J. Byun, K. Shim, C. Kim: "BitNet: Learning-Based Bit-Depth Expansion", Proc. of the Asian Conference on Computer Vision, pp. 67–82 (2019).
4) J. Liu, W. Sun, Y. Su, P. Jing, X. Yang: "BE-CALF: Bit-Depth Enhancement by Concatenating All Level Features of DNN", IEEE Trans. on Image Processing, Vol. 28, No. 10, pp. 4926–4940 (2019).
5) Y. Zhao, R. Wang, W. Jia, W. Zuo, X. Liu, W. Gao: "Deep Reconstruction of Least Significant Bits for Bit-Depth Expansion", IEEE Trans. on Image Processing, Vol. 28, No. 6, pp. 2847–2859 (2019).
6) Y. Zhao, R. Wang, Y. Chen, W. Jia, X. Liu, W. Gao: "Lighter but Efficient Bit-Depth Expansion Network", IEEE Trans. on Circuits and Systems for Video Technology, Vol. 31, No. 5, pp. 2063–2069 (2021).
7) A. Punnappurath, M. S. Brown: "A Little Bit More: Bitplane-Wise Bit-Depth Recovery", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 44, No. 12, pp. 9718–9724 (2022).
8) Y. Liu, Q. Jia, J. Zhang, X. Fan, S. Wang, S. Ma, W. Gao: "Learning Weighting Map for Bit-Depth Expansion within a Rational Range", arXiv preprint arXiv:2204.12039 (2022).
9) X. Huang, S. Belongie: "Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization", Proc of the IEEE International Conference on Computer Vision, pp. 1501–1510 (2017).
10) E. Perez, F. Strub, H. de Vries, V. Dumoulin, A. C. Courville: "FiLM: Visual Reasoning with a General Conditioning Layer", Proc. of the AAAI Conference on Artificial Intelligence (2017).
11) X. Wang, K. Yu, C. Dong, C. C. Loy: "Recovering Realistic Texture in Image Super-Resolution by Deep Spatial Feature Transform", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 606–615 (2018).
12) R. A. Ulichney, S. Cheung: "Pixel Bit-Depth Increase by Bit Replication", Proc. of Society of Photo-Optical Instrumentation Engineers, Vol. 3300, pp. 232–241 (1998).
13) C.-H. Cheng, O. C. Au, C.-H. Liu, K.-Y. Yip: "Bit-Depth Expansion by Contour Region Reconstruction", Proc. of the IEEE International Symposium on Circuits and Systems, pp. 944–947 (2009).
14) P. Wan, O. C. Au, K. Tang, Y. Guo, L. Fang: "From 2D Extrapolation to 1D Interpolation: Content Adaptive Image Bit-Depth Expansion", Proc. of the IEEE International Conference on Multimedia and Expo, pp. 170–175 (2012).
15) J. Liu, G. Zhai, A. Liu, X. Yang, X. Zhao, C. W. Chen: "IPAD: Intensity Potential for Adaptive De-Quantization", IEEE Trans. on Image Processing, Vol. 27, No. 10, pp. 4860–4872 (2018).
16) K. He, X. Zhang, S. Ren, J. Sun: "Deep Residual Learning for Image Recognition", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016).
17) A. L. Maas: "Rectifier Nonlinearities Improve Neural Network Acoustic Models", Proc. of the International Conference on Machine Learning, Vol. 30, No. 1, p. 3 (2013).
18) J. Liu, W. Sun, Y. Liu: "Bit-Depth Enhancement via Convolutional Neural Network", Proc. of the International Conference on Digital TV and Wireless Multimedia Communication, pp.

255–264 (2018).

19) J. Johnson, A. Alahi, L. Fei-Fei: "Perceptual Losses for Real-Time Style Transfer and Super-Resolution", Proc. of the European Conference on Computer Vision, pp. 694–711 (2016).
20) D. J. Butler, J. Wulff, G. B. Stanley, M. J. Black: "A Naturalistic Open Source Movie for Optical Flow Evaluation", Proc. of the European Conference on Computer Vision, pp. 611–625 (2012).
21) V. Bychkovsky, S. Paris, E. Chan, F. Durand: "Learning Photographic Global Tonal Adjustment with a Database of Input / Output Image Pairs", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp.97—104, (2011)
22) B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, A. Torralba: "Scene Parsing through ADE20K Dataset", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5122–5130 (2017).
23) T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick: "Microsoft COCO: Common Objects in Context", Proc. of the European Conference on Computer Vision, pp. 740–755 (2014).
24) D. P. Kingma, J. Ba: "Adam: A Method for Stochastic Optimization", Proceedings of the International Conference on Learning Representations (2015).
25) Z. Liu, X. Li, P. Luo, C. C. Loy, X. Tang: "Deep Learning Markov Random Field for Semantic Segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 40, pp. 1814–1828 (2016).

(Received April 16, 2024)
(Revised October 5, 2024)

Taishi IRIYAMA (*Member*)
He received the B.E., M.E. and Ph.D. degrees in electronic information engineering from Tamagawa University, Tokyo, Japan, in 2017, 2019 and 2022, respectively. Currently, he is assistant professor of mathematics, electronics and informatics at Saitama University.

Yuki WATANABE
He recieved the B.E. degree in information and conmputer sciences from the Saitama University, Saitama, Japan, in 2023. He is currently working as a software engineer at DTS corporation.

Takashi KOMURO
He received the B.E., M.E., and Ph.D. degrees in mathematical engineering and information physics from the University of Tokyo in 1996, 1998, and 2001, respectively. At present he is a professor of mathematics, electronics and informatics at Saitama University.

IIEEJ Survey Paper

Corporate Efforts for R&D on Video Coding and Its Practical Implementation (Part–1) : R&D Evolution from Delta Modulation through H.320/H.261

Toshio KOGA *(Fellow)* †

† NEC Corporation (1971 - 2000) and Yamagata University (2000 - 2012)

<Summary> Digital video signal processing for varieties of purposes including storage as well as transmission is currently playing a very important role in our daily life. NEC Corporation has been actively doing R&D on video coding since its dawn and incessantly making every effort for implementation of practical codecs, terminals, and systems. They include Delta Modulation for digitization, HO-DPCM for high quality TV transmission, and interframe coding for a broad spectrum of audiovisual services. The author believes it is quite beneficial to introduce a combined history of coding algorithm improvements and practical codecs based on them, since they will provide the overview of the technical history from theoretical and also practical aspects, and definitely there is no such review published so far. This survey paper consists of two parts and reviews the corporate R&D efforts with respect to video coding algorithm improvements and our contribution to progress in digital TV/video transmission services over the world for three decades since the mid-1960s. Part -1 (Chapter 1and 2) mainly handles continued improvement of Delta Modulation and coding algorithms on intraframe and interframe prediction as well as entropy coding. Part - 2 (from Chapter 3 to 6) mainly shows effectiveness and significance of video coding through various application examples of codecs/terminals as well as standardization activities in which we were deeply involved. In addition, Part-2 includes a brief history of NEC's CODECs in conjunction with continued algorithm improvements and several examples of practical use in businesses.

Keywords: video coding, NTSC signal, TDM signal, adaptive prediction, motion compensation, entropy coding, H.120/Part 3, H.320/H.261, interoperability test

1. Introduction

Pulse Code Modulation (PCM) was invented by A. H. Reeves in 1937 and theoretically established by C. E. Shannon in 1948. It is the most fundamental digital means for representation of every kind of information and for transmission as well as storage. First PCM transmission experiment was conducted by AT&T in 1949, using 4 GHz radio links.

In parallel with the progress in digitization of network facilities, signal processing was also going digital, targeting at telephone at first then at video signals. In the 1960s, Delta Modulation (ΔM) was a unique and a realistic possible solution for digital video communication. In the 1970s, research activities on video coding quickly shift to interframe coding worldwide[1]. NEC also started R&D activities on interframe coding at the same time and in parallel extended works on ΔM and intraframe coding to NTSC Color TV signal. Early in the 1980s, Motion Compensation in real-time became possible and its improvement was discussed extensively[2]. Along with the development of coding algorithms, necessity of international standardization was recognized worldwide in the mid-1980s. Meanwhile, VLSIs were deployed in manufacturing codecs[3].

In early days of progress in digital technology represented by networks and computers, NEC advocated "Declaration of C & C ", a concept of "fusion of computers and communications technologies." The Declaration told us that "early in the 21st century, it will be possible to talk and see between any persons, at any time, at any place on earth, and that this will require an integration of technologies for communications, computers, and television." Our efforts for R&D on video coding and its practical implementation are in harmony with the Declaration.

2. R&D Activities on Video Coding in NEC

Digitization is an entrance to digital video coding and transmission. However, Analog-to-Digital (A/D) converters for video signals were not realistic in the 1960s. It led us to study ΔM for videophone signals. When A/D converters were commercially available in the 1970s, predictive coding for NTSC TV signal, both intraframe and interframe, was included in our objectives.

Once digitized at a certain sampling frequency, video signals are coded by appropriate algorithms. Basically, video source coding consists of decorrelation, quantization, and entropy coding. Prediction, transform coding, and their combination are major means for the decorrelation. Prediction can be improved by changing prediction functions adaptively corresponding to local properties in input signals. Quantization reduces the number of levels or their equivalents resulting from the decorrelation. Entropy coding is applied to the quantization

output, coding parameters, house-keeping information, etc. There are two major means for increasing or decreasing an amount of coded information, one being quantization characteristics and the other sampling frequency. We studied all these items above and they are described in what follows.

2.1 Digitization of TV/video signals by Delta Modulation

If direct digitization of black-and-white video signal or NTSC Color TV signal by Delta Modulation (ΔM) is assumed, a very high sampling frequency is necessary. An expected solution was to lower the frequency so that commercially available components could be used. Our R&D activities on ΔM for black-and-white videophone signals started in the mid-1960s and continued extensively to ΔM for NTSC signal in the early 1970s.

2.1.1 ΔM for black-and-white videophone signals

The study on the transmission of video signals by ΔM had been done. Picture quality degradation intrinsic to ΔM had been analyzed and its countermeasure discussed so far. Typically, false contouring is visible in flat areas, while blurriness and edge busyness appear on edges with sharp brightness change. In 1969, we developed a ΔM CODEC [4] for 1 MHz black-and-white videophone signals with a sampling frequency (*fs*) at 8 MHz, i.e., equivalent to 8 Mbit/sec. Asymmetrical DC Off-Set was found very effective in reducing false contouring by using this codec. Edge busyness was also reduced based on an analytical study[5].

ΔM to DPCM conversion

In general, DPCM is better than ΔM with respect to coding efficiency. It needs digitization of input signals at first. As an alternative for expensive A/D converters, a digital filter was used in combination with ΔM to generate DPCM signals.

If Single Integration ΔM (SI-ΔM) is used to generate directly PCM signals, a required frequency (*fs*) is as high as 160 MHz. Therefore, it is quite important to select *fs* as low as possible. As a means for lowering the sampling frequency, we proposed an idea of applying a sharp-cut-off filter to out-of-band frequency components in Double Integration ΔM (DI-ΔM) output [6]. This resulted in reduction of the sampling frequency to 16 MHz for ΔM and to 2 MHz for DPCM, respectively. The linear DPCM output can be easily converted to four-bit nonlinear DPCM, which corresponds to 8 Mbit/sec.

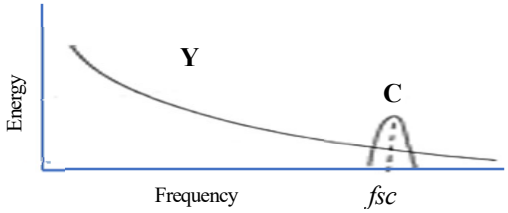


Fig.1 Frequency spectrum of NTSC color TV signal

2.1.2 ΔM for NTSC Color TV signal

Efforts were also made for coding NTSC Color TV signal by ΔM. NTSC signal had been used in U.S., Japan, and many other countries for longer than half a century. NTSC signal with a nominal bandwidth of 4.2 MHz is an analog composite one in which color subcarrier *fsc* (3.58 MHz) is modulated by two color components (R-Y, B-Y) and multiplexed in the vicinity of the subcarrier. As a result, interference between luminance (Y) and chrominance (C) is minimized, as shown briefly in **Fig.1**.

In coding, a de-emphasis filter is applied to suppress color S/N degradation by spectrum shaping of ΔM quantization error. A resonance circuit is implemented as the de-emphasis filter in the second integrator of a DI-ΔM coder. In decoding, a regular SI-ΔM decoder is used and the decoded signal is applied further to an emphasis filter to reproduce the NTSC signal. Based on this idea, what may be called a Higher-Order ΔM CODEC was implemented[7] in 1975. This research study was extensively succeeded by Higher-Order intraframe coding (HO-DPCM).

2.2 Predictive coding of color TV signals

The principle of predictive coding is shown in **Fig.2**. P(*z*) is a prediction function. Prediction error (e) in Encoder is difference between Input signal (X) and Predictor output. It is quantized (ê) and transmitted to Decoder after Entropy Coding. Quantization is applied to the prediction error to reduce the number of allowable levels or their equivalents. This greatly helps reduce data amount. At Decoder, the quantized prediction error (ê) is added to Predictor output to reproduce the video signal ($\hat{X}$).

There are three kinds of correlation in video signals, i.e., horizontal, vertical, and temporal. When a prediction function P(*z*) consists of a sample memory working as a pixel delay, it means that horizontal correlation is used. It is the most fundamental prediction for black-and-white video signals. When a line memory is available, vertical correlation can be used. In addition, when a frame memory is available, temporal correlation can be also used.

There are two choices in handling color TV/video signals depending on applications, one for high quality transmission such as CATV or TV broadcasting, and the other for wider

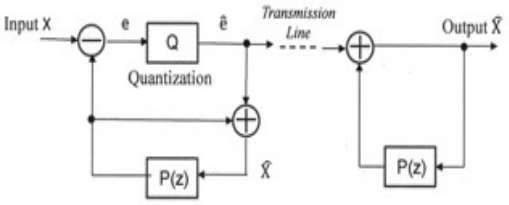


Fig.2 Predictive coding principle

applications such as audiovisual services, image information retrieval, distance learning, etc. It is very important to handle color signals appropriately, taking applications into account.

2.2.1 Composite coding for TV program broadcasting

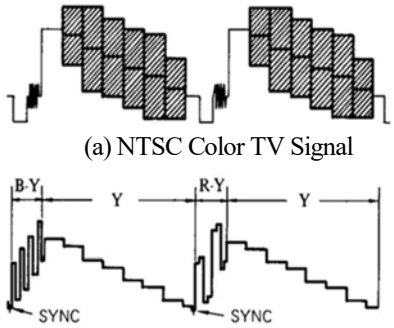
As an example of NTSC Color TV signal, two successive lines of Color Bar test signal are shown in **Fig. 3 (a)**. Each line consists of horizontal sync, color subcarrier, and a frequency-multiplexed video part. The subcarrier phase is inverted alternately line by line. Therefore, color phase of the pixels should be fully taken care of in prediction, intraframe or interframe.

(1) Composite intraframe prediction and sampling frequency

Relationship between luminance (Y) and chrominance (C) in NTSC signal is briefly shown in **Fig. 4** for video parts of two successive lines, when it is sampled at 3×*fsc*. In this example, the same color phase is seen every three samples. This leads to choosing a three-previous sample for prediction, i.e., third-order prediction, which can be described as $P_b(z) = z^{-3}$ in the z-transform notation. This is appropriate for predicting chrominance pixels. Frequency response $H_b(z)$ of this prediction is given by $H_b(z) = 1 - z^{-3}$. If a term $(1 - 0.5z^{-1})$ is added to improve luminance prediction[8], the response results in Eq.(1).

$$H_c(z) = (1 - 0.5z^{-1})(1 - z^{-3}) \qquad (1)$$

For reference, let us add $H_a(z)$, a frequency response for previous sample prediction. It is described as $H_a(z) = 1 - z^{-1}$.



(a) NTSC Color TV Signal

(b) TDM Color TV Signal

Fig.3 Waveform examples of two Color TV formats

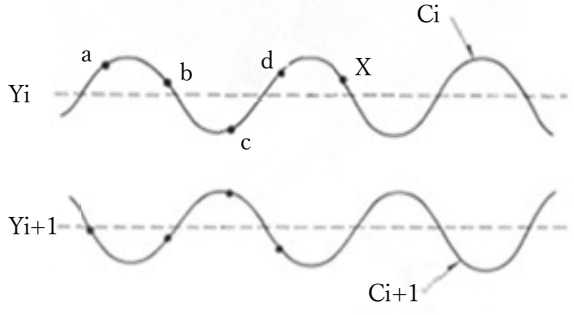


Fig.4 Color phase inversion between two successive lines

These three responses are compared in **Fig. 5,** where the vertical axis corresponds to amplitude and the horizontal one to normalized frequency (f/fs). In addition to the results at f/fs=0, responses for $H_b(z)$ and $H_c(z)$ are also zero at f/fs=1/3, i.e., zero at $f = fsc$, while not for $H_a(z)$. When $H_c(z)$ is compared with $H_b(z)$ between 0 and 2 in the normalized frequency, it shows apparently smaller amplitude values than $H_b(z)$. That is, $H_c(z)$ is much better, and therefore it is used as a basis in HO-DPCM 45A series codecs.

Sampling frequency selection

For many years, a sampling frequency for composite TV signals was usually bound by the subcarrier frequency, such as multiple integer times *fsc*, typically 3×*fsc* or 4×*fsc* in NTSC. Both meet the Nyquist rate which is defined as twice the signal frequency bandwidth. There are applications such as CATV which may not necessarily require highest quality. If a lower sampling frequency is allowed, data amount reduction becomes much easier and it helps produce inexpensive codecs, while coded video quality should be maintained at an acceptable level.

As an example, let us choose *fs* to be 2×*fsc*, i.e., 7.2 MHz. It is a little bit lower than the Nyquist rate and therefore called sub-Nyquist rate sampling. In this case, it is very important to avoid or suppress fold-over effect, a kind of interference, between luminance and chrominance. In addition to the frequency-interleave sampling, a comb filter is applied to suppress luminance existing in high frequency region dominated by chrominance. The sub-Nyquist sampling was deployed in HO-DPCM 32A in 1975 for 32 Mbit/sec transmission[9]. Frequency response $H(z)$ in this case is given by Eq.(2).

$$H(z) = (1 - 0.5z^{-1})(1 - z^{-2}) \qquad (2)$$

It is very flexible if the sampling frequency can be chosen freely or with less constraint. Generalization or relaxation of constraint on the sampling frequency selection was also studied[10]. Using a parameter "α", let us rewrite $H_c(z)$ with a slight change and obtain $H_d(z)$ defined by the following equation,

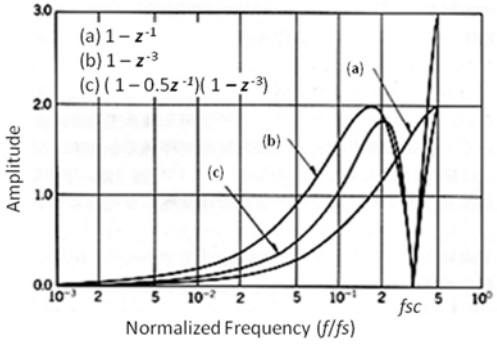


Fig.5 Frequency response of Higher-Order DPCM

$H_d(z) = (1- 0.5\, z^{-1})\,(1- z^{-1})\,(1+ (1+\alpha)\, z^{-1} + z^{-2})$.

What is most important in the third term of $H_d(z)$ is to find relation between α and a sampling frequency *fs* on the condition that the frequency response is zero at $f = fsc$.

Consequently, a simple relation is obtained as follows.

$$1 + \alpha + 2\cos(2\pi\, fsc / fs) = 0,$$

therefore, $fs = 2\pi\, fsc / \cos^{-1}(-(1+\alpha)/2)$.

As an example, let us choose *fs* to be 8.8 MHz. An appropriate value of α is calculated to be 0.65896. If it is approximated to be 0.65625, it is expressed by (1/2 + 1/8 + 1/32), quite suitable for binary implementation.

This somewhat generalized sampling frequency of 8.8 MHz was deployed in HODPCM-45B in 1985, designed for industrial use at 45 Mbit/sec.

(2) Composite interframe prediction with reversible Y/C separation

In the early 1970s, it was not clearly known whether interframe prediction could realize transmission of high quality composite color TV signals such as NTSC. It was mainly because separation of luminance (Y) and chrominance(C) is likely to degrade color quality, and it may be perceptible after their synthesis back to NTSC. To cope with the difficulty, we proposed an idea that linear (reversible) transform is applied for color phase adjustment before interframe prediction[11]. Color subcarrier phase is alternately inverted line by line as shown in Fig. 4. The same is true between two successive frames since there are 525 lines in a frame. That's why the subcarrier phase should further be taken into account in interframe prediction.

As an example, a pair of lines in NTSC in Fig. 3(a), L_{2m} and L_{2m+1}, are applied to the following equation Eq.(3) to produce another pair of lines, Y_m and C_m, each corresponding to luminance and color, respectively. The equation is a kind of Orthogonal Transform (OTF) and equivalent to Hadamard Transform of the 2nd order.

$$\begin{bmatrix} Y_m \\ C_m \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ (-1)^n & -(-1)^n \end{bmatrix}\cdot\begin{bmatrix} L_{2m} \\ L_{2m+1} \end{bmatrix} \quad (3)$$

A term $(-1)^n$ is applied so that subcarrier phase is inverted every other frame for phase adjustment between frames, where 'n' can be a frame number or any appropriate serial integer number. Once the subcarrier phase is adjusted between frames, interframe prediction can be applied.

At Decoder, subcarrier phase restoration is made using the following Eq.(4) to reproduce NTSC signal after interframe decoding. The term $(-1)^n$ is applied so that subcarrier phase is inverted every other frame and then the initial phase relation at Encoder is restored.

$$\begin{bmatrix} L_{2m} \\ L_{2m+1} \end{bmatrix} = \begin{bmatrix} 1 & (-1)^n \\ 1 & -(-1)^n \end{bmatrix}\cdot\begin{bmatrix} Y_m \\ C_m \end{bmatrix} \quad (4)$$

With this adjustment, interframe prediction can be applied. Intraframe prediction was applied to interframe difference to cope with abrupt changes likely to occur between frames. This is called Combinational Difference prediction. Third-order prediction (z^{-3}) can be used as intraframe prediction when the sampling frequency is 3×*fsc*. When data amount increases abruptly, sub-Nyquist sampling mode (2×*fsc*) is evoked to suppress the increase and corresponding Higher-Order intraframe prediction ($0.5z^{-1} + z^{-2} - 0.5\, z^{-3}$) is used without interframe prediction.

This algorithm was implemented in NETEC-22H(prototype). Its Encoder (left) and Decoder (right) are shown in **Fig. 6**. A paper on NETEC-22H was presented at IEEE National Telecommunications Conference (currently, GLOBECOM) in 1976 to show effectiveness of digital transmission of high quality broadcast TV signals[11]. Fortunately, "1976 NTC Best Paper Award" was given to this presentation as shown in **Fig.7**.

2.2.2 Component interframe coding for wide applications

Economical transmission is more important in many applications than highest quality as required in TV broadcasting. Component coding can be a better solution, since it allows many more techniques for improvement than composite one, and lower rate transmission as well.

(1) Video format conversion to Line-sequential TDM signal

NTSC Color TV format exemplified in Fig. 3 (a) is converted to another one in which luminance (Y) and two color components (R-Y, B-Y) are separated at first and then rearranged in time slots as shown in Fig. 3 (b). In this example, the two color components (R-Y, B-Y) are subsampled, and each component after grouping is placed alternately line-by-line in Horizontal Sync parts. Then, interframe coding can be applied as if input signals were black-and-white.

This digital format is named Line-sequential Time Division Multiplexed (TDM) Color signal[12]. An active frame of this TDM signal consists of 480 lines/frame and 510 pixels/line. Each line includes 6 (sync) + 84 (C1/C2) + 420 (Y) in pixels.

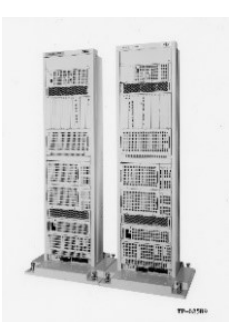


Fig.6 NETEC-22H

Fig.7 NTC Best Paper Award

Color components C1 and C2 correspond to (R-Y) and (B-Y), respectively, and they are 5-to-1 subsampled.

(2) Implementation of interframe codec for TDM signal

When NEC started R&D on interframe coding in 1971, our target was transmission of 1-MHz black-and-white videophone signals at 1.5 Mbit/sec. Soon, the target was raised to transmission of 4-MHz Color TV signal at 6.3 Mbit/sec. NETEC-6/16 was developed in 1974 based on interframe prediction applied to the TDM Color signal[13]. It was the first practical interframe codec ever developed in the world. Our R&D efforts for further improvement continued since then.

2.3 Adaptive prediction

Any adaptive prediction makes use of more than one prediction functions, expecting improvement. There are three kinds of correlation in video signals, i.e., horizontal, vertical, and temp-oral. When these three are appropriately used in combination, it will result in higher coding efficiency.

The first adaptive algorithm proposed by R. E. Graham in 1958 was based only on intraframe prediction, using previous sample (horizontal) and previous line prediction (vertical) [14]. That is, Graham's algorithm is an adaptive prediction in spatial domain. In the 1970s, an interframe coding algorithm was proposed by J. C. Candy et al. [15] , and its improvement was carried out in many places over the world [1]. Essence of interframe prediction lies in the third correlation, i.e., temporal one. Adaptive algorithms combining spatial and temporal correlation had been also very important research items. We have studied the adaptive algorithms intensively since then, particularly in the1980s.

2.3.1 Pixel-based adaptation

A basic pixel-based adaptation scheme without selection information transmission is depicted in **Fig. 8**. Here, two prediction functions, P1 and P2, are assumed and one of the two estimated to be better is selected for prediction at the next sample time. What can be used in the estimation is everything, so long as it is available both at Encoder and Decoder, i.e., P1(i), P2(i), $\hat{x}(i)$, and $\hat{e}(i)$ in Fig. 8. This is essential, because selection information need to be transmitted unless otherwise. What is also important in this scheme is that the actual selection is done at least one sample time later than the estimation. Therefore, a sample memory or delay "D" is quite essential for this purpose.

Two types of adaptive prediction are possible in this scheme, i.e., Type I using solely $\hat{e}(i)$, while Type II using P1(i), P2(i), and $\hat{x}(i)$. Here, let us assume P1(i) and P2(i) represent prediction values for P1 and P2 at a time (i), respectively. P(i) is an output value of the selected prediction. Selection at SW is carried out in obedience to the estimation result J(i-1).

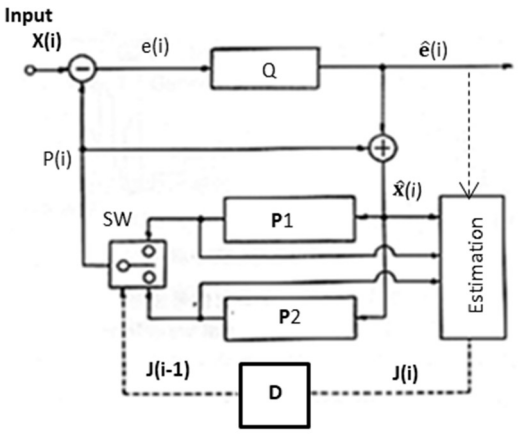


Fig.8 Basic scheme for Pixel-based Adaptive Prediction

Type I: Comparison with threshold value

Only quantized prediction error $\hat{e}(i)$ is used. Let us assume that P1(i) is selected at a sample time (i). If $Abs[\hat{e}(i)] \leq Th$, then P(i+1) = P1(i+1), else P(i+1) = P2(i+1). The "Th" value, fixed or variable, should be identical in both Encoder and Decoder at any instance. If "Th" is variable, an appropriate algorithm is necessary to find its optimum or sub-optimum value, which should work always in synchronization on both sides.

Type II : Comparison of difference values

Two prediction values given by P1(i) and P2(i) plus locally decoded sample $\hat{x}(i)$ are used, where $\hat{x}(i) = P(i) + \hat{e}(i)$.

Better prediction is estimated by the following comparison. If $Abs[P1(i) - \hat{x}(i)] \leq Abs[P2(i) - \hat{x}(i)]$, then P(i+1) = P1(i+1), else P(i+1) = P2(i+1) . Here, Abs[•]means an absolute value.

Let us call this type "Two-diff" in what follows, since two difference values are compared for the estimation. The number of prediction functions can be extended to more than two.

A theoretical analysis was made on Type I adaptive algorithm in terms of prediction error entropy, based on a mathematical model that both intraframe and interframe prediction error signals can be represented by the Exponential Distribution. According to the analysis, Type I is shown to be as effective as interframe prediction for pictures moving at a speed of 1 pixel/frame or lower. At faster speeds than 1 pixel/frame, Type I is as good as intraframe prediction. Therefore, Type I is more efficient than Combinational Difference prediction. This is confirmed theoretically as well as experimentally [16].

Change in prediction function in Type I takes place when the prediction error value exceeds "Th", regardless of which prediction was used at the time. That is, exceeding "Th" shows that the prediction did not work well but it does not necessarily mean the other was better. As for Type II, two difference values are compared, one for P1 and the other for P2. When either difference value is smaller, its corresponding prediction is likely

to be relevant. The relevance information in the coded pixel parts in two dimensions helps improve coding efficiency, since spatial correlation can be used. As a whole, Type II can be preferred for its algorithmic simplicity as well as for performance stability.

(1) Composite intraframe coding

Combining predictions based on horizontal and vertical correlation can be a good basis for adaptative algorithms in intraframe coding. In addition, Type II adaptation was adopted in our composite intraframe coding algorithms.

Here, let us define the basic prediction, Higher-Order DPCM (HO-DPCM), as $(0.5z^{-1}+z^{-3}-0.5z^{-4})$ in the case of $fs=3\times fsc$. Its frequency response is denoted by $Hc(z)$ and shown in Fig. 5. In "Two-diff" for composite coding, HO-DPCM can be used as P1 and z^{-2H} as P2. The latter is two-previous line prediction in which the color subcarrier phase is identical (see **Table 1**).

If we assume that vertical distance is "two lines" in the two-previous line prediction, then the distance in interfield prediction (z^{-262H}) is a quarter of two lines, i.e., "a half line." In general, the smaller the distance, the higher the correlation, resulting in better prediction, so long as the subcarrier phase is identical between lines of interest. Therefore, it is expected that interfield prediction (z^{-262H}) can improve coding efficiency. After confirming that interfield prediction improves efficiency by 4 ~ 5 %, it was added to the "Two-diff", resulting in "Three-diff". It was implemented as P3 in HO-DPCM 45A in 1982 [17]. "Four-diff" is an improved algorithm for "Three-diff," i.e., previous sample prediction (z^{-1}) is further added as P4 so that it can show excellent performance even when black-and-white signals are included in input signals. It was implemented in Broadcaster 45 in 1992 for use at DS-3 rate in U.S., and also in Broadcaster 52 in 1993 for use at 52 Mbit/sec in Synchronous Digital Hierarchy networks. It is also possible to include two-previous frame prediction[18] for the highest coding efficiency in still background parts since the subcarrier phase is identical. However, it may better be classified in interframe prediction. Type II adaptive prediction algorithms based on HO-DPCM are summarized in Table 1, in which HO-DPCM CODECs are shown at the bottom.

Table 1 HO-DPCM-based Adaptive Prediction

Scheme	Fixed	Two-diff	Three-diff	Four-diff
P1	HO-DPCM	HO-DPCM	HO-DPCM	HO-DPCM
P2	-	z^{-2H}	z^{-2H}	z^{-2H}
P3	-	-	z^{-262H}	z^{-262H}
P4	-	-	-	z^{-1}
CODEC	Expr ('73)	Proto ('80)	45A ('82) ~ 45AIII ('88)	B-45 ('92) B-52 ('93)

(2) Composite interframe coding

The prototype composite interframe codec was based on Combinational Difference prediction. However, it was theoretically clarified that an adaptive prediction based on Type I is more effective than that for a broad range of "Th" values[16]. Therefore, Type I was implemented with a preselected fixed value for "Th" in NETEC-22H (Product) with other functions unchanged from the prototype.

(3) Component interframe coding

Our first interframe codec, NETEC-6/16, was based on simple interframe prediction for TDM color signals. To improve coding efficiency, Type I was compared with "Two-diff" in Type II, in which P1 was interframe and P2 previous sample prediction. As a result, "Two-diff" in Type II had been employed in NETEC-series products until Recommendation H.320/H.261 was issued.

2.3.2 Block-based adaptation in component coding

In addition to pixel-based adaptation, we can also mention block-based one. Block-matching type motion compensation is the most typical candidate.

(1) Block-based motion compensation (MC)

Motion compensation (MC) has long been expected for higher performance in video coding [2]. MC can be realized in two ways, i.e., pixel-basis or block-basis. Both methods had been studied in many places worldwide in the 1970s through 1980s. However, to the author's best knowledge, nothing but block-based MC has been implemented so far in practice.

Block Matching Algorithm (BMA) is the most typical block-based method and a kind of pattern matching between two 2-dimensional blocks. A block (Block A) in the current frame is compared with a block (Block B_k) which is displaced by a candidate vector ($\vec{V}_k$). The comparison is repeated with respect to many candidate vectors within an MC search range in the previous frame. A parameter SUM ($\vec{V}_k$) is defined by Eq.(5) as a measure of similarity between Block A and Block B_k. $C[a(i,j) - b_k(i,j)]$ in Eq.(5) is a cost function for the difference, where a(i, j) is a pixel in Block A and $b_k(i, j)$ in Block B_k, respectively. Assuming that (e) is the difference, C[e] outputs either one of Abs[e], e^2, word-length of a VLC for (e), etc. Let us also assume here that each block consists of M lines × N pixels.

$$\mathrm{SUM}(\vec{V}_k) = \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} C[a(i,j) - b_k(i,j)] \qquad (5)$$

The smaller the SUM value, the higher the similarity. The relative position or displacement showing the highest similarity is chosen as a motion vector ($\vec{V}_{opt}$). BMA has been shown to be

very effective so that it can reduce information amount approximately to 1/2 - 1/3 for videoconference signals[19].

(2) Pixel-based enhancement to MC prediction

MC does not necessarily work well when temporal correlation is poor, e.g., large objects moving very rapidly or beyond MC range as well as scene change in the worst case. To cope with this difficulty, spatial correlation also helps. "Two-diff" type prediction combining previous sample prediction (z^{-1}) and MC prediction (z^{-MC}) is still helpful. However, there is room for improvement. "Two-diff" type requires simple comparison of two difference values to estimate better prediction. Instead of the comparison, a kind of look-up table can play a similar role. The table consists of logical calculation results for several past sample points in two dimensions, which are obtained from statistics on many videoconference signals. That is, the table outputs an estimation result taking spatial correlation into account. It was implemented in NETEC-X1MC in 1983 [20].

(3) MC combined with DCT

As far as BMA is used, it is natural to combine MC and 2D-DCT, since the latter is also block-based. It was shown by N. Ahmed et al. (Univ. of Texas) that 2D-DCT is a very effective tool in image coding[21]. It is more efficient than spatial prediction such as previous sample and/or previous line prediction. When MC does not work well, 2D-DCT is expected to help MC instead of spatial prediction. On the contrary, when MC does, it may not be necessary but does not degrade efficiency, either. Consequently, they can be combined in a fixed manner. This combination was first implemented in VL-3000 codec in 1989.

2.4 Entropy coding

When NETEC-6/16 was developed in 1974, prediction error was almost all the information to be transmitted and represented by two groups of fixed-length code sets, short (FLC1) and long (FLC2). It may look far from entropy coding, but practical implementation of variable-length code set (VLC) was not possible at the time, mainly because of insufficient ability of hardware, particularly in memory capacity of Programmable ROMs.

2.4.1 Code conversion of prediction error information

In general, there are two types of quantization characteristics, mid-riser usually used in intraframe and mid-tread in interframe prediction. There is no zero output in the mid-riser type, while very small prediction error is suppressed to zero in the mid-tread type. Effective representation of the quantized error information is quite important and executed by Entropy Coding.

(1) VLC for intraframe prediction error

An experimental HO-DPCM encoder was developed in 1976[22], equipped with 22-level mid-riser type nonlinear quantization characteristics. Its VLC set consists of codewords, each being 2, 3, 5, 6, 8, and 9 bits including a sign bit.

According to a study on appropriate maximum length of VLC codewords, it was clarified that average code lengths were almost equal to theoretical entropy values, when the maximum length was about 14 or 15 bits[17]. This study shows that an optimum VLC code set can be designed with a maximum code length of 15 bits. The result was reflected in successors such as HO-DPCM 45AII and thereafter.

(2) Adaptive VLC for interframe prediction error

Interframe difference is usually very small in background parts. Therefore, there are many zeros appearing when the mid-tread type quantization is applied. Effective representation of the zero information is quite important.

Representation of zero prediction error

Basically, there are two representation methods for this purpose, one being block-based and the other pixel-based.

If a block includes non-zero pixelwise quantized error, the block is classified as significant. Positions of the significant blocks can be easily represented in many ways. The pixelwise quantized error, zero or non-zero, in the significant blocks is coded with VLC or FLC. A sequence of zeros is called a run in pixel-based representation. The run can be very long so that it covers a whole horizontal line. The most popular Run-length Coding (RLC) method is Modified Huffman (MH) which is an international standard for Facsimile image coding. In the MH scheme, a multiple of 64 zeros is expressed with a Make-up (MK) code (j) and a run shorter than 64 is with a Terminating code (Y). Any run-length (R) can be expressed as follows.

$$R = 64 \times j + Y + 1 \quad (63 \geq Y \geq 0)$$

Adaptive conversion of non-zero prediction error

As for non-zero error, it is beneficial to avoid successive appearance of long VLC codewords. An adaptive conversion by a switching scheme between VLC(V0~V28) and FLC(F0~F28) is introduced. If a VLC codeword selected exceeds a specified length, then FLC is used at the next sample time. On the contrary, if an FLC one selected corresponds to a VLC one shorter than the specified length, then switched to VLC at the next sample time. This is shown in the upper half of **Fig.9**. A combination of significant/insignificant block addressing and this adaptive VLC/FLC code conversion in the significant blocks was implemented in several NETEC CODECs.

Adaptive transition among FLC, VLC, and RLC

Furthermore, it may be better to incorporate RLC with the adaptive VLC/FLC switching scheme. If a VLC codeword (V0) for zero is found, transition to RLC takes place. If End-of-Run is

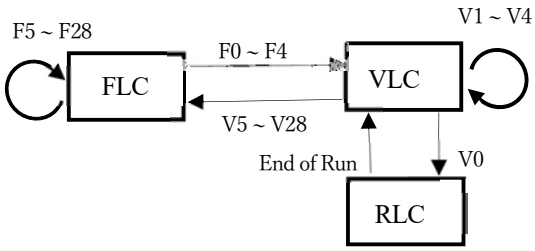


Fig.9 Transition among VLC, FLC, and RLC

found, transition back to VLC. This transition scheme among VLC, FLC, and RLC is shown in Fig. 9. It was implemented in NETEC-X1MC[23].

2.4.2 VLC for DCT coefficients [24]

Usually, there are many zero coefficients in high frequency parts after 2D-DCT is applied to video signals. Therefore, coding for zero coefficient runs is quite effective when diagonal zigzag scanning is applied, in which every coefficient is scanned from DC coefficient (upper-left corner) to the end (lower-right) across the diagonal axis in every block. Information amount for nonzero coefficients is the same regardless of scanning methods. The following four methods were compared in designing VLC code set.

(1) Basic Method (Reference)

Diagonal zigzag scanning is applied to all the coefficients within a block. Run-length coding is used for successive zero coefficients, while VLC for nonzero ones.

(2) Zone coding

"Zone" is a minimum rectangle including DC coefficient and all nonzero coefficients. There are 64 kinds of zone shape with a minimum size of 1 × 1 while a maximum size of 8 × 8. Zigzag scanning is assumed within this zone. Zone shape information must be transmitted.

(3) Scan Length coding

Scan Length is the number of coefficients from the beginning to the last nonzero coefficient and is used to indicate that the encoding has finished for the block. This is transmitted instead of the last zero run.

(4) End of Block coding

EOB (end of block) code is added immediately after the last nonzero coefficients to indicate that encoding is finished in the block. This is transmitted instead of the last zero run.

With respect to entropy values for the four methods above, Zone Coding shows the highest efficiency, although difference among them is small. Therefore, End of Block (EOB) coding is preferred since it is very simple and easy to implement. It was implemented in VisuaLink-3000 and included later in H.261 Specification.

2.4.3 VLC for motion vector information [24]

Basically, motion vector is two-dimensional and represented by two parameters, *Vx* (pels/frame) in horizontal and *Vy* (lines/frame) in vertical direction, respectively. When MC range covers an area in ± *H* samples and ± *V* lines, the number of motion vectors in the range is given by $(2H+1) \times (2V+1)$, e.g., 225 for $H = V = 7$. In a two-dimensional expression (*2D*), a single codeword is assigned to a combination of *Vx* and *Vy* , resulting in 225 codewords needed to specify a motion vector within this range. For the sake of simplicity, two kinds of approximation can be mentioned using one-dimensional expressions. In the first one-dimensional expression (*1D-A*), two independent code sets are prepared, one for *Vx* and the other for *Vy,* respectively. The former code set is designed from the statistics of *Vx* and the latter from that of *Vy*. Each code set consists of 15 codewords in this example. In the second one-dimensional expression (*1D-B*), the two components *Vx* and *Vy* are put together to produce a combined distribution irrespective of directions. Then, a single code set consisting of 15 codewords is derived from the distribution and used in common for encoding both *Vx* and *Vy*.

There are two choices in coding motion vector information i.e., coding is applied directly to vectors or to differential ones between two adjacent blocks. In practice, there is very little difference between the two choices for video signals with normal motion. However, considerable improvement can be expected of difference vectors when translational motion covers many blocks or camera is panned.

All the three VLC methods above were implemented for motion vectors in practice, although applied to difference vectors. That is, (*2D*) was deployed in NETEC-X1MC[23] while (*1D-A*) in VisuaLink -3000 (VL-3000). (*1D-B*) was adopted in H.261[25] and also implemented in VL-5000[26].

References (Part-1)

1) Hisashi Kaneko, T. Ishiguro: “Digital Television Transmission Using Bandwidth Compression Techniques”, IEEE Communications Magazine vol.18, No.4, pp.14 – 22 (July 1980).

2) T. Ishiguro, K. Iinuma: “Television Bandwidth Compression Transmission by Motion-compensated Interframe Coding”, IEEE Communications Magazine, vol.20, No.6, pp.24 – 30 (Nov. 1982).

3) T. Ishiguro: “VLSI in Picture Coding”, Kruwer Academic Publishers, Journal of VLSI Signal Processing, Vol.5, pp.115 –120 (1993).

4) T. Ishiguro, T. Ohshima, Y. Iijima, A. Tomozawa: “Delta Modulation CODEC for Video Phone Signals”, IECE, CS69 - 32, pp.1–16 (July 1969). (in Japanese)

5) T. Oshima, T. Ishiguro: “Reduction of Edge Busyness in Delta Modulation”, IEEE Trans. on Communications, Vol. COM - 23, pp.550 – 554 (May 1975).

6) T. Ishiguro, T. Oshima, Hisashi Kaneko: “Digital DPCM Codec for TV Signals Based on ΔM/DPCM Digital Conversions”, IEEE Trans. on Communications. Vol. COM-22, No.7, pp. 970 – 976 (July 1974).
7) T. Ohshima, T. Ishiguro: “Higher Order ΔM-CODEC for NTSC Color Television Signals”, IECE, CS75-9, pp.1 – 8 (May 1975). (in Japanese)
8) Y. Iijima, T. Ishiguro: “DPCM of NTSC Color Television Signals”, IECE, CS73 - 44, pp.1 – 9 (July 1973). (in Japanese)
9) T. Ishiguro, N. Suzuki, Y. Iijima, N. Kawachi: “32 M b/s Higher Order DPCM of NTSC Color Television Signals”, IECE, CS75 - 69, pp. 9 – 15 (Mar. 1975). (in Japanese)
10) Y. Iijima, N. Suzuki: “Experiments on Higher Order DPCM for NTSC Color Television Signals”, IECE, CS74 - 63, pp.29 – 38 (Aug. 1974). (in Japanese)
11) T. Ishiguro, K. Iinuma, Y. Iijima, T. Koga, S. Azami, T. Mune: “Composite Interframe Coding of NTSC Color Television Signals”, Proc. of IEEE National Telecommunications Conference (NTC’76), Dallas (TX), pp.6.4.1 – 6.4.5 (Nov. 1976).
12) K. Iinuma, Y. Iijima, T. Ishiguro, H. Kaneko, S. Shigaki: “Interframe Coding for 4-MHz Color Television Signals”, IEEE Trans. on Communications, Vol.COM-23, No.12, pp.1461 – 1466 (Dec. 1975).
13) T. Ishiguro, K. Iinuma, Y. Iijima, T. Koga, H. Kaneko: “NETEC system : Interframe Encoder for NTSC Color Television Signals”, Proc. of 3rd International Conference on Digital Satellite Communications, Kyoto (Japan), pp. 309 – 314 (Nov. 1975).
14) R. E. Graham: “Predictive Quantizing of Television Signals”, IRE WESCON Convention Record, Part - 4, pp.147 – 156 (Aug. 1958).
15) J. C. Candy, M. A. Franke, B. G. Haskell, F. W. Mounts: “Transmitting Television as Clusters of Frame-to-Frame Differences”, Bell System Technical Journal (BSTJ), Vol. 50, pp.1889 – 1919 (July - Aug. 1971).
16) T. Koga, K. Iinuma: “Study of Adaptive Interframe Coding Efficiency for TV Signals”, IECE, CS80 - 79/IE80 - 58, pp. 63 – 70 (July 1980). (in Japanese)
17) N. Suzuki, K. Iinuma, T. Ishiguro: “Information Preserving Coding for Broadcast Television Signals”, Proc. of IEEE Global Telecommunications Conference (GLOBECOM’82), Miami (FL), pp.B6.7.1 – B6.7.5 (Nov. 29 - Dec. 2, 1982).
18) T. Shibuya, N. Suzuki, N. Kawayachi, T. Koga: “Adaptive DPCM for Broadcast Quality TV Transmission of Composite NTSC Signal”, Proc. of SPIE Visual Communications and Image Processing (VCIP'93), Eds. B. G. Haskell and H.-M. Hang, vol. 2094, pp.434 – 443 (Nov. 1993).
19) T. Koga, K. Iinuma, A. Hirano, Y. Iijima, T. Ishiguro: “Motion-compensated Interframe Coding for Videoconferencing”, Proc. of IEEE National Telecommunications Conference (NTC’81), New Orleans (LA), pp. G5.3.1 – G5.3.5 (Nov. 1981).
20) T. Koga, A. Hirano, Y. Iijima, K. Iinuma: “Motion-compensated Adaptive Intra-Interframe Prediction Coding Algorithm”, Proc. IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP’85), Tampa (FL), pp.10.7.1 – 10.7.4 (Mar. 1985).
21) N. Ahmed, T. Natarajan, K. R. Rao: “Discrete Transform”, IEEE Trans. on Computer, Vol. C-23, No.1, pp.90 – 93 (Jan. 1974).
22) N. Suzuki, Y. Iijima, T. Ishiguro: “Variable Length Code Higher Order DPCM”, IECE (Japan), CS76 - 47, pp.33 – 40, (July 1976). (in Japanese)
23) T. Koga, K. Iinuma, Y. Iijima, K. Niwa: “Motion-compensated Interframe and Intraframe Adaptive Prediction and Entropy Coding”, Journal of ITE, vol.39, No.10, pp.963(113) – 971(121) (Oct. 1985). (in Japanese)
24) T. Koga, M. Ohta: “Entropy Coding for a Hybrid Scheme with Motion Compensation in Subprimary Rate Video Transmission”, IEEE Journal on Selected Areas in Communications, Vol. SAC - 5, No. 7, pp. 1166 – 1174 (Aug. 1987).
25) NTT, KDD, NEC, FUJITSU: “MOTION VECTOR CODING”, CCITT Study Group XV, Working Party XV/1, Specialists Group on Coding for Visual Telephony, Doc. #153, pp.1 – 4 (Nov. 1986).
26) Y. Endo, M. Nishiwaki, T. Yasuda, Y. Nakano, T. Koga, T. So, M. Sato: “p × 64 Standard Video Codec with LSI Technology”, ITU 6th World Telecommunications Forum : Part 2 Technical Symposium, Geneva, pp.303–307 (Oct. 1991).

Appendix to Part-1 : Contents of Part-2

（Received Oct. 21, 2024）
（Revised Dec. 2, 2024）

Toshio KOGA *(Fellow)*
He received B. S., M. S., and Ph. D. degrees from Kyushu University, Fukuoka, Japan, in 1969, 1971, and 1989, respectively. Since joining NEC Corporation in 1971, he had been engaged in R & D on video coding and video communication systems. In addition, he participated in the standardization of H.261. As an extension of the activities, he was involved in the HATS interoperability tests for H.320/H.261 terminals as a chairman. In 2000, he joined Yamagata University in Yamagata Prefecture, Japan.
He is Fellow of IIEEJ (Japan), Fellow of IEICE (Japan), Fellow of SPIE (U.S.A.), and Life Fellow of IEEE.
He is a recipient of Commendation by the Minister of State for Science and Technology (Persons of Scientific and Technological Research Merits) in 1993. He is also a corecipient of The Emmie (1990-1991), Outstanding Paper Award in the 3rd DIGISAT Conference in 1975, Best Paper Award in IEEE ICC’76 Conference in 1976, The Ichimura Prize in Industry for Excellent Achievement in 1994, and several other awards.

IIEEJ Survey Paper

Corporate Efforts for R&D on Video Coding and Its Practical Implementation (Part–2) : Standardization Activities and Practical Contributions to Video Coding World

Toshio KOGA *(Fellow)* †

† NEC Corporation (1971 - 2000) and Yamagata University (2000 - 2012)

<Summary> Digital video signal processing for varieties of purposes including storage as well as transmission is currently playing a very important role in our daily life. NEC Corporation has been actively doing R&D on video coding since its dawn and incessantly making every effort for implementation of practical codecs, terminals, and systems. They include Delta Modulation for digitization, HO-DPCM for high quality TV transmission, and interframe coding for a broad spectrum of audiovisual services. The author believes it is quite beneficial to introduce a combined history of coding algorithm improvements and practical codecs based on them, since they will provide the overview of the technical history from theoretical and also practical aspects, and definitely there is no such review published so far. This survey paper consists of two parts and reviews the corporate R&D efforts with respect to video coding algorithm improvements and our contribution to progress in digital TV/video transmission services over the world for three decades since the mid-1960s. Part -1 (Chapter 1and 2) mainly handles continued improvement of Delta Modulation and coding algorithms on intraframe and interframe prediction as well as entropy coding. Part -2 (from Chapter 3 to 6) mainly shows effectiveness and significance of video coding through various application examples of codecs/terminals as well as standardization activities in which we were deeply involved. In addition, Part-2 includes a brief history of NEC's CODECs in conjunction with continued algorithm improvements and several examples of practical use in businesses.

Keywords: video coding, NTSC signal, TDM signal, adaptive prediction, motion compensation, entropy coding, H.120/Part 3, H.320/H.261, interoperability test

3. Pioneering Experiences to Verify and Enhance Effectiveness of Video Coding

We had quite valuable experiences since the development of NETEC-6/16, which is our first codec, in 1974. They include local live demonstration, real-time video transmission through existing radio networks, and transmission by various flexible ways including terrestrial and satellite channels. Furthermore, we contributed to realization of real-time Motion Compensation.

3.1 Algorithm implementation using premature ICs

It was in 1966 that SN7400, a 14-pin discrete IC package with 2-input Quad NAND gates, was commercially available from Texas Instruments for the first time in the world.

Eight years later, NETEC-6/16 was developed. It was designed to encode TDM Color TV signal consisting of 480 lines by 510 pixels/line in a frame, approximately amounting to 2Mbits. Surprisingly, about 2,000 DRAM packages were needed for a single frame memory, since memory capacity of a commercially available DRAM then was 1 kbit/package. In addition, about 2,000 discrete ICs such as 4-bit full adders, 4-bit registers, etc., were necessary for arithmetic and/or logical computation. Complexity of the NETEC-6/16 decoding algorithm was almost the same as its coding algorithm. As a result, more than 8,000 IC packages were used in total.

Every function needed was implemented in wired-logic hardware, resulting in a huge codec set as shown in **Fig. 10**. Both Encoder and Decoder are about 2 m high. Two frame memories, each being about 1 m high, are not equipped inside but standing next to Encoder and Decoder, respectively.

3.2 Live demonstration of interframe-coded video

Soon after the development of NETEC-6/16, an international conference on digital satellite communications (DIGISAT) was held in 1975 in Kyoto, Japan[13]. The codec set was transported to the conference site, since it was a good opportunity to show how much coded video transmission is promising. Video signals taken by TV camera operated on the site were encoded, transmitted back-to-back, and decoded on the spot. Satellite engineers were strongly impressed by possibility of the coded

Fig.10 NETEC-6/16

Fig.11 Outstanding Paper Award

video transmission, i.e., remarkable effectiveness of bandwidth reduction in satellite channels. It was the first live exhibition of interframe-coded video in the world. The impact by the paper presentation as well as the exhibition on the conference site may have contributed to our reception of "Outstanding Paper Award" shown in **Fig. 11**.

3.3 Coded video transmission

Our codecs were provided with flexible transmission capability such as straightforward, inverse-multiplexed, and multiplexed, taking into considerations customers' network facilities available.

3.3.1 Transmission through existing radio networks

A long distance transmission experiment was conducted in 1976[27] jointly by National Police Agency Japan (NPA) and NEC. Coded video was sent from Osaka to Tokyo, 550 km distant, using NETEC-6/16. Encoder was placed at a communication facility in NPA Osaka District building, while Decoder in Tokyo, since there was only a single NETEC-6/16 codec set available. Coded video was transmitted through NPA's digital radio networks capable of transmission at 7.876 or 2 × 7.876 Mbit/sec. In spite of a prototype codec, NETEC-6/16 was ready for the transmission, since fundamental functions were already equipped for communication such as digital network interface and Error Correction capability. That is, Double Error Correcting BCH (239/255) code was implemented.

3.3.2 Transmission through multiple low speed lines

Even in the 1970s, digital networks were not flexible for public users in that there were few choices with respect to transmission speeds. PCM-24 or T1 line was widely used in U.S. for transmission at 1.5 Mbit/sec. However, 3 or 6 Mbit/sec may be required for better video quality. It is convenient if these rates are realized in effect by combining several T1 channels. What is called Inverse MUX/DMUX is introduced for this purpose. In Inverse MUX at Encoder, coded information is divided into several streams, say, two or four, and each stream is assigned a single low speed channel (T1). In Inverse DMUX at Decoder, each stream from the two or four channels is once buffered and delay time difference is adjusted among the channels by frame aligners. And then, they are combined into a single stream to reproduce the initial stream information. The Inverse MUX/DMUX function was implemented in NETEC-6[28] and NETEC-6/3[29].

3.3.3 Simultaneous transmission of plural TV programs through a satellite[30]

(1) Adaptive bit-sharing (ABS) multiplexed transmission

An example is shown in **Fig. 12** for multiplexing three channels. ABS-MUX accepts buffer memory occupancy (BMO) values, BMO1 through BOM3, from the three encoders. Each BMO shows information amount being generated in each encoder and corresponds to request for a necessary transmission rate. DATA1 through DATA3 from each encoder are already partitioned into blocks and ABS-MUX assigns the number of blocks to be sent out from each encoder. CLOCK1 through CLOCK3 correspond to permission for sending the blocks. The number of blocks coming from each encoder is changing at any instance while the total number of blocks is always kept constant in the transmission line. The block size is chosen to be 64 bytes including Header information. This is quite close to Packet size defined in Asynchronous Transfer Mode of B-ISDN. This adaptive bit-sharing system is a good example of packet video transmission. **Figure 13** shows an ABS-MUX/DMUX system at Los Angels Station (LA), operated by Western Union, a U.S. satellite communication company. Shown are, (from left to right), three Encoders, Satellite MODEM, ABS-MUX/DMUX, and two Decoders, i.e., three outgoing video channels are multiplexed, and two incoming channels are demultiplexed. The satellite transponder allows transmission at 60 Mbit/sec.

(2) Effectiveness verification for ABS multiplexing[31]

Assignment decision is made instantaneously by ABS-MUX/DMUX corresponding to BMO values. The assignment was measured in real-time using commercial TV programs for two kinds of multiplexing, two- and three-channel cases, both with a total transmissin rate of 60 Mbit/sec.

The assignment results in **Fig. 14** show that effectiveness in the three-channel case is apparent where the probability of visible degradation such as 30 ~ 40 dB reduces to almost 1/10. The

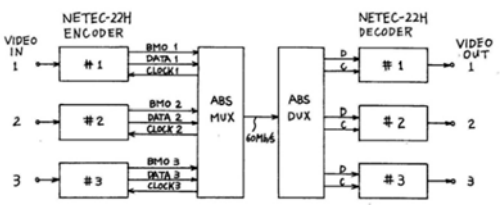


Fig.12 ABS-MUX/DMUX System Configuration

Fig.13 ABS System

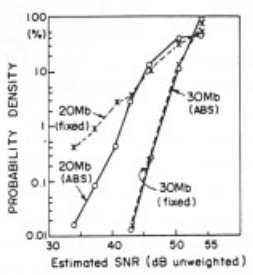


Fig.14 Effectiveness

effectiveness is less visible with the two-channel case. It seems transmission at 30 Mbits/sec may not need multiplexing.

3.4 Real-time motion vector detection

Motion vector detection by BMA in real time requires prohibitively high volume of computation. Let us roughly calculate the amount for Full Search algorithm in which all candidate vectors are evaluated and compared. MC Range is assumed to cover ± S_H pixels and ± S_V lines, respectively.

When a similarity measure SUM($\vec{V}_k$) defined in Eq.(5) is calculated, execution of {subtraction, C[e], summation} is necessary for every pixel in a block. If C[e] corresponds to an Instruction, three Instructions are carried out in the execution.

The calculation for a SUM value needs 3×(M×N) Instructions. In case of Full search, it is repeated $(2S_H+1)\times(2S_V+1)$ times for all the candidate vectors in MC Range. Lastly, the number of comparison necessary to find minimum SUM($\vec{V}_k$) value is added, which is equal to $(2S_H+1)\times(2S_V+1)$. Total amount of calculation to be completed within a pixel time is given by the following calculation,

$$3\times\{(2S_H+1)\times(2S_V+1)\}+\{(2S_H+1)\times(2S_V+1)\}\ /\ (M\times N).$$

If we assume Block size (M×N) to be 8 lines × 8 pixels, and in addition, $S_H = S_V = 8$, then, the number of calculations in a pixel time amounts to 871.5 Instructions. If the sampling frequency is 8 MHz, it is equal to 6,972 Mega Instructions/sec, which presumably corresponds to computational power of a supercomputer in the 1970s. If the MC range is expanded, higher speed is required approximately in proportion to the expansion ratio. Change in the block size does not matter.

3.4.1 Fast algorithm for motion vector detection[19)]

A fast algorithm is mandatory to realize MC in a practical size using commercially available components. In general, difference between SUM($\vec{V}_{opt}$) and SUM($\vec{V}_k$) values decreases monotonously as the norm difference between the two vectors becomes smaller. This property suggests an algorithm which looks for the smallest SUM value in several steps in a coarse-to-fine manner.

As shown in **Fig. 15**, nine candidate vectors are coarsely spaced with one of the vectors located at (0, 0) in the first step and compared one by one to find which vector gives the smallest SUM value. A vector located at (0, 0) is $\vec{V}_0$, equivalent to interframe prediction. In this example, let us assume that a vector $\vec{V}_1$ in the north-east direction shows the smallest SUM in the first step. Then, another nine vectors are less coarsely spaced, where $\vec{V}_1$ (= $\vec{V}_{10}$) is included at the center and other eight vectors surround it. Eight corresponding SUM values are compared with SUM($\vec{V}_{10}$) to find smaller SUM, if any. This is the second step. If a vector $\vec{V}_{12}$ placed above $\vec{V}_{10}$ is assumed to give the smallest SUM, then another nine vectors in its closest vicinity are placed in a similar manner to the previous step, and further compared to find the smallest SUM. A vector ($\vec{V}_{123}$) corresponding to the smallest SUM is chosen to be the motion vector ($\vec{V}_{opt}$). This is the last step in the case of Three-Step Search. The number of calculations is reduced approximately to 25/169 in this example. Further multi-step search is also possible for higher accuracy.

3.4.2 Hardware implementation

Summation of Cost Function C[e] outputs is necessary to calculate SUM($\vec{V}_k$) in Eq. (5) and it may be normally carried out by using M×N Cost Function cards working in parallel. However, it is desirable that the hardware is as small as possible. The number of pixel points for C[e] calculation is reduced by 2-to-1 subsampling the points vertically as well as horizontally, resulting in 4-to-1 reduction. Candidate Vector Generator is equipped to indicate and control the vector selection in each step after comparing SUM values for each $\vec{V}_k$.

The implemented Vector Detector (H:25, W:55, D:20 in cm) is shown in **Fig. 16**, in which sixteen C [e] cards work in parallel for the MC block size of 8 by 8. One of them is shown in **Fig. 17**, which includes eight 1-kbit static RAMs (placed in two columns and each seen as a white-gray-white pattern). The static RAMs can provide an MC search range covering 17 pixels × 17 lines in the case of $S_H = S_V = 8$.

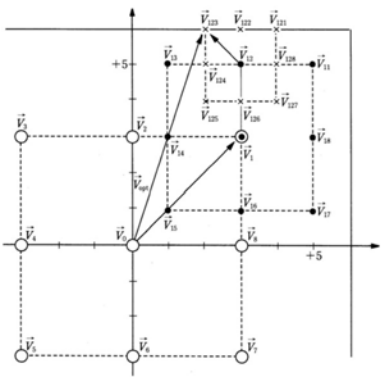


Fig.15 Three-Step Search

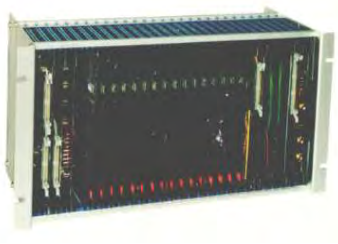


Fig.16 Motion Vector Detector

Fig.17 C [e] Card

4. Contribution to Standardization Activities

4.1 Broadcast TV program transmission project in U.S.

A project started in U.S. in 1987 by the Committee T1. Sub-committeeT1Y1.1, later T1A1.1, was responsible for the project "Digital Encoding of System M-NTSC Television Signals for Broadcast Quality Transmission at the DS-3 Rate."
Many companies such as ABL, Grass Valley Group, NEC, Northern Telecom Inc., and Telettra took part in this project. NEC provided HO-DPCM 45 and its successors for video coding algorithm comparison as well as for video quality evaluation during 1987 through 1991.

However, this activity faded out, when it was widely recognized that MPEG-2 would come up soon.

4.2 H.120/Part-3 for audiovisual services at 1.5 Mbit/sec

In 1983, a Working Party in CCITT SG-XV started to create a common video coding algorithm H.120/Part-3 for transmission of video conference signals at a primary rate (1.5 Mbit/sec) in NTSC countries. PAL/SECAM countries selected separately another algorithm H.120/Part-1.

H.120/Part-3 is basically interframe prediction adaptively combined with intraframe one. However, the interframe part consists of two types, one for MC and the other for Background prediction. The latter was proposed by H. Kuroda et al., (NTT, Japan)[32], since MC does not work in uncovered background parts. That is, H.120/Part-3 is an adaptive algorithm among three prediction functions, selected on a pixel-by-pixel basis. As for its entropy coding, NETEC-X1MC is fully referenced. Therefore, H.120/Part-3 can be regarded as equivalent to the specification of NETEC-X1MC with back-ground prediction capability.

Recommendation H.120/Part-3 was officially issued by CCITT in 1988. However, as a matter of fact, it didn't come to common use due to significant delay against its initial schedule and, in addition, expectation for H.261 to come up soon.

4.3 H.320/H.261 for audiovisual services in ISDN

H.261 is a block-based video coding algorithm consisting of MC interframe prediction, DCT, quantization, and entropy coding[33]. A prototype H.261 codec (H:210, W:56, D:61 in cm) was developed in 1988[34] to verify video coding performance. It was called "Flexible Hardware" or "n×384 codec" shown in **Fig. 18**. A prototype H.320/H.261 terminal (H:75, W:56, D:61 in cm) was developed in 1989 for overall function verification. It was called "p×64" and is shown in **Fig. 19.**

"Flexible Hardware" was implemented with many printed circuit boards, each consisting of discrete ICs and other circuit components. DCT/IDCT circuit boards were also made in the same manner. Therefore, we could directly touch almost any

Fig.18 Flexible Hardware

Fig.19 p×64 Terminal

pins of ICs and circuit components. When we intentionally let a part of DCT/IDCT circuit in Encoder or IDCT in Decoder short-circuited to the ground in a moment, mismatch in DCT/IDCT calculation takes place between Encoder and Decoder. This unusual experiment reminded us of a possible DCT/IDCT mismatch problem, if H.320/H.261 terminals are produced by different manufacturers without any common specification[35]. It contributed to find a practical solution, i.e., DCT/IDCT accuracy specification within a certain tolerance in combination with cyclic refresh by Intra DCT mode.

4.4 Interoperability tests for H.320/H.261 terminals

When terminals based on H.320/H.261 are delivered to customers, every one of them should be able to talk to each other. Therefore, interoperability tests were planned inside Japan at first and then internationally. Promotion Conference of Harmonization of Advanced Telecommunications Systems (HATS Conference Japan) supported by Ministry of Posts and Telecommunications (then), was responsible for proliferation of international telecommunication standards throughout Japan. Upon request by HATS, TTC (The Telecommunication Technology Committee, supported and organized by industries) completed a guideline for the test in 1994[36]. The author was asked to chair HATS Digital Video Conference and Videophone WG responsible for the tests, including the guideline preparation.

First round tests started in Japan in 1991 and the tests were carried out eight times in total until 1996. During these tests, 17 vendors or organizations participated. After considerable experience in executing the tests, we started revision of the guideline and, in addition, preparation for international tests to come. Almost at the same time when it was completed, we planned a test with European countries. As a first international interoperability test, we had an opportunity to work with Belgium through a framework of EJIX (Europe-Japan ISDN Experiment Program). The first test was conducted in March 1994 and the second in July 1994. The ratio of successful calls

in total was 70 % (56/80 calls). Later in 1995, another test was also conducted between European countries and Japan in a similar manner through a framework of EVE (European Videotelephony Experiment). It consists of six Telecom organizations from U.K., France, Sweden, Germany, Italy, and The Netherlands. PTT Telecom (The Netherlands) played a central role in European countries in that all the terminals in European side were prepared in its facility and the test was carried out successfully.

Encouraged by these experiences, HATS thought it was appropriate to propose the test between U.S. and Japan. Representing HATS WG, the author proposed the test in July 1994 to T1A1.1 meeting, a subcommittee of the Committee T1[37]. The test was carried out on a voluntary basis with the participation of three U.S. and nine Japanese vendors in March 1995. The number of calls placed in total was 54 in which 49 calls were successful, i.e., success ratio was 90.7%. The author also reported the result in the T1A1.1 meeting in Aug. 1995 [38].

All the test results above were reported in an international conference and published in the conference proceedings[39].

5. Overview of NEC's Video CODECs

Throughout the three decades from the mid-1960s, NEC developed many codecs and terminals largely in three application categories; broadcast quality TV transmission, high quality services in CCTV/CATV, and wide applications such as audio-visual services. Those codecs/terminals were based on our proprietary algorithms, since our R&D started more than ten years before the dawn of international standardization activities. However, we also developed those conforming to recommendations immediately when officially issued.

5.1 Broadcast quality TV transmission use

NETEC-22H was only one composite interframe codec developed for NTSC Color TV transmission with high quality. Several NETEC-22H codecs were used in satellite networks, where two or three Broadcast TV signals were adaptively multiplexed and transmitted through a satellite[30]. This system is exemplified in Figs. 12 and 13.

HO-DPCM 45-series codecs were based on composite intraframe coding and paved way to digital terrestrial broadcasting of NTSC Color TV. These codecs were delivered to one of three major TV broadcasting networks in US. and used in practice in a news program connecting New York and Washington DC. In addition, several HO-DPCM-series codecs and their successor Broadcaster 45 were provided to official evaluation tests for establishing digital TV program transmission specification in U.S. conducted by T1Y1.1/T1A1.1. Later, Broadcaster-52 was developed in 1994 for operation at 52 Mbit/sec with STM-0 line interface. It was equipped with lossless coding capability. Codecs based on HO-DPCM are summarized in Table 1.

Other codecs conforming to standards for broadcasting were also developed. Broadcaster-34 is a compo-nent codec developed in 1983, conforming to ITU-R 723 /ETSI (ETS 300174), whose adaptive prediction consists of MC interframe, interfield, and intraframe prediction. It is used for PAL/NTSC transmission at 34 Mbit/sec. Broadcaster-140, developed in 1989, is a codec without compression for NTSC/PAL/SECAM TV signals transmission. It conforms to ITU-R721.

5.2 High quality CATV/CCTV use

This application requires high quality but not so high as broadcast TV, while lower transmission rates and inexpensive codecs are favored. As a solution for these requirements, a lower sampling frequency is chosen and followed by a simple algorithm. HO-DPCM 32 is an experimental codec developed in 1975 for 32 Mbit/sec transmission, based on sub-Nyquist sampling and higher-order prediction appropriate for $fs = 2 \times fsc$, i.e., 7.2 MHz[9]. Frequency response $H(z)$ in this case is given by Eq. (2). Prediction error is quantized with 31-level characteristics and code-converted with 5-bit FLC codewords[9]. By removing partially Horizontal Blanking intervals, video signals along with audio are transmitted at 32 Mbit/sec. HO-DPCM 45B, developed in 1985 for industrial use at 45 Mbit/sec, and its successors are based on higher order prediction for video signals sampled at 8.8 MHz, a somewhat generalized frequency different from multiple integer times fsc. Prediction error is handled with FLC in a similar way to HO-DPCM 32A.

5.3 Audiovisual communication use

Based on our NETEC-6/16 development experience, many NETEC-series codecs were dev. eloped for applications such as videotelephone, audiovisual services, distance learning, etc. (see **Table 2**). NETEC-6 was the first commercial model and designed for use at 6 Mb/sec by incorporating four T1 lines in parallel[28].This is Inverse Multiplexing/Demultiplexing in **3.3.2.** The prediction algorithm is almost the same as that of NETEC-6/16. NETEC-6/3 was developed for transmission at 6 or 3M bit/sec, using 16-kbit DRAMs for frame memories[29]. It is in effect the first commercial product in NETEC-series codecs. Its performance was evaluated by Bell Systems in U.S. for their video communication services. "Two-diff" type adaptive prediction was implemented in NETEC-X1 for the first time and was named so after its targeted transmission rate, i.e., a single T1 line.

When MC was incorporated into NETEC-X1, it was renamed NETEC-X1MC[20) 23)]. Its algorithm is "Two-diff" type adaptive prediction with MC. In addition, the motion vector detection method is improved here. In view of what is entropy-coded and transmitted in essence, it must be better to take into consideration an estimated amount of information on both motion vector and corresponding MC prediction error[40)]. That is, what should be minimized is not a similarity measure SUM but SUM + C2[$\vec{V}_k$], where C2 [$\vec{V}_k$] is a cost function for $\vec{V}_k$. Minimum amount of information to be transmitted is preferred to similarity.

This change contributes to information amount reduction by 7 ~ 15 % on the average for normal videoconference signals and for signals with large moving objects as well as TV camera panning. NETEC-X1MC was introduced in a private satellite education network of a major IT company in U.S. in 1983. Its successor was developed in 1985 for transmission at Sub-T1 (384 kbit/sec) through T1 (1.5 Mbit/sec)[41)]. A simplified logical decision is introduced in the adaptation algorithm, instead of Look-up Table used in NETEC-X1MC. That is, intraframe prediction is selected only when intraframe prediction was judged to be better both at a previous pixel and a previous line position. Otherwise, interframe prediction is used.

In addition, improvement was made in subsampling and freeze picture modes for lower rate operation. Quincunx subsampling is employed to reduce artifacts. Skipped pixels are reproduced by interpolation using four neighboring pixels, i.e., upper, lower, left, and right. Freeze picture may take place more often as transmission rates become lower. It seems better to resume the normal coding process gradually upon restart to suppress frequent repetition of freeze mode[42)]. For example, a block-line consisting of eight successive lines is coded in a frame immediately after restart. The number of coded block-lines increases frame by frame. Thus, the decoded video part becomes wider and wider, just like a window shutter being opened, though downwards.

NETEC-series CODECs production came to an end when H.320/261 was officially recommended in 1989. During these days, VLSI technology made a great progress[3)]. NEC developed Video/Image Signal Processor (VISP)[43)], a 16-bit fixed-point

Table 2 Coding Algorithms in NEC Interframe CODECs

Prediction	Fixed Prediction		Adaptive Prediction				
	Pixel-based (e)		Pixel-based (e)		Block-based Motion Compensation		
					Pixel-based (e)	Block-based (e)	
	Fixed FF (1)	Fixed FF (2)	Adapt-FF (1)	Adapt-FF (2)	MC+Adapt-FF (3)	MC + DCT	
	Isolated Pel Removal	Combinational Difference + Non-linear Circuit	Threshold (Type II)	Two-diff (z^{-1}, z^{-F})	MC Two-diff (z^{-1}, z^{-MC})	2D-DCT	
Entropy Coding							
(1) Prediction Error Information							
	Prediction Error Amplitude					DCT coeff	
= 0	Block-based				Pixel-based	Block-based	
≠ 0	(FLC1, FLC2)	VLC	Adaptive (FLC, VLC)			zig-zag+VLC	
(2) Motion Vector Information					diff Vector	diff Vector	
				= 0	RLC	RLC	
				≠ 0	VLC (*2D*)	VLC (*1D*)	
CODEC	N-6/16 ('75)	N-22H ('76)	N-6/3 ('79)	N-22H ('78)	N-X1 ('81)	N-X1MC ('83)	VL-1000 ('89)
	N-6 ('77)					N-XV ('85)	VL-3000 ('89)
	(N:NETEC, VL:VisuaLink)						VL-5000 ('93)

processor with Instruction Cycle time of 25 nsec. A single board processor VSP (Video Signal Processor) consists of six VISPs as shown in **Fig. 20**. Since we started to develop a new codec conforming to H.320/H.261 in parallel with the standardization process, it was impossible to implement it by wired-logic hardware as was usually done before. That is, software-based codec became a must in order to make it flexible to possible changes in the process of standardization. Then, a new software-based codec was developed in 1989 using several VSP boards. It is VL-3000 (H:75, W:52, D:70 in cm), shown in **Fig. 21**. VL-3000 was the first VisuaLink-series VSP-based codec. As an example of its system application, a satellite-based multipoint videoconference network was introduced by a major pharmaceutical company in Japan in 1991[44]. An ASIC-based successor of VL-3000, named VL-50000[26], was developed in 1991 in a small box (H:22, W:42.5, D:45 in cm) with weight of 25 kg. Reduction ratio is 1/6.5 in volume and 1/4 in weight compared with VL-3000. VL-5000 EX was made further smaller in 1994, resulting in 1/3 in volume and 1/2 in weight compared with VL-5000. This newest model was also used in the interoperability tests inside Japan and in international tests as well. The next and the last VL-series CODEC was VL-7000. However, it was based on MPEG-2 SP@ML algorithm.

NETEC- and VisuaLink-series interframe CODECs above are briefly summarized in Table 2 with their individual video coding algorithms. **"**FF" is used here for "interframe".

Achievement Award was given to Hisashi Kaneko, Tatsuo Ishiguro, and Kazumoto Iinuma in 1986 for "R&D and Practical Implementation of Interframe Codecs for Television Signals," by Institute of Electronics, Information, and Communications Engineers (IEICE, Japan).

6. Conclusion

This survey paper is to describe how intensively a private company has made efforts for R&D on video coding and their implementation in practice toward utilization worldwide. The author would like to emphasize that all the description here is based on nothing but facts without exaggeration, i.e., just what and how we did or experienced. He also expects the readers, young readers in particular, to understand or imagine how different or poor the R&D environment in several tens of years ago was.

In view of R&D environment during three decades from the mid-1960s, computer power at the time was not so adequate that simulation works should be highly effective or less time-consuming. H.261 is very impressive in that the algorithm consists of a small number of "essential" or "very influential" parameters. It is a result from a common understanding, "Divergence and Convergence." This attitude may have changed due to progress in semiconductor technology in the mid-90s. It seems, like in MPEG specifications, many parameters are included in the algorithms so long as they are advantageous, even if they require a considerable amount of computation. This is a significant change before MPEG and thereafter. Furthermore, a huge change can be seen in implementation, in that algorithms can be realized using only a single or several microprocessors, i.e., transition from wired logic to software.

The Emmy (1990 - 1991) shown in **Fig.22** was presented to NEC in 1991 by The National Academy of Television Arts and Sciences for "Pioneering Work and Implementation of Data Compression Techniques for Real-time Television Transmission." This is one of the best proofs of our incessant efforts.

Acknowledgments

The author would like to appreciate deeply Hisashi Kaneko, Yasuo Katoh, Tatsuo Ishiguro, Haruo Kaneko, Kazumoto Iinuma, Hiroshi Iijima, Toshio Ohshima, Kunihiko Niwa, Yukihiko Iijima, Takayoshi Mune, Shuzo Tsugane, Toshiyuki Onaka, and many other NEC researchers/engineers for their continued leadership, guidance, encouragements, discussions, and collaboration.

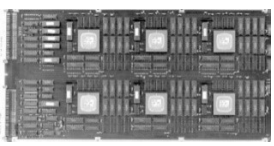


Fig.20 VSP Board

Fig.21 VL-3000

Fig.22 The Emmy (1990 - 1991)

References (Part-2)

27) S. Obara, M. Tan ge, H. Kaneko, K. Iinuma: “Experiments on Encoded Color TV Transmission through Long-distance Digital Radio Links”, IECE, CS77-50, pp.95 –102 (July 1977). (in Japanese)

28) K. Iinuma, Y. Iijima, T. Ishiguro, T. Koga, S. Tanaka, H. Kaneko: “NETEC-6 : Interframe Encoder for Color Television Signals”, NEC Research & Development, No.44, pp.92–96 (Jan. 1977).

29) Hisashi Kaneko, Y. Iijima, T. Ishiguro, K. Iinuma: “NETEC-6/3 Video Transmission Equipment for Teleconference”, Proc. of INTELCOM’79, Dallas (TX), pp. 1– 4 (March 1979).

30) Hisashi Kaneko, T. Ishiguro, M. Sugiyama: “Digital Television Transmission through The Satellite”, Proc. of INTELCOM’77, Atlanta (GA), pp. 1 –7 (Oct. 1977).

31) T. Koga, Y. Iijima, K. Iinuma, T. Ishiguro: “Statistical Performance Analysis of an Interframe Encoder for Broadcast Television Signals”, IEEE Trans. on Communications, Vol. COM - 29, No. 12, pp.1868 ー 1876 (Dec. 1981).

32) H. Kuroda, H. Hashimoto, S. Ohkubo, N. Mukawa: “An Interframe Coding System for Video Teleconferencing (TRIDEC 1.5) ”, IEICE, IE84 - 4, pp.25 – 31 (Apl. 1984). (in Japanese)

33) R. Plompen, Y. Hatori, W. Geuen, J. Guichard, M. Guglielmo, H. Brusewitz: “Motion video coding in CCITT SG-XV - The Video Source Coding”, Proc. IEEE Global Telecommunications Conference (Globe-com’88), Miami (FL), pp.31.2.1 – 31.2.8, Nov. 1988.

34) J. Guichard, D. Devimuex, T. Koga, D. G. Morrison, N. Randall, J. Speidel: “Motion Video Coding in CCITT SG-XV - Hardware Trials”, Proc. of IEEE Global Telecommunications Conference (Globecom’88), Miami (FL), pp.2.2.1 – 2.2.6 (Nov. 1988).

35) M. Ohta, T. Omachi, T. Koga: “Estimation and Countermeasure of IDCT Mismatch Error for Interframe DCT Coding”, IEICE, IE88 - 44, pp.55 – 62 (July 1988). (in Japanese)

36) TTC-G-001-V1E : TTC Guideline for Execution of Interconnection Tests, “Guideline for Executing Interconnection Tests of Digital Visual Telephone Terminals”, (Apr. 1994).

37) T1A1.5/94 - 135: “Proposal on Interconnection Test for H.320 Terminals between U.S. and Japan”, (July 1994). (Contribution document)

38) TlAl.5/95 - 216: “Results of First Interconnection Tests for H.320 Terminals between U.S. and Japan”, (Aug. 1995). (Contribution document)

39) T. Koga: “Interconnectivity Testing in Japan for Equipments based on ITU-T Recommendation for Audiovisual Services”, Ed. K. R. Rao, SPIE Optical Engineering Press, Vol.CR60, “Standardization and Common Interfaces for Video Information Systems”, pp.178 – 205 (Oct. 1995).

40) T. Koga, A. Hirano, K. Iinuma, Y. Iijima, T. Ishiguro: “A 1.5 Mb/s Interframe Codec with Motion Compensation”, Proc. of IEEE International Conference on Communications (ICC’83), Boston (MA), pp. D8.7.1 –– D8.7.5 (June 1983).

41) T. Koga, K. Iinuma, K. Niwa, S. Tsugane, M. Nishiwaki, Y. Iijima: “ Sub-T1 Rate Motion Video Codec for Teleconference”, Proc. of IEEE International Conference on Communications (ICC’85), Chicago (IL), pp.2.4.1 – 2.4.5 (June 1985).

42) Koga, K. Niwa, Y. Iijima, K. Iinuma: “Low Bit Rate Motion Video Coder/Decoder for Teleconferencing”, SPIE Optical Engineering, Vol.26, No.7, pp.590 – 595 (July 1987).

43) T. Ishida, Y. Nakano, Y. Iijima, M. Yano, T. Mochizuki, J. Ohki, T. Nishitani: “Development of a 64 kbps Video CODEC: NETEC VisuaLink 1000”, NEC Research and Development, No.95, pp.69 – 78 (Oct. 1989).

44) T. Sato, Y. Ikeda, Y. Isoe, T. Sato, M. Noda, R. Hatanaka: “Multi-point Satellite Video Conference System for Taisho Pharmaceutical Co. Ltd.”, NEC Technical Journal, Vol. 44, No.6, pp.44 – 51 (Aug. 1991). (in Japanese)

Appendix to Part-2: Contents of Part1

（Received Oct. 21, 2024）
（Revised Dec. 2, 2024）

Toshio KOGA *(Fellow)*

He received B. S., M. S., and Ph. D. degrees from Kyushu University, Fukuoka, Japan, in 1969, 1971, and 1989, respectively. Since joining NEC Corporation in 1971, he had been engaged in R & D on video coding and video communication systems. In addition, he participated in the standardization of H.261. As an extension of the activities, he was involved in the HATS interoperability tests for H.320/H.261 terminals as a chairman. In 2000, he joined Yamagata University in Yamagata Prefecture, Japan.

He is Fellow of IIEEJ (Japan), Fellow of IEICE (Japan), Fellow of SPIE (U.S.A.), and Life Fellow of IEEE.

He is a recipient of Commendation by the Minister of State for Science and Technology (Persons of Scientific and Technological Research Merits) in 1993. He is also a corecipient of The Emmie (1990-1991), Outstanding Paper Award in the 3rd DIGISAT Conference in 1975, Best Paper Award in IEEE ICC’76 Conference in 1976, The Ichimura Prize in Industry for Excellent Achievement in 1994, and several other awards.

Report of MoU Ceremony between IEEE CTSoc and IIEEJ

IEEE CTSoc and IIEEJ have exchanged the MoU on sister society relationship on Oct.29, 2024. The idea of establishing such relationship has started at GCCE 2022 held on October 2022. GCCE is one of the biggest conferences of CTSoc, which has been held in Japan, and several BoG members of IIEEJ have been working as Committee members of GCCE for long time. On October 2023, at GCCE2023, IEEE CTSoc President and IIEEJ President have reached to the basic agreement. Then, IIEEJ has approved the content of MoU in their BoG meeting held on May 2024, and CTSoc has approved the same content of MoU in the BoG meeting on September 2024.

The content of MoU includes mutual discount of membership fees, exchange of plans and schedules for international conferences and major events, and forming partnerships for the co-sponsorship of international technical meetings, if agreed. It is also encouraged to execute joint activities including active links on each society's web site to the web site of the other, and reciprocal advertisements and address swapping for promotional purposes.

Based on these approvals, the signing ceremony was held at Kokura, Japan, on Oct.29, 2024, from 12:40 to 13:00, where GCCE2024 was held in parallel. To the ceremony, Prof. Wen-Chung Kao, President of CTSoc, and Prof. Naoki Kobayashi, past President of IIEEJ, attended representing each organization.

At the ceremony Prof. Naoki Kobayashi introduced the message from Prof. Seishi Takamura, current President of IIEEJ, as follows. "It is an honor to mark the beginning of a new chapter of cooperation. This partnership between our two esteemed societies will be a significant step in promoting technological innovation and its societal benefits. We look forward to leveraging each other's strengths and embarking on this journey together".

Prof. Wen-Chung Kao, President of CTSoc, announced his statement as follows: "It is our pleasure to have sister society relation between IIEEJ and IEEE CTSoc. We hope this partnership will further activate both societies and bring a lot of benefits to their members."

The ceremony was witnessed by Ms. Charlotte Kobert, Adminisyrator of CTSoc, Prof. Yu-Cheng Fan, National Taipei University of Technology, and Prof. Fumitaka Ono, BoG member of CTSoc and former President of IIEEJ. The term of this agreement will be until December 2026, and the renewal is contingent upon approval by both societies.

The photos of the Signing Ceremony (**Photo 1**) and the attendants to the Ceremony (**Photo 2**) are shown below.

Concerning this agreement and its ceremony, both societies will express sincere thanks to Prof. Tomohiro Hase and Prof. Takako Nonaka for their continuing encouragements and assistances.

Photo1 The Signing Ceremony

Photo 2 Attendants to the Ceremony

Call for Papers

Special Issue on

Image Electronics Technologies Related to AI

IIEEJ Editorial Committee

The rapid advancements in artificial intelligence (AI) technologies in recent years have profoundly accelerated and enhanced various image-electronics-related fields, including image and video processing, recognition, and generation. These technologies have vast potential applications, spanning autonomous driving, medical image diagnostics, facial recognition systems, anomaly detection in industrial settings, surveillance cameras, drones, and beyond. At the same time, addressing societal challenges arising from the misuse of these technologies—such as the generation of fake images—is expected to become a pressing issue. Nonetheless, it is evident that AI technologies will continue to grow in importance, playing an increasingly pivotal role in the field of image electronics technologies.

This special issue invites a broad range of submissions focusing on research advancements in AI and their impact on image-electronics-related technologies, as well as evaluations of their practical applications. Accepted contributions may include research papers, system development papers, practice-oriented papers, and survey papers.

1. Topics covered include but are not limited to
- Application of AI in image processing (recognition, classification, generation)
- Image recognition technologies using machine learning and deep learning
- Improvements in video compression and transmission technologies using AI
- Computer vision technologies using AI
- Image generation and editing using generative AI
- Fusion technologies between natural language processing and image processing
- Application of AI in medical image processing
- AI in automated driving
- Application of AI in video analysis using surveillance cameras and drones

2. Treatment of papers

   The submission paper style format and double-blind peer review process are the same as the regular paper. If the number of accepted papers is less than the minimum number for the special issue, the acceptance paper will be published as the regular contributed paper. We ask for your understanding and cooperation.

3. Publication of Special Issue:

   IIEEJ Transactions on Image Electronics and Visual Computing Vol.14, No.1 (June 2026)

4. Submission Deadline:

   Friday, October 31, 2025

5. Contact details for Inquiries:

   IIEEJ Office E-mail: hensyu@iieej.org

6. Online Submission URL: http://www.editorialmanager.com/iieej/

Revised: January 6, 2017
Revised: July 6, 2018
Revised: Dec. 10, 2024

Guidance for Paper Submission

1. Submission of Papers

(1) Preparation before submission

- The authors should download "Guidance for Paper Submission" and "Style Format" from the "Academic Journals", "English Journals" section of the Society website and prepare the paper for submission.
- Two versions of "Style Format" are available, TeX and MS Word. To reduce publishing costs and effort, use of TeX version is recommended.
- There are four categories of manuscripts as follows:
  - Ordinary paper: It should be a scholarly thesis on a unique study, development or investigation concerning image electronics engineering. This is an ordinary paper to propose new ideas and will be evaluated for novelty, utility, reliability and comprehensibility. As a general rule, the authors are requested to summarize a paper within eight pages.
  - Short paper: It is not yet a completed full paper, but instead a quick report of the partial result obtained at the preliminary stage as well as the knowledge obtained from the said result. As a general rule, the authors are requested to summarize a paper within four pages.
  - System development paper: It is a paper that is a combination of existing technology or it has its own novelty in addition to the novelty and utility of an ordinary paper, and the development results are superior to conventional methods or can be applied to other systems and demonstrates new knowledge. As a general rule, the authors are requested to summarize a paper within eight pages.
  - Data Paper: A summary of data obtained in the process of a survey, product development, test, application, and so on, which are the beneficial information for readers even though its novelty is not high. As a general rule, the authors are requested to summarize a paper within eight pages.
  - Survey Paper: A summary of existing Research and Developments, organized under some viewpoint, compared for the sake of positioning purpose, observed as the changes in generations. Comprehensive references, overall perspective, objective evaluation, are needed without advertising specific organizations. It is also appreciated that the status and problems of the field, and the effect of them to the researchers and concerned people are undersood by the author, and the resulant paper encourages the new entry into the field, accelerates further development of related technologies, and prompts the development in even other fields or brand new researches. As a general rule, the authors are requested to summarize a paper within eight pages.

- To submit the manuscript for ordinary paper, short paper, system development paper, or data paper, at least one of the authors must be a member or a student member of the society.
- We prohibit the duplicate submission of a paper. If a full paper, short paper, system development paper, or data paper with the same content has been published or submitted to other open publishing forums by the same author, or at least one of the co-authors, it shall not be accepted as a rule. Open publishing forum implies internal or external books, magazines, bulletins and newsletters from government offices, schools, company organizations, etc. This regulation does not apply to a preliminary draft to be used at an

annual meeting, seminar, symposium, conference, and lecture meeting of our society or other societies (including overseas societies). A paper that was once approved as a short paper and being submitted again as the full paper after completion is not regarded as a duplicate submission.

(2) Submission stage of a paper

- Delete all author information at the time of submission. However, deletion of reference information is the author's discretion.
- At first, please register your name on the paper submission page of the following URL, and then log in again and fill in the necessary information. Use the "Style Format" to upload your manuscript. An applicant should use PDF format (converted from dvi of TeX or MS Word format) for the manuscript. As a rule, charts (figures and tables) shall be inserted into the manuscript to use the "Style Format". (a different type of data file, such as audio and video, can be uploaded at the same time for reference.)

  http://www.editorialmanager.com/iieej/
- If you have any questions regarding the submission, please consult the editor at our office.

Contact:
Person in charge of editing
The Institute of Image Electronics Engineers of Japan
3-35-4-101, Arakawa, Arakawa-Ku, Tokyo 116-0002, Japan
E-mail: hensyu@iieej.org
Tel: +81-3-5615-2893, Fax: +81-3-5615-2894

2. Review of Papers and Procedures

(1) Review of a paper

- A manuscript is reviewed by professional reviewers of the relevant field. The reviewer will deem the paper "acceptance", "conditionally acceptance" or "returned". The applicant is notified of the result of the review by E-mail.
- Evaluation method

  Ordinary papers are usually evaluated on the following criteria:
  - ✓ Novelty: The contents of the paper are novel.
  - ✓ Utility: The contents are useful for academic and industrial development.
  - ✓ Reliability: The contents are considered trustworthy by the reviewer.
  - ✓ Comprehensibility: The contents of the paper are clearly described and understood by the reviewer without misunderstanding.

A short paper can be evaluated by having a quickness on the research content and evaluated to have new knowledge with results even if that is partial or for specific use, apart from the novelty and utility of an ordinary paper.

A system development paper is evaluated based on the following criteria, apart from the novelty and utility of an ordinary paper.

- ✓ Novelty of system development: Even when integrated with existing technologies, the novelty of the combination, novelty of the system, novelty of knowledge obtained from the developed system, etc. are recognized as the novelty of the system.
- ✓ Utility of system development: It is comprehensively or partially superior compared to similar systems. Demonstrates a pioneering new application concept as a system. The combination has appropriate optimality for practical use. Demonstrates performance

limitations and examples of performance of the system when put to practical use.

A data paper is considered novel if new deliverables of test, application and manufacturing, the introduction of new technology and proposals in the worksite have any priority, even though they are not necessarily original, apart from the novelty and utility of an ordinary paper. Also, if the new deliverables are superior compared to the existing technology and are useful for academic and industrial development, they should be evaluated.

A survey paper is evaluated by comprehensiveness, overviewing point, and objectiveness apart from the novelty of an ordinary paper. Reliability, comprehensibility, completeness of reference papers are common to those in an ordinary paper. Utility is evaluated how the paper will enlighten the readers in the target fields.

(2) Procedure after a review

- In case of acceptance, the author prepares a final manuscript (as mentioned in 3.).
- In the case of acceptance with comments by the reviewer, the author may revise the paper in consideration of the reviewer's opinion and proceed to prepare the final manuscript (as mentioned in 3.).
- In case of conditional acceptance, the author shall modify a paper based on the reviewer's requirements by a specified date (within 60 days), and submit the modified paper for approval. The corrected parts must be colored or underlined. A reply letter must be attached that carefully explains the corrections, assertions and future issues, etc., for all of the acceptance conditions.
- In case a paper is returned, the author cannot proceed to the next step. Please look at the reasons the reviewer lists for the return. We expect an applicant to try again after reviewing the content of the paper.

(3) Review request for a revised manuscript

- If you want to submit your paper after conditional acceptance, please submit the reply letter to the comments of the reviewers, and the revised manuscript with revision history to the submission site. Please note the designated date for submission. Revised manuscripts delayed more than the designated date be treated as new applications.
- In principle, a revised manuscript willl be reviewed by the same reviewer. It is judged either acceptance or returned.
- After the judgment, please follow the same procedure as (2).

3. Submission of final manuscript for publication

(1) Submission of a final manuscript

- An author, who has received the notice of "Acceptance", will receive an email regarding the creation of the final manuscript. The author shall prepare a complete set of the final manuscript (electronic data) following the instructions given and send it to the office by the designated date.
- The final manuscript shall contain a source file (TeX edition or MS Word version) and a PDF file, eps files for all drawings (including bmp, jpg, png), an eps file for author's photograph (eps or jpg file of more than 300 dpi with length and breadth ratio 3:2, upper part of the body) for authors' introduction. Please submit these in a compressed format, such as a zip file.
- In the final manuscript, write the name of the authors, name of an organizations, introduction of authors, and if necessary, an appreciation acknowledgment. (cancel macros in the Style file)

- An author whose paper is accepted shall pay a page charge before publishing. It is the author's decision to purchase offprints. (ref. page charge and offprint price information)

(2) Galley print proof

- The author is requested to check the galley (hard copy) a couple of weeks before the paper is published in the journal. Please check the galley by the designated date (within one week). After making any corrections, scan the data and prepare a PDF file, and send it to our office by email. At that time, fill in the Offprint Purchase Slip and Copyright Form and return the scanned data to our office in PDF file form.
- In principle, the copyrights of all articles published in our journal, including electronic form, belong to our society.
- You can download the Offprint Purchase Slip and the Copyright Form from the journal on our homepage. (ref. Attachment 2: Offprint Purchase Slip, Attachment 3: Copyright Form)

(3) Publication

- After final proofreading, a paper is published in the Academic journal or English transaction (both in electronic format) and will also be posted on our homepage.

IIEEJ Trans. on Image Electronics and Visual Computing

Vol. 12 No. 2 2024

Editor in Chief: Osamu Uchida
The Institute of Image Electronics Engineers of Japan
3-35-4-101, Arakawa, Arakawa-ku, Tokyo 116-0002, Japan

Print: ISSN 2188-1898
Online: ISSN 2188-1901
CD-ROM: ISSN 2188-191x